



Notes on Probability and Statistics (DRAFT)

Course Notes for CEE 3770

Compiled and edited by **Jorge A. Laval**



Contentsmkboth CONTENTS

I		Probability
1	Basics of Probability	9
1.1	Histograms	10
1.2	Terminology and set theory	10
1.2.1	Basic Laws	14
1.3	Combinatorics: counting strategies when outcomes are equally likely	17
1.3.1	More problems in combinatorics (optional)	23
1.4	Axioms of Probability	27
1.5	Addition rule	28
1.6	Conditional Probability	31
1.7	Independent Events	37
1.8	Law of Total Probability	41
1.9	Bayes' Theorem	44
1.9.1	Updating probability estimates	48
1.10	More Problems	51
2	Random Variables	55
2.1	Probability distribution function	56
2.2	Quantiles (aka percentiles)	56
2.3	Discrete Random Variables	57

2.4	Expectation	62
2.4.1	Expectation of a function of X	64
2.4.2	Variance	66
2.5	Jointly Distributed Discrete Random Variables	69
2.5.1	Chapter 1 results in PMF notation	71
2.5.2	Expectation with two random variables	78
2.6	Covariance	81
2.6.1	Interpretation of Covariance and Correlation	82
2.6.2	Covariance of linear combinations	84
3	Continuous Random Variables	101
3.1	Joint Continuous Variables	107
3.1.1	Exercises	118
4	Special Distributions	121
4.1	Uniform Random Variable	121
4.2	Normal Distribution	122
4.2.1	The Central Limit Theorem for sums	125
4.2.2	How to read normal probability tables	126
4.2.3	The “68-95-99.7 Rule”	127
4.3	Lognormal Distribution	135
4.3.1	The Central Limit Theorem for products	138
4.4	Bernoulli Family of Random Variables	144
4.4.1	Binomial random variable	145
4.4.2	The Multinomial distribution	152
4.4.3	Geometric Random Variables	153
4.4.4	Negative Binomial Random Variable	158
4.4.5	Hypergeometric Random Variable	159
4.5	Poisson Random Variables	161
4.6	Exponential Random Variable	166
4.7	Gamma (Erlang) distributions are sums of exponentials	169
4.8	The beta distribution: finite interval sample space	170
4.9	The Bivariate Normal Distribution	170
5	Function of Random Variables	173
5.1	One Random Variable	173
5.1.1	Single Discrete Random Variable	174
5.1.2	Single Continuous Random Variable	177
5.2	Two Random Variables	180
5.2.1	What if the distribution of X is unknown? (not covered)	182

5.3	Important distributions for statistics	185
5.3.1	The chi-square distribution with r degrees of freedom	185
5.3.2	Student's t -distribution	186

II

Statistics

6	Normal Random Samples	191
6.1	Theoretical building blocks	193
6.1.1	The Z, T and C^2 -statistics	193
6.1.2	Estimators	196
6.2	Confidence intervals	197
6.2.1	Confidence intervals for μ	198
6.2.2	Confidence intervals for σ^2	204
6.3	Hypothesis Testing	205
6.3.1	Basic t -test about μ	206
6.3.2	The magic 5% significance level (or p -value of 0.05)	209
6.3.3	Paired t -test	212
6.3.4	The two-sample t -test	213
6.3.5	Pearson's χ^2 test (goodness-of-fit test)	214
7	Linear regression	217
7.1	The regression model	217
7.2	Matrix notation	218
7.3	The method of ordinary least squares (OLS)	221
7.4	Testing the significance of coefficients	222
7.5	Goodness-of-fit: R^2	223
7.5.1	Adjusted R^2	223
7.5.2	One-way ANOVA	224
7.6	Assessing the model	224
7.7	Model selection	231
7.8	Making predictions	237
7.8.1	Prediction of the mean response at \mathbf{x}_0 , $E(Y \mathbf{x}_0) = \mathbf{x}_0\boldsymbol{\beta}$.	238
7.8.2	Prediction of a particular realization of $Y_0 = \mathbf{x}_0\boldsymbol{\beta} + \varepsilon_0$	238
7.9	Simple linear regression	239
7.9.1	Predictions	241
7.10	Problems	246

Probability

1 Basics of Probability 9

- 1.1 Histograms
- 1.2 Terminology and set theory
- 1.3 Combinatorics: counting strategies when outcomes are equally likely
- 1.4 Axioms of Probability
- 1.5 Addition rule
- 1.6 Conditional Probability
- 1.7 Independent Events
- 1.8 Law of Total Probability
- 1.9 Bayes' Theorem
- 1.10 More Problems

2 Random Variables 55

- 2.1 Probability distribution function
- 2.2 Quantiles (aka percentiles)
- 2.3 Discrete Random Variables
- 2.4 Expectation
- 2.5 Jointly Distributed Discrete Random Variables
- 2.6 Covariance

3 Continuous Random Variables .. 101

- 3.1 Joint Continuous Variables

4 Special Distributions 121

- 4.1 Uniform Random Variable
- 4.2 Normal Distribution
- 4.3 Lognormal Distribution
- 4.4 Bernoulli Family of Random Variables
- 4.5 Poisson Random Variables
- 4.6 Exponential Random Variable
- 4.7 Gamma (Erlang) distributions are sums of exponentials
- 4.8 The beta distribution: finite interval sample space
- 4.9 The Bivariate Normal Distribution

5 Function of Random Variables .. 173

- 5.1 One Random Variable
- 5.2 Two Random Variables
- 5.3 Important distributions for statistics

1. Basics of Probability

Probability has its origins in correspondence discussing **the mathematics of games of chance** between Blaise Pascal and Pierre de Fermat in the 17th century, and was formalized and rendered axiomatic as a distinct branch of mathematics by Andrey Kolmogorov in the 20th century.



The first attempt at mathematical rigour in the field of probability, championed by Pierre-Simon Laplace, is now known as the **classical definition**: probability is shared equally between all the possible outcomes, provided these outcomes can be deemed equally likely.

Two interpretations of probability :

- a) Frequentist (this course) If we denote by n_a the number of occurrences of an event A in n

trials, then if

$$\lim_{n \rightarrow +\infty} \frac{n_a}{n} = p \quad (1.1)$$

we say that the probability of A is p , or $P(A) = p$.

- Data are a repeatable random sample - there is a frequency
 - **Parameters (e.g. p) are fixed and unknown**
- b) Bayesian (not covered here)
- Parameters are *random variables*
 - **Data are fixed** but can be updated.
 - More info...

→ NOVA episode: Prediction by the numbers

1.1 Histograms

For a specific set of experimental data, a histogram shows the relative frequencies of the different observed values of a single variable. They may be constructed as follows.

1. select a range on the x-axis that is sufficient to cover the largest and smallest values among the set of data,
2. divide this range in “convenient” intervals or *bins*.
3. the y-axis can be either (i) the number of observations within each bin among the total number of observations, or (ii) the fraction of the total number.

→ Google image search for histograms in engineering

→ Galton machine video

→ Online histogram maker

→ Google public data

1.2 Terminology and set theory

To understand probability it helps to understand basic set theory. To illustrate the definitions below, we consider the example of a dice roll that generates outcomes in the set $\{1, 2, 3, 4, 5, 6\}$.

Experiment: An action with an uncertain outcome, e.g. a dice roll.

Sample space, S : The set of all possible outcomes of an experiment. In the example. $S = \{1, 2, 3, 4, 5, 6\}$.

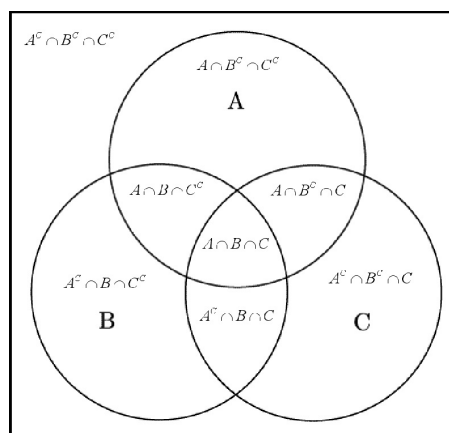
Event: Any subset of the sample space, $A \subset S$ would be an event, e.g. $A = \{1, 3, 5\}$. We say that the event has occurred if any of the outcomes in the event has happened.

Elementary event, ω : an event which contains only a single outcome in the sample space, e.g. $\omega = \{2\}$.

aka: basic outcome or simple event.

Venn diagrams

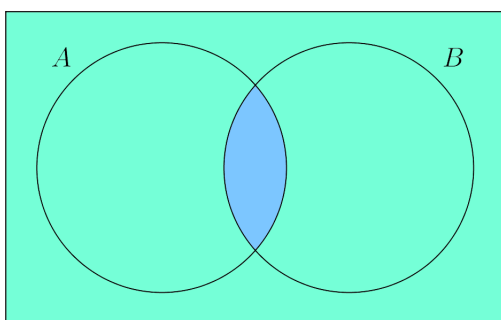
A Venn diagram shows all possible logical relations between a finite collection of different sets. These diagrams depict basic outcomes as points in the plane, and events as regions inside closed curves.



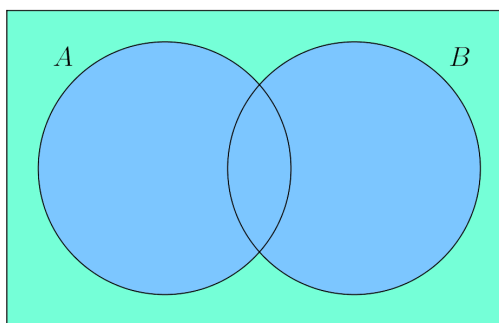
For the definitions below, let A and B be the events $A = \{1, 2, 3, 4\}$ and $B = \{4, 5, 6\}$.

Intersection of sets, $A \cap B$. The set of all outcomes that are both in A and B is called the intersection of A and B . In the example $A \cap B = \{4\}$

Union of sets, $A \cup B$. The set of all outcomes that are in either of A and B is called the union of A and B . In the example $A \cup B = \{1, 2, 3, 4, 5, 6\}$



■ $A \cap B$



■ $A \cup B$

Complement of set, A^c . The set of all outcomes that are **not in** A , but are in S is called the complement of A . In the example $A^c = \{5, 6\}$.

Note: \bar{A} and A' are also common notation for complement.

Mutually exclusive (ME) events. Events A and B are mutually exclusive if

$$A \cap B = \emptyset$$

where \emptyset is the empty set.

Collectively exhaustive (CE) events. Events A and B are collectively exhaustive if

$$A \cup B = S$$

Division. Events A_1, \dots, A_n form a division of the sample space S if

$$\bigcup_{k=1}^n A_k = S, \quad \text{and} \quad A_i \cap A_j = \emptyset \quad i \neq j.$$

Notation. For events A_1, \dots, A_n :

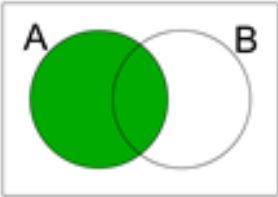
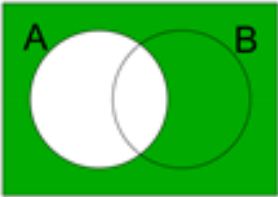
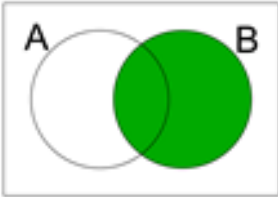

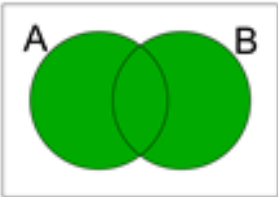
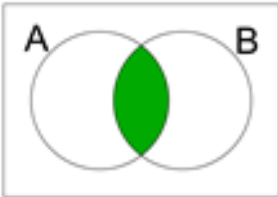
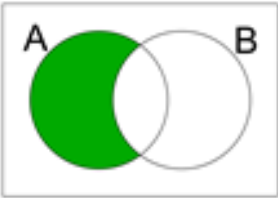
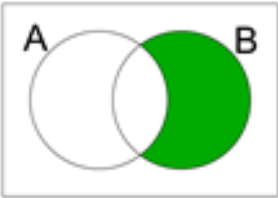
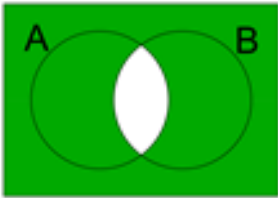

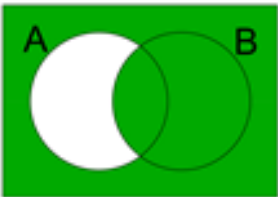



$$\bullet \bigcup_{k=1}^n A_k = A_1 \cup A_2 \cdots \cup A_n \quad \text{and} \quad \bigcap_{k=1}^n A_k = A_1 \cap A_2 \cdots \cap A_n.$$

For variables x_1, \dots, x_n :

$$\bullet \sum_{k=1}^n x_k = x_1 + x_2 \cdots + x_n \quad \text{and} \quad \prod_{k=1}^n x_k = x_1 x_2 \cdots x_n.$$

→ [image source](#)

2 Circle Venn Diagram Shading

 <p>A</p>	 <p>A'</p>	 <p>B</p>
 <p>B'</p>	 <p>$A \cup B$</p>	 <p>$A \cap B$</p>
 <p>$A \cap B'$</p>	 <p>$A' \cap B$</p>	 <p>$(A \cap B)'$ or $A' \cup B'$</p>
 <p>$A' \cap B'$ or $(A \cup B)'$</p>	 <p>$A' \cup B$ or $(A \cap B)'$</p>	 <p>$A \cup B'$ or $(A' \cap B)'$</p>
 <p>$(A' \cap B') \cup (A \cap B)$</p>	 <p>$(A \cap B') \cup (A' \cap B)$</p>	

Example 1. A coin is tossed twice. Let H stand for heads and T for tails. So:
 a) the elementary events are HH, HT, TH and TT

b) the sample space is $S = \{HH, HT, TH, TT\}$

Example 2. Suppose the travel time between two major cities A and B by air is 7 or 8 hr if the flight is nonstop; however, if there is one stop, the travel time would be 10, 11, or 12 hr. A nonstop flight between A and B would cost \$1000, whereas with one stop the cost is only \$650. Then, between cities B and C, all flights are nonstop requiring 2 or 3 hours at a cost of \$250. (There is no flight from A to C)

For a passenger wishing to travel from city A to city C,

- (a) What is the possibility space or sample space of his travel times from A to B? From A to C?
 (b) What is the sample space of his travel cost from A to B?
 (c) If T =travel time from city A to city C, and S =cost of travel from A to C, what is the sample space of T and S ?

Solution: (a) Sample space of travel time from A to B = $\{7, 8, 10, 11, 12\}$

Sample space of travel time from A to C = $\{9, 10, 11, 12, 13, 14, 15\}$

(b) Sample space of travel cost from A to B = $\{650, 1000\}$

(c) Sample space of $T = \{9, 10, 11, 12, 13, 14, 15\}$

Sample space of $S = \{900, 1250\}$

Sample space of T and $S = \{\{9, 1250\}, \{10, 1250\}, \{11, 1250\}, \{12, 900\}, \{13, 900\}, \{14, 900\}, \{15, 900\}\}$ \square

1.2.1 Basic Laws

Commutative Laws

$$A \cup B = B \cup A$$

$$A \cap B = B \cap A$$

Associative Laws

$$(A \cup B) \cup C = A \cup (B \cup C)$$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

Distributive Laws

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$$

$$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$$

Note

$$A \cup \emptyset = A$$

$$A \cap \emptyset = \emptyset$$

$$A \cup S = S$$

$$A \cap S = A$$

$$A \cup A = A$$

$$A \cap A = A$$

$$(A^c)^c = A$$

The intersection “is like” multiplication, and the union “is like” addition, but there is no double counting: $A \cup A \neq A$. If there are no parentheses, the intersection takes precedence over the union.

Example 3. — Simplify: $(A \cup C)(B \cup C)$

Solution:

$$\begin{aligned} (A \cup C)(B \cup C) &= AB \cup AC \cup BC \cup CC \\ &= AB \cup AC \cup BC \cup C \\ &= AB \cup AC \cup C \\ &= AB \cup C \end{aligned}$$

□

De Morgan's Laws

$$(A_1 \cup \dots \cup A_n)^c = A_1^c \cap \dots \cap A_n^c$$

$$(A_1 \cap \dots \cap A_n)^c = A_1^c \cup \dots \cup A_n^c$$

Proof.

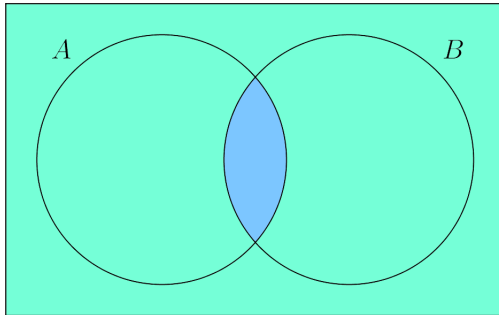
$$\begin{aligned} \omega \in (A_1 \cup \dots \cup A_n)^c &\iff \omega \notin A_1 \cup \dots \cup A_n \\ &\iff \omega \notin A_1 \text{ and } \omega \notin A_2 \text{ and } \dots \text{ and } \omega \notin A_n \\ &\iff \omega \in A_1^c \text{ and } \omega \in A_2^c \text{ and } \dots \text{ and } \omega \in A_n^c \\ &\iff \omega \in A_1^c \cap \dots \cap A_n^c \end{aligned}$$

Similarly,

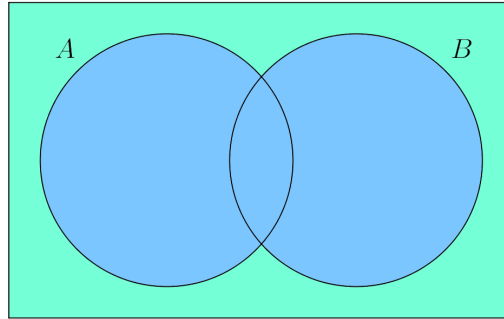
$$\begin{aligned} \omega \in (A_1 \cap \dots \cap A_n)^c &\iff \omega \notin A_1 \cap \dots \cap A_n \\ &\iff \omega \notin A_1 \text{ or } \dots \text{ or } \omega \notin A_n \\ &\iff \omega \in A_1^c \text{ or } \dots \text{ or } \omega \in A_n^c \\ &\iff \omega \in A_1^c \cup \dots \cup A_n^c \end{aligned}$$

■

For 2 events:



$A \cap B$
 $(A \cap B)^c = A^c \cup B^c$



$A \cup B$
 $(A \cup B)^c = A^c \cap B^c$

Not (A and B) is the same as Not A or Not B.

Not (A or B) is the same as Not A and Not B.

Example 4. — A chain Consider a simple chain consisting of two links. The chain will fail to carry a given load if either link breaks. Let:

A = the breakage of link 1

B = the breakage of link 2

Then,

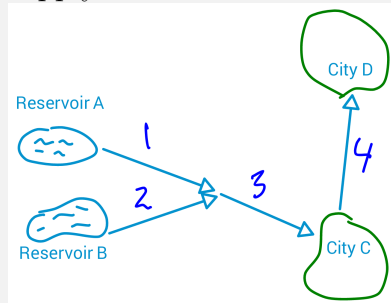
the chain fails = $A \cup B$

No failure of the chain, therefore, is the complement $(A \cup B)^c$. However, no failure of the chain also means that both links survive (no breakage); that is,

the chain does not fail = $A^c \cap B^c$

which is a demonstration of the validity of De Morgan's rule.

Example 5. — Water supply system The water supply for two cities C and D comes from the two sources A and B as shown in the figure. Water is transported by pipelines consisting of branches 1, 2, 3, and 4. Assume that either one of the two sources, by itself, is sufficient to supply the water for both cities.



Denote: E_i = failure of branch $i = 1, 2, 3, 4$. Failure of a pipe branch means there is serious leakage or rupture of the branch. Using these events, express the events

a) water shortage in city C

- b) no water shortage in city C (simplify using De Morgan)
- c) water shortage in city D
- d) no water shortage in city D

Solution: We have:

- a) water shortage in city C: $E_1 E_2 \cup E_3$
- b) no water shortage in city C: $(E_1 E_2 \cup E_3)^c = (E_1 E_2)^c E_3^c = (E_1^c \cup E_2^c) E_3^c$
- c) water shortage in city D: $E_1 E_2 \cup E_3 \cup E_4$
- d) no water shortage in city D: $(E_1 E_2 \cup E_3 \cup E_4)^c = (E_1^c \cup E_2^c) E_3^c E_4^c$

□

1.3 Combinatorics: counting strategies when outcomes are equally likely

Motivation for combinatorics:

Fact 1.7 — Probability when outcomes are equally likely. If all outcomes of an experiment are equally likely, the probability of an event A happening is:

$$P(A) = \frac{\text{number of outcomes favorable to } A}{\text{number of outcomes}} = \frac{|A|}{|S|}$$

where $|A|$ is the size of set A .

→ Combinatorics helps calculate $|A|$ and $|S|$.

Example 6. At **Coco's restaurant** you have

- a) two choices for appetizers: soup or juice;
- b) three for the main course: a meat, fish, or vegetable dish; and
- c) two for dessert: ice cream or cake.

How many possible choices do you have for your complete meal?

Solution: We illustrate the possible meals by a tree diagram, where we see that the total number of choices is the product of the number of choices at each stage. In this examples we have $2 \cdot 3 \cdot 2 = 12$ possible menus. □

Our menu example illustrates the following general counting technique.

Basic Rule in combinatorics In an experiment with k steps, if

- the 1st step has n_1 possible outcomes,
- the 2nd step has n_2 possible outcomes,
- ...
- the k th step has n_k possible outcomes, then there are:

$$n_1 \times n_2 \times \cdots \times n_k \tag{1.2}$$

possible outcomes for the whole experiment.

Example 7. A finite set Ω has n elements. Show that if we count the empty set and Ω as subsets, there are 2^n subsets of Ω .

Solution: The experiment of generating a subset of Ω can be broken down in n steps, one for each element in Ω , and for each element we have 2 choices: we either pick or do not pick the element. \square

Example 8. — Permutations. The English alphabet has 26 letters. How many 5-letter “words” are there if:

- a) repeated letters are allowed (experiment with replacement)
- b) repeated letters are not allowed (without replacement)

Solution:

- a) repeated letters are allowed: $26 \times 26 \times 26 \times 26 \times 26 = 26^5$
- b) repeated letters are not allowed: $26 \times 25 \times 24 \times 23 \times 22$

\square

Example 9. In example 8, how many 5-letter “words” contain NO letter “S”, if:

- a) repeated letters are allowed (with replacement)
- b) repeated letters are not allowed (without replacement)

Solution:

- a) repeated letters are allowed: 25^5
- b) repeated letters are not allowed: $25 \times 24 \times 23 \times 22 \times 21$

\square

Example 10. In example 8, how many 5-letter “words” contain at least 1 letter “S”, if:

- a) repeated letters are allowed (with replacement)
- b) repeated letters are not allowed (without replacement)

Solution:

- a) repeated letters are allowed: $26^5 - 25^5$
- b) repeated letters are not allowed: $26 \times 25 \times 24 \times 23 \times 22 - 25 \times 24 \times 23 \times 22 \times 21$

\square

Example 11. In example 8, how many 5-letter “words” contain exactly one letter “S”, if:

- a) repeated letters are allowed (with replacement)
- b) repeated letters are not allowed (without replacement)

Solution:

- a) repeated letters are allowed: 5×25^4
(first step: place letter S in any of the five spots, second step: fill any of the four spots available with any of the 25 remaining letters,...)
- b) repeated letters are not allowed: $5 \times 25 \times 24 \times 23 \times 22$

\square

Example 12. In example 8, how many 5-letter “words” contain exactly 1 letter “S” and 1 letter “O”, if:

- a) repeated letters are allowed (with replacement)
- b) repeated letters are not allowed (without replacement)

Solution:

- a) repeated letters are allowed: $5 \times 4 \times 24^3$
(first step: place letter S in any of the five spots, second step: place letter O in any of the four remaining spots, third step: fill any of the 3 remaining spots with any of the 24 remaining

letters,...)

b) repeated letters are not allowed: $5 \times 4 \times 24 \times 23 \times 22$

□

Ordered sequence (word) is a list of elements where the order matters, e.g. $\{1, 2, 3, 2\} \neq \{1, 2, 2, 3\}$.

Permutations: The number of distinct **ordered sequences** with k elements that can be chosen from a set with n elements.

Fact 1.9 The number of **permutations with replacement** of k objects out of n is

$$n^k$$

and gives the number of *ordered sequences* possible when selecting k objects out of n *with replacement*.

Fact 1.10 The number of **permutations without replacement** of k objects out of n ,

$${}^n P_k = n(n-1)(n-2)\dots(n-k+1) = \frac{n!}{(n-k)!}$$

gives the number of *ordered sequences* possible when selecting k objects out of n *without replacement*.

We can alter this formula to disregard ordering by eliminating each ordering of each set of objects. Since we are choosing k objects from a set of n objects, those k objects can be ordered in $k!$ ways. So, if we simply divide ${}^n P_k$ by $k!$, we then have the number of ways we can select k objects from n without replacement and without regard for order. This is called the number of combinations of n taken k at a time:

Example 13. — Combinations. How many **groups** of 3 students can be made in a class of 4 students?

Solution: There are ${}^4 P_3$ **ordered** sequences of 3 students:

$$\begin{vmatrix} \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{2} & \mathbf{2} & \mathbf{2} & \mathbf{2} & \mathbf{2} & \mathbf{2} & \mathbf{3} & \mathbf{3} & \mathbf{3} & \mathbf{3} & \mathbf{3} & \mathbf{3} & \mathbf{4} & \mathbf{4} & \mathbf{4} & \mathbf{4} & \mathbf{4} & \mathbf{4} \\ \mathbf{2} & \mathbf{2} & \mathbf{3} & \mathbf{3} & \mathbf{4} & \mathbf{4} & \mathbf{1} & \mathbf{1} & \mathbf{3} & \mathbf{3} & \mathbf{4} & \mathbf{4} & \mathbf{1} & \mathbf{1} & \mathbf{2} & \mathbf{2} & \mathbf{4} & \mathbf{4} & \mathbf{1} & \mathbf{1} & \mathbf{2} & \mathbf{2} & \mathbf{3} & \mathbf{3} \\ \mathbf{3} & \mathbf{4} & \mathbf{2} & \mathbf{4} & \mathbf{2} & \mathbf{3} & \mathbf{3} & \mathbf{4} & \mathbf{1} & \mathbf{4} & \mathbf{1} & \mathbf{3} & \mathbf{2} & \mathbf{4} & \mathbf{1} & \mathbf{4} & \mathbf{1} & \mathbf{2} & \mathbf{2} & \mathbf{3} & \mathbf{1} & \mathbf{3} & \mathbf{1} & \mathbf{2} \end{vmatrix}$$

Since 3 students can be shuffled in $3! = 6$ different ways, the answer is ${}^4 P_3 / 3!$. □

An unordered sequence (group) is a list of elements where the order DOES NOT matter, e.g. the group $\{1, 2, 3, 4\}$ is equivalent to $\{1, 4, 2, 3\}$.

Combinations: The number of distinct **unordered sequences** with k elements that can be chosen, without replacement, from a set with n elements is denoted by $\binom{n}{k}$, and is pronounced “ n choose k .” The number $\binom{n}{k}$ is called a binomial coefficient.

Fact 1.11 The number of **combinations without replacement** of k objects out of n ,

$$\binom{n}{k} = {}^n P_k / k! = \frac{n!}{(n-k)!k!}$$

gives the number of *unordered sequences* possible when selecting k objects out of n *without replacement*.

Note: $\binom{n}{k} = \binom{n}{n-k}$

The sampling table gives the number of possible samples (sequences) of size k out of a population of size n , depending on how the sample is collected.

	Permutations (order matters)	Combinations (not matter)
With Replacement	n^k	$\binom{n+k-1}{k}$
Without Replacement	${}^n P_k = \frac{n!}{(n-k)!}$	$\binom{n}{k}$

→ More info on combinations with replacement...

Example 14. — **Georgia lottery Powerball** Pick 6 different numbers:

- 5 between 1-69 and
- 1 PowerBall number between 1-26.

what are the chances of winning?

Solution: $|S| = \binom{69}{5} \binom{26}{1} = 292,201,338 \rightarrow P(\text{winning}) = 1/292,201,338$. □

Example 15. * There are n students in a classroom. Assuming 365 days in a year, what is the probability that everyone has distinct birthdays, ignoring the year?

Solution: There are 365 options for each person's birthday. The sample space is all possible birthday sequences of length n . Therefore,

$$|S| = 365^n$$

Let A be the event that everyone has distinct birthdays.

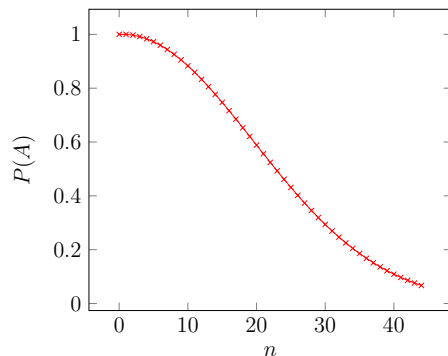
If $n > 365$, at least two persons must share a birthday.

If $n \leq 365$,

$$|A| = 365 \times 364 \dots (365 - n + 1) = {}^{365} P_n$$

Therefore,

$$\begin{aligned} P(A) &= \frac{{}^{365} P_n}{365^n} \\ &= \frac{365!}{(365-n)! 365^n} \end{aligned}$$



NOTE: the probability of at least two persons sharing a birthday is A^c , and

$$|A^c| = |S| - |A|, \quad \text{and therefore:}$$

$$P(A^c) = \frac{|S| - |A|}{|S|} = 1 - P(A)$$

(1.3)

□

Important note on identical objects: Consider n objects of which k are identical of type A and the remaining $(n - k)$ of type B. The number of ways one can arrange these n objects is

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}.$$

Why? label each of the n objects with a unique number so that all objects are different. These n objects can be arranged in $n!$ different ways. If we remove the labels, we would see that each particular arrangement is repeated $k!(n - k)!$ times (recall that k distinct objects can be ordered in $k!$ ways).

For example, the number of arrangements of 3 A's and 2 B's **after labeling** is $5!=120$:

$\{A_1, A_2, A_3, B_1, B_2\}$	$\{A_1, A_2, A_3, B_2, B_1\}$	$\{A_1, A_2, B_1, A_3, B_2\}$	$\{A_1, A_2, B_1, B_2, A_3\}$	$\{A_1, A_2, B_2, A_3, B_1\}$
$\{A_1, A_2, B_2, B_1, A_3\}$	$\{A_1, A_3, A_2, B_1, B_2\}$	$\{A_1, A_3, A_2, B_2, B_1\}$	$\{A_1, A_3, B_1, A_2, B_2\}$	$\{A_1, A_3, B_1, B_2, A_2\}$
$\{A_1, A_3, B_2, A_2, B_1\}$	$\{A_1, A_3, B_2, B_1, A_2\}$	$\{A_1, B_1, A_2, A_3, B_2\}$	$\{A_1, B_1, A_2, B_2, A_3\}$	$\{A_1, B_1, A_3, A_2, B_2\}$
$\{A_1, B_1, A_3, B_2, A_2\}$	$\{A_1, B_1, B_2, A_2, A_3\}$	$\{A_1, B_1, B_2, A_3, A_2\}$	$\{A_1, B_2, A_2, A_3, B_1\}$	$\{A_1, B_2, A_2, B_1, A_3\}$
$\{A_1, B_2, A_3, A_2, B_1\}$	$\{A_1, B_2, A_3, B_1, A_2\}$	$\{A_1, B_2, B_1, A_2, A_3\}$	$\{A_1, B_2, B_1, A_3, A_2\}$	$\{A_2, A_1, A_3, B_1, B_2\}$
$\{A_2, A_1, A_3, B_2, B_1\}$	$\{A_2, A_1, B_1, A_3, B_2\}$	$\{A_2, A_1, B_1, B_2, A_3\}$	$\{A_2, A_1, B_2, A_3, B_1\}$	$\{A_2, A_1, B_2, B_1, A_3\}$
$\{A_2, A_3, A_1, B_1, B_2\}$	$\{A_2, A_3, A_1, B_2, B_1\}$	$\{A_2, A_3, B_1, A_1, B_2\}$	$\{A_2, A_3, B_1, B_2, A_1\}$	$\{A_2, A_3, B_2, A_1, B_1\}$
$\{A_2, A_3, B_2, B_1, A_1\}$	$\{A_2, B_1, A_1, A_3, B_2\}$	$\{A_2, B_1, A_1, B_2, A_3\}$	$\{A_2, B_1, A_3, A_1, B_2\}$	$\{A_2, B_1, A_3, B_2, A_1\}$
$\{A_2, B_1, B_2, A_1, A_3\}$	$\{A_2, B_1, B_2, A_3, A_1\}$	$\{A_2, B_2, A_1, A_3, B_1\}$	$\{A_2, B_2, A_1, B_1, A_3\}$	$\{A_2, B_2, A_3, A_1, B_1\}$
$\{A_2, B_2, A_3, B_1, A_1\}$	$\{A_2, B_2, B_1, A_1, A_3\}$	$\{A_2, B_2, B_1, A_3, A_1\}$	$\{A_3, A_1, A_2, B_1, B_2\}$	$\{A_3, A_1, A_2, B_2, B_1\}$
$\{A_3, A_1, B_1, A_2, B_2\}$	$\{A_3, A_1, B_1, B_2, A_2\}$	$\{A_3, A_1, B_2, A_2, B_1\}$	$\{A_3, A_1, B_2, B_1, A_2\}$	$\{A_3, A_2, A_1, B_1, B_2\}$
$\{A_3, A_2, A_1, B_2, B_1\}$	$\{A_3, A_2, B_1, A_1, B_2\}$	$\{A_3, A_2, B_1, B_2, A_1\}$	$\{A_3, A_2, B_2, A_1, B_1\}$	$\{A_3, A_2, B_2, B_1, A_1\}$
$\{A_3, B_1, A_1, A_2, B_2\}$	$\{A_3, B_1, A_1, B_2, A_2\}$	$\{A_3, B_1, A_2, A_1, B_2\}$	$\{A_3, B_1, A_2, B_2, A_1\}$	$\{A_3, B_1, B_2, A_1, A_2\}$
$\{A_3, B_1, B_2, A_2, A_1\}$	$\{A_3, B_2, A_1, A_2, B_1\}$	$\{A_3, B_2, A_1, B_1, A_2\}$	$\{A_3, B_2, A_2, A_1, B_1\}$	$\{A_3, B_2, A_2, B_1, A_1\}$
$\{A_3, B_2, B_1, A_1, A_2\}$	$\{A_3, B_2, B_1, A_2, A_1\}$	$\{B_1, A_1, A_2, A_3, B_2\}$	$\{B_1, A_1, A_2, B_2, A_3\}$	$\{B_1, A_1, A_3, A_2, B_2\}$
$\{B_1, A_1, A_3, B_2, A_2\}$	$\{B_1, A_1, B_2, A_2, A_3\}$	$\{B_1, A_1, B_2, A_3, A_2\}$	$\{B_1, A_2, A_1, A_3, B_2\}$	$\{B_1, A_2, A_1, B_2, A_3\}$
$\{B_1, A_2, A_3, A_1, B_2\}$	$\{B_1, A_2, A_3, B_2, A_1\}$	$\{B_1, A_2, B_2, A_1, A_3\}$	$\{B_1, A_2, B_2, A_3, A_1\}$	$\{B_1, A_3, A_1, A_2, B_2\}$
$\{B_1, A_3, A_1, B_2, A_2\}$	$\{B_1, A_3, A_2, A_1, B_2\}$	$\{B_1, A_3, A_2, B_2, A_1\}$	$\{B_1, A_3, B_2, A_1, A_2\}$	$\{B_1, A_3, B_2, A_2, A_1\}$
$\{B_1, B_2, A_1, A_2, A_3\}$	$\{B_1, B_2, A_1, A_3, A_2\}$	$\{B_1, B_2, A_2, A_1, A_3\}$	$\{B_1, B_2, A_2, A_3, A_1\}$	$\{B_1, B_2, A_3, A_1, A_2\}$
$\{B_1, B_2, A_3, A_2, A_1\}$	$\{B_2, A_1, A_2, A_3, B_1\}$	$\{B_2, A_1, A_2, B_1, A_3\}$	$\{B_2, A_1, A_3, A_2, B_1\}$	$\{B_2, A_1, A_3, B_1, A_2\}$
$\{B_2, A_1, B_1, A_2, A_3\}$	$\{B_2, A_1, B_1, A_3, A_2\}$	$\{B_2, A_2, A_1, A_3, B_1\}$	$\{B_2, A_2, A_1, B_1, A_3\}$	$\{B_2, A_2, A_3, A_1, B_1\}$
$\{B_2, A_2, A_3, B_1, A_1\}$	$\{B_2, A_2, B_1, A_1, A_3\}$	$\{B_2, A_2, B_1, A_3, A_1\}$	$\{B_2, A_3, A_1, A_2, B_1\}$	$\{B_2, A_3, A_1, B_1, A_2\}$
$\{B_2, A_3, A_2, A_1, B_1\}$	$\{B_2, A_3, A_2, B_1, A_1\}$	$\{B_2, A_3, B_1, A_1, A_2\}$	$\{B_2, A_3, B_1, A_2, A_1\}$	$\{B_2, B_1, A_1, A_2, A_3\}$
$\{B_2, B_1, A_1, A_3, A_2\}$	$\{B_2, B_1, A_2, A_1, A_3\}$	$\{B_2, B_1, A_2, A_3, A_1\}$	$\{B_2, B_1, A_3, A_1, A_2\}$	$\{B_2, B_1, A_3, A_2, A_1\}$

If we remove the labels, we see that each pattern is repeated $3!2!=12$ times:

$\{A, A, A, B, B\}^*$	$\{A, A, A, B, B\}^*$	$\{A, A, B, A, B\}$	$\{A, A, B, B, A\}$	$\{A, A, B, A, B\}$
$\{A, A, B, B, A\}$	$\{A, A, A, B, B\}^*$	$\{A, A, A, B, B\}^*$	$\{A, A, B, A, B\}$	$\{A, A, B, B, A\}$
$\{A, A, B, A, B\}$	$\{A, A, B, B, A\}$	$\{A, B, A, A, B\}$	$\{A, B, A, B, A\}$	$\{A, B, A, A, B\}$
$\{A, B, A, B, A\}$	$\{A, B, B, A, A\}$	$\{A, B, B, A, A\}$	$\{A, B, A, A, B\}$	$\{A, B, A, B, A\}$
$\{A, B, A, A, B\}$	$\{A, B, A, B, A\}$	$\{A, B, B, A, A\}$	$\{A, B, B, A, A\}$	$\{A, A, A, B, B\}^*$
$\{A, A, A, B, B\}^*$	$\{A, A, B, A, B\}$	$\{A, A, B, B, A\}$	$\{A, A, B, A, B\}$	$\{A, A, B, B, A\}$
$\{A, A, A, B, B\}^*$	$\{A, A, A, B, B\}^*$	$\{A, A, B, A, B\}$	$\{A, A, B, B, A\}$	$\{A, A, B, A, B\}$
$\{A, A, B, B, A\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$
$\{A, B, B, A, A\}$	$\{A, B, B, A, A\}$	$\{A, B, A, A, B\}$	$\{A, B, A, B, A\}$	$\{A, B, A, A, B\}$
$\{A, B, A, B, A\}$	$\{A, B, B, A, A\}$	$\{A, B, B, A, A\}$	$\{A, A, A, B, B\}^*$	$\{A, A, A, B, B\}^*$
$\{A, B, A, A, B\}$	$\{A, A, B, B, A\}$	$\{A, A, B, B, A\}$	$\{A, A, B, B, A\}$	$\{A, A, B, B, A\}$
$\{A, A, A, B, B\}^*$	$\{A, A, B, A, B\}$	$\{A, A, B, A, B\}$	$\{A, A, B, A, B\}$	$\{A, A, B, A, B\}$
$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$
$\{A, B, B, A, A\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$	$\{A, B, A, A, B\}$
$\{A, B, B, A, A\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$
$\{B, A, A, B, A\}$	$\{B, A, B, A, A\}$	$\{B, A, B, A, A\}$	$\{B, A, A, A, B\}$	$\{B, A, A, B, A\}$
$\{B, A, A, A, B\}$	$\{B, A, A, B, A\}$	$\{B, A, B, A, A\}$	$\{B, A, B, A, A\}$	$\{B, A, A, A, B\}$
$\{B, A, A, B, A\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, B, A\}$
$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$
$\{B, B, A, A, A\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$
$\{B, A, B, A, A\}$	$\{B, A, B, A, A\}$	$\{B, A, A, A, B\}$	$\{B, A, A, B, A\}$	$\{B, A, A, A, B\}$
$\{B, A, A, B, A\}$	$\{B, A, B, A, A\}$	$\{B, A, B, A, A\}$	$\{B, A, A, A, B\}$	$\{B, A, A, B, A\}$
$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$	$\{B, A, A, A, B\}$
$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$	$\{B, B, A, A, A\}$

So there are only $\binom{5}{k} = \binom{5}{3} = 5!/(3!2!) = 10$ different patterns:

$\{A, A, A, B, B\}$ $\{A, A, B, A, B\}$ $\{A, A, B, B, A\}$ $\{A, B, A, A, B\}$ $\{A, B, A, B, A\}$
 $\{A, B, B, A, A\}$ $\{B, A, A, A, B\}$ $\{B, A, A, B, A\}$ $\{B, A, B, A, A\}$ $\{B, B, A, A, A\}$

In general:

Several groups of identical objects: Consider n_i identical objects of type $i = 1, 2, \dots, m$. The number of ways one can arrange these $n = \sum_{i=1}^m n_i$ objects is

$$\frac{n!}{n_1!n_2!\dots n_m!}$$

Example 16. A *Bernoulli* trial is an experiment that can result in either a success or a failure. In how many ways can the results of n Bernoulli trials be arranged such that there are k successes?

Solution: $\binom{n}{k}$ □

Example 17. The Georgia Tech football team played 8 games in a season, winning 3, losing 3, and ending 2 in a tie. In how many ways could this have happened?

Solution: $\frac{8!}{3!3!2!}$ □

Example 18. — Books on a shelf In how many ways can we arrange 3 books on a shelf with capacity for 4 books?

Solution: $|A| = \binom{4}{3} = 4$ □

Example 19. A nickel is tossed 4 times. What is the probability of obtaining 3 heads?

Solution: $|S| = 2^4 = 16$

The number of outcomes that yield 3H is identical to the number of ways we can arrange 3 books on a shelf with capacity for 4 books, so:

$|A| = \binom{4}{3} = 4$ and $P(A) = 4/16$. □

Example 20. A nickel is tossed 4 times. What is the probability of obtaining 3 heads if the coin is biased with a probability of heads of 0.53?

Solution: Since the coin is biased outcomes are not equally likely and therefore combinatorics techniques cannot be applied. Later we will see that this is given by the binomial distribution. □

Example 21. 8 books are to be arranged on 2 shelves, of capacities 3 and 5 respectively. Out of the 8 books, 2 books are special. Find the probability that the two special books end up on the same shelf.

Solution: [1] If the special books are to be placed on the longer shelf, the possible combinations are $\binom{5}{2}$.

If the special books are to be placed on the shorter shelf, the possible combinations are $\binom{3}{2}$.

To total possible arrangements are $\binom{8}{2}$.

Therefore, the required probability is $\frac{\binom{5}{2} + \binom{3}{2}}{\binom{8}{2}} = 13/28$. □

Solution: [2] Let the special books be placed first.

If the first special book is placed on the longer shelf, then it has 5 available positions, and the

second special book has 4 available positions.

If the first special book is placed on the shorter shelf, then it has 3 available positions, and the second special book has 2 available positions.

In either case, the number of ways of arranging the remaining 6 books in the remaining positions is $6!$.

Therefore, the total number of arrangements satisfying the conditions is $(5 \cdot 4 \cdot 6! + 3 \cdot 2 \cdot 6!)$. The total number of arrangements is $8!$. Therefore, the required probability is $\frac{5 \cdot 4 \cdot 6! + 3 \cdot 2 \cdot 6!}{8!} = 13/28$. \square

Example 22. — * The state of GA has license plates showing three numbers and four letters. How many different license plates are possible

- if the numbers must come after the letters?
- if there is no restriction on where the letters and numbers appear?
- as in part b) but replacing all numbers by an “A” and all letters by “B”?
- BONUS: as in part b) but replacing all numbers < 5 by an “A”, all numbers ≥ 5 by “B” and all letters by “C”?

Solution:

- $10^3 \times 26^4$
- There are $\binom{7}{3}$ possible arrangements of letters and numbers, each having the same number of outcomes as in part a), so the answer is $\binom{7}{3} \times 10^3 \times 26^4$.
- $\binom{7}{3}$
- $\sum_{n_A=0}^3 \frac{7!}{n_A!(3-n_A)!4!}$

\square

Example 23. Letters and mailboxes.]

- How many ways can six indistinguishable letters be put in three mail boxes?
- Using part a) above, show that r indistinguishable objects can be put in n boxes in

$$\binom{n+r-1}{n-1} = \binom{n+r-1}{r}$$

different ways.

Solution:

- One representation of this is given by the sequence $|LL|L|LLL|$ where the $|$'s represent the partitions for the boxes and the L's the letters. Any possible way can be so described. Note that we need two bars at the ends and the remaining two bars and the six L's can be put in any order. In this way, the problem boils down to shuffling six identical objects and two identical objects, therefore the answer is $\binom{8}{2}$ or $\binom{8}{6}$. Both give the same answer
- same logic as above using r letters n mailboxes.

\square

1.3.1 More problems in combinatorics (optional)

Example 24. Three balls are to be randomly selected, without replacement, from an urn containing 20 balls numbered 1 to 20. If Alice bets that at least one of the balls drawn has a number as large as or larger than 17, what is the probability that Alice wins the bet?

Solution: Let X be the largest number selected.

Therefore, X is a random variable which has a value from $\{3, \dots, 20\}$.

Let the value of the highest valued ball be i . Therefore, the number of ways to select the remaining two balls is $\binom{i-1}{2}$.

Therefore, the probability of the value of the highest valued ball being i is

$$P(X = i) = \frac{\binom{i-1}{2}}{\binom{20}{3}}$$

Therefore,

$$\begin{aligned} P(X \geq 17) &= P(X = 17) + P(X = 18) + P(X = 19) + P(X = 20) \\ &= \frac{\binom{16}{2} + \binom{17}{2} + \binom{18}{2} + \binom{19}{2}}{\binom{20}{3}} \end{aligned}$$

□

Example 25. A president, a treasurer, and a secretary, all different, are to be chosen from a club consisting of 10 people. How many different choices of office bearers are possible if

1. There are no restrictions.
2. Alice and Bob cannot serve together.
3. Charlie and David can serve together or not at all.
4. Eve must be an officer.
5. Frank can serve only if he is the president.

Solution:

1. The possible choices are ${}^{10}P_3$.
2. If neither Alice nor Bob are office bearers, there are 8P_3 possible choices.
If one of Alice and Bob is an office bearer, there are three possible posts for the selected person. The number of choices for the rest of the posts are 8P_2 . Same for Bob. Therefore, the total number of choices are ${}^8P_3 + 2 \cdot 3 \cdot {}^8P_2$.
3. If both Charlie and David are chosen, the number of choices is $3 \cdot 2 \cdot {}^8P_1$. If neither Charlie nor David are chosen, the number of choices is 8P_3 . Therefore, the total number of choices are $3 \cdot 2 \cdot \binom{8}{1} + {}^8P_3$.
4. There are three possible posts for Eve. Therefore, the total number of choices are $3 \cdot {}^9P_2$.
5. If Frank is the president, the number of choices is 9P_2 . If Frank is not the president, the number of choices is 9P_3 . Therefore, the total number of choices is ${}^9P_2 + {}^9P_3$.

□

Example 26. a different balls are divided randomly into n different cells. Find the probability that all cells are non-empty when

- a) $a = n$
- b) $a = n + 1$

Solution: We have:

a)

$$\begin{aligned} |S| &= n^a \\ &= n^n \end{aligned}$$

If all cells are to be non-empty, the number of combinations is $|A| = n!$. Therefore, the probability is $\frac{n!}{n^n}$.

b)

$$\begin{aligned} |S| &= n^a \\ &= n^{n+1} \end{aligned}$$

The number of combinations to select 2 balls is $\binom{n+1}{2}$. Let these two balls be glued together and be treated as one.

The number of arrangements of these n balls are $n!$.

Therefore, the total number of combinations are $\binom{n+1}{2}n!$.

Therefore, the probability is $\frac{\binom{n+1}{2}}{n!}$.

□

Example 27. — Poker: why a four of a kind beats a full house? A poker hand is a random subset of 5 elements from a deck of 52 cards.

- a) How many hands have four of a kind?
- b) How many hands have a full house?

Solution:

- a) How many hands have four of a kind? There are 13 ways that we can specify the value for the four cards. For each of these, there are 48 possibilities for the fifth card. Thus, the number of four-of-a-kind hands is $13 \cdot 48 = 624$. Since the total number of possible hands is $\binom{52}{5} = 2598960$, the probability of a hand with four of a kind is $624/2598960 = .00024$.
- b) Full house: There are 13 choices for the value which occurs three times; for each of these there are $\binom{4}{3} = 4$ choices for the particular three cards of this value that are in the hand. Having picked these three cards, there are 12 possibilities for the value which occurs twice; for each of these there are $\binom{4}{2} = 6$ possibilities for the particular pair of this value. Thus, the number of full houses is $13 \cdot 4 \cdot 12 \cdot 6 = 3744$, and the probability of obtaining a hand with a full house is $3744/2598960 = .0014$.

Thus, while both types of hands are unlikely, you are six times more likely to obtain a full house than four of a kind. □

Example 28. 5 cards are taken out randomly from a 52 card deck. Consider the following events.

- a) A : All cards are higher than 10.
- b) B : All cards are hearts.
- c) C : All cards have different numbers.
- d) D : All cards are consecutive numbers.

Assuming ace to have value 1, find the probabilities of A , B , C , and D .

Solution:

$$|S| = \binom{52}{5}$$

- a) There are 12 cards with numbers higher than 10. Therefore,

$$|A| = \binom{12}{5}$$

Therefore,

$$P(A) = \frac{|A|}{|S|} = \frac{\binom{12}{5}}{\binom{52}{5}}$$

b) There are 13 heart cards. Therefore,

$$|B| = \binom{13}{5}$$

Therefore,

$$P(B) = \frac{|B|}{|S|} = \frac{\binom{13}{5}}{\binom{52}{5}}$$

c) There are $52 \times 48 \times 44 \times 40 \times 36$ ways to have all cards with different numbers, but since order does not matter:

$$|C| = 52 \times 48 \times 44 \times 40 \times 36 / 5! = 1,317,888$$

Alternatively, the number of ways of selecting 5 different numbers out of 13 is $\binom{13}{5}$. For each of the selected number, there are 4 cards, of which exactly one has to be selected. Therefore,

$$|C| = \binom{13}{5} 4^5 = 1,317,888$$

Therefore,

$$P(C) = \frac{1,317,888}{\binom{52}{5}}$$

d) There are 9 sequences of consecutive numbers. Each of the numbers have 4 corresponding cards each. Therefore,

$$|D| = 9 \cdot 4^5$$

Therefore,

$$P(D) = \frac{9 \cdot 4^5}{\binom{52}{5}}$$

□

Example 29. A dice is tossed 3 times. Consider the following events.

a) A : The sum of all three numbers is even.

Find the probability of A .

Solution: Every time the dice is rolled, there are 6 possible outcomes. Therefore,

$$|S| = 6^3$$

a) For the sum of three numbers to be even, exactly 0 or 2 of them must be odd.

There is $\binom{3}{0} = 1$ combination for all three numbers to be even. Each of these even numbers has 3 options. Therefore, the total number of combinations following the restriction is 1×3^3 .

There are $\binom{3}{2} = 3$ combinations for exactly 2 numbers to be odd. Each of the odd numbers has 3 options, and the even number has 3 options. Therefore, the total number of combinations following the restriction is 3×3^3 .

Therefore, $|A| = 3^3 + 3^4$ and $P(A) = \frac{3^3 + 3^4}{6^3}$. \square

1.4 Axioms of Probability

Probability: The probability of an event E , denoted by $P(E)$, is a function that satisfies the three basic axioms:

Axiom 1:

$$0 \leq P(E) \leq 1$$

Axiom 2:

$$P(S) = 1$$

Axiom 3: For any sequence of mutually exclusive events A_1, A_2, \dots ,

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Fact 1.12

$$P(\emptyset) = 0$$

Fact 1.13 For a finite collection of mutually exclusive events A_1, \dots, A_n ,

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{i=1}^n P(A_i)$$

Probability of the complement

$$P(A^c) = 1 - P(A)$$

Proof.

$$A \cap A^c = \emptyset$$

Therefore, A and A^c are mutually exclusive. Therefore,

$$\begin{aligned} P(A) + P(A^c) &= P(A \cup A^c) \\ &= P(S) \\ &= 1 \\ \therefore P(A^c) &= 1 - P(A) \end{aligned}$$

■

Example 30. — Three traffic signals A lady crosses three traffic signals, with red and green lights only, on the way to her dog's hairdresser. The probabilities of encountering 0, 1, and 2 red lights are 0.4, 0.1, 0.2 respectively.

Find the probabilities of

- Encountering at least one red light.
- Encountering at least one green light.
- Encountering an odd number of red lights.

Solution:

- Encountering at least one red light.

$$\begin{aligned} P(1 \text{ red}) + P(2 \text{ red}) + P(3 \text{ red}) &= 1 - P(0 \text{ red}) \\ &= 1 - 0.4 \\ &= 0.6 \end{aligned}$$

- Encountering at least one green light.

$$\begin{aligned} P(1 \text{ green}) + P(2 \text{ green}) + P(3 \text{ green}) &= P(0 \text{ red}) + P(1 \text{ red}) + P(2 \text{ red}) \\ &= 0.4 + 0.1 + 0.2 \\ &= 0.7 \end{aligned}$$

- Encountering an odd number of red lights.

$$\begin{aligned} P(1 \text{ red}) + P(3 \text{ red}) &= 1 - P(0 \text{ red}) - P(2 \text{ red}) \\ &= 1 - 0.4 - 0.2 \\ &= 0.4 \end{aligned}$$

□

1.5 Addition rule

Suppose that you have two finite sets A and B . We can find the size of their union using

$$|A \cup B| = |A| + |B| - |A \cap B|$$

because when you work out $|A| + |B|$ the elements of $A \cap B$ are being 'counted twice'. You compensate for this by subtracting $|A \cap B|$.

Fact 1.15 — **Addition rule (aka inclusion-exclusion formula).**

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Example 31. Let A be the event “even number” and B the event “number > 3 ” when a fair dice is thrown. $P(A \cup B)$?

Solution: $P(A) = P(B) = \frac{3}{6}$ and $P(A \cap B) = \frac{2}{6}$. Hence $P(A \cup B) = \frac{3}{6} + \frac{3}{6} - \frac{2}{6} = \frac{4}{6}$, i.e. $P(\{2, 4, 5, 6\}) = \frac{2}{3}$. \square

Example 32. Ben is going to celebrate the beginning of the year of the dragon. He lives close to two pubs. The probability that he would go to pub A is 0.5 and the probability that the would go to pub B is 0.4. In addition, the probability that he would go to at least one of the two venues is 0.8.

1. What is the sample space (in terms of A and B)?
2. What is the probability that he would go to both pubs?
3. What is the probability that he would go to exactly one pub?

Solution:

1. Let A be the event that he would go to pub A, and let B be the event that he goes to pub B. Therefore,

$$S = \{A \cap B^c, A^c \cap B, A \cap B, A^c \cap B^c\}$$

2. The probability that he would go to both pubs is

$$\begin{aligned} P(A \cap B) &= P(A) + P(B) - P(A \cup B) \\ &= 0.5 + 0.4 - 0.8 \\ &= 0.1 \end{aligned}$$

3. The probability that he would go to exactly one pub is

$$\begin{aligned} P(A \cup B) - P(A \cap B) &= 0.8 - 0.1 \\ &= 0.7 \end{aligned}$$

\square

Example 33. — **A fair coin is flipped 3 times** . What is the probability of obtaining heads on the first flip OR the third flip?

Solution: the sample space is given by

$$\begin{array}{l} H \ H \ H \\ H \ H \ T \\ H \ T \ H \\ H \ T \ T \\ T \ H \ H \\ T \ H \ T \\ T \ T \ H \\ T \ T \ T \end{array}$$

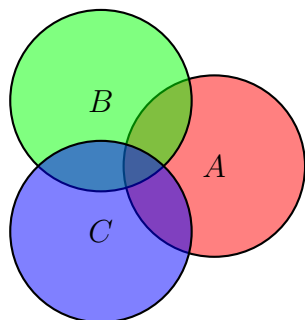
Let H_1 and H_3 denote the events that the first flip results in heads and the third flip results in heads, respectively. By the inclusion-exclusion formula, we have

$$\begin{aligned} P(H_1 \cup H_3) &= P(H_1) + P(H_3) - P(H_1 \cap H_3) \\ &= \frac{1}{2} + \frac{1}{2} - \frac{1}{4} \\ &= \frac{3}{4} \end{aligned}$$

□

Fact 1.16 — Addition rule for 3 events.

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - (P(A \cap B) + P(A \cap C) + P(B \cap C)) + P(A \cap B \cap C)$$



Fact 1.17 — Addition rule for n events.

$$\begin{aligned} P\left(\bigcup_{i=1}^n A_i\right) &= \sum_{i=1}^n P(A_i) - \sum P(\text{all 2-event intersections}) \\ &\quad + \sum P(\text{all 3-event intersections}) \\ &\quad - \sum P(\text{all 4-event intersections}) \\ &\quad \dots \end{aligned}$$

Fact 1.18 — Addition rule for n events using DeMorgan's law.

$$P\left(\bigcup_{i=1}^n A_i\right) = 1 - P\left(\bigcap_{i=1}^n A_i^c\right)$$

For example,

$$P(A \cup B \cup C) = 1 - P(A^c \cap B^c \cap C^c)$$

Example 34. — The weatherman said that:

$$P(\text{rain Mon}) = 30\%, \quad P(\text{rain Tue}) = 40\%, \quad P(\text{rain Wed}) = 50\%.$$

From experience, we know that

$$P(\text{rain 2 days in a row}) = 20\%, \quad P(\text{rain 3 days in a row}) = 10\%,$$

$$P(\text{rain Mon, Wed}) = 5\%.$$

Show that there is an 85% percent chance of rain anytime from Monday to Wednesday.

Solution: Let A, B and C be the events that it rains Monday, Tuesday and Wednesday, respectively. Then,

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) - (P(A \cap B) + P(A \cap C) + P(B \cap C)) + \\ &\quad + P(A \cap B \cap C) \\ &= .3 + .4 + .5 - (.2 + .05 + .2) + .1 = .85 \end{aligned}$$

□

Example 35. — **3 dice are rolled.** What is the probability that (at least) one of the dice results in 4?

Solution:[1] Let $F_i, i \in \{1, 2, 3\}$ be the event that the i th dice results in a 4. We are interested in $P(F_1 \cup F_2 \cup F_3)$. By the inclusion-exclusion formula we have

$$P(F_1 \cup F_2 \cup F_3) = P(F_1) + P(F_2) + P(F_3) - P(F_1 \cap F_2) - P(F_1 \cap F_3) - P(F_2 \cap F_3) + P(F_1 \cap F_2 \cap F_3)$$

Using combinatorics (\rightarrow see spreadsheet):

$$\begin{aligned} |F_i| &= 6^2, i = 1, 2, 3 \\ |F_i \cap F_j| &= 6, j \neq i = 1, 2, 3 \\ |F_1 \cap F_2 \cap F_3| &= 1 \end{aligned}$$

and thus

$$\begin{aligned} |F_1 \cup F_2 \cup F_3| &= |F_1| + |F_2| + |F_3| - |F_1 \cap F_2| - |F_1 \cap F_3| - |F_2 \cap F_3| + |F_1 \cap F_2 \cap F_3| \\ &= 6^2 + 6^2 + 6^2 - 6 - 6 - 6 + 1 = 91 \end{aligned}$$

Since $|S| = 6^3$, the requested probability is $91/216$.

□

Solution:[2, using De Morgan]

$$P(F_1 \cup F_2 \cup F_3) = 1 - P(F_1^c \cap F_2^c \cap F_3^c)$$

Using combinatorics, $|F_1^c \cap F_2^c \cap F_3^c| = 5^3$ because none of the three dice should land on 4. Therefore, $P(F_1^c \cap F_2^c \cap F_3^c) = 5^3/6^3$ and the requested probability is $1 - 5^3/6^3 = 91/216$.

□

1.6 Conditional Probability

The conditional probability of an event A is the probability that the event will occur given the knowledge that an event B has already occurred. This probability is written $P(A|B)$, the probability of B given A .

For example the probability of 7 when rolling two die is $1/6 (= 6/36)$ because the sample space consists of 36 equi-probable elementary outcomes of which 6 are favorable to the event of getting 7 as the sum of two die. Denote this event A: $P(A) = 1/6$. Consider another event B which is having at least one 2. There are still 36 elementary outcomes of which 11 are favorable to B; therefore, $P(B) = 11/36$. Question is what happens to the probability of A under the assumption that B took place?

The assumption that B took place reduces the set of possible outcomes to 11. Of these, only two - 2+5 and 5+2 - are favorable to A. Since this is reasonable to assume that the 11 elementary outcomes are equi-probable, the probability of A under the assumption that B took place equals $2/11$. This probability is denoted $P(A|B)$ - the probability of A assuming B: $P(A|B) = 2/11$.

For two events A and B in sample space S , where $P(B) > 0$, the conditional probability, i.e. the probability that A will occur after B has already occurred is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Key idea: we divide by $P(B)$ to account for the new sample space, B .

Conditional Probabilities are just like the ordinary probabilities, only on a new sample space. Thus, they satisfy all the formulas we already know, e.g.:

- $P(A \cup B|C) = P(A|C) + P(B|C) - P(A \cap B|C)$
- $P(A^c|B) = 1 - P(A|B)$
- $P(F_1 \cup F_2 \cup F_3|C) = 1 - P(F_1^c \cap F_2^c \cap F_3^c|C)$

Example 36. A dice is rolled once. Consider the following events.

A : The result is even.

B : The result is greater than 3.

What is the probability that the result is even, if it is known that result is greater than 3?

Solution:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(\{2, 4, 6\} \cap \{4, 5, 6\})}{P(\{4, 5, 6\})} \\ &= \frac{P(\{4, 6\})}{P(\{4, 5, 6\})} \\ &= \frac{1/3}{1/2} \\ &= \frac{2}{3} \end{aligned}$$

□

Example 37. A coin is flipped twice. What is the probability of getting ‘Heads’ on both flips, given that the first flip results in ‘Heads’.

Solution:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Let A be the event of getting two ‘Heads’. Therefore,

$$A = \{(H, H)\}$$

Let B be the event that the first flip results in ‘Heads’. Therefore,

$$B = \{(H, T), (H, H)\}$$

Therefore,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{1/4}{1/2} \\ &= \frac{1}{2} \end{aligned}$$

□

Example 38. A coin is flipped twice. What is the probability of ‘Heads’ on both flips, given that at least one flip results in ‘Heads’.

Solution:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

Let A be the event of getting two ‘Heads’. Therefore,

$$A = \{(H, H)\}$$

Let B be the event that at least one flip results in ‘Heads’. Therefore,

$$B = \{(H, T), (T, H), (H, H)\}$$

Therefore,

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{1/4}{3/4} \\ &= \frac{1}{3} \end{aligned}$$

□

Example 39. The probability that a new car battery functions for over 10,000 miles is 0.8, the probability that it functions for over 20,000 miles is 0.4, and the probability that it functions for over 30,000 miles is 0.1. If a new car battery is still working after 10,000 miles, what is the probability that

- its total life will exceed 20,000 miles,
- its additional life will exceed 20,000 miles?

Consider the following events to answer the questions:

L_{10} : event that the battery lasts for more than 10K miles.

L_{20} : event that the battery lasts for more than 20K miles.

L_{30} : event that the battery lasts for more than 30K miles.

Solution:

We know that $P(L_{10}) = 0.8$, $P(L_{20}) = 0.4$ and $P(L_{30}) = 0.1$. We are interested in calculating $P(L_{20}|L_{10})$ and $P(L_{30}|L_{10})$.

$$\begin{aligned} P(L_{20}|L_{10}) &= \frac{P(L_{20} \cap L_{10})}{P(L_{10})} \\ &= \frac{P(L_{20})}{P(L_{10})} \\ &= \frac{0.4}{0.8} \\ &= \frac{1}{2} \end{aligned}$$

By doing similar calculations it is easy to verify that $P(L_{30}|L_{10}) = \frac{1}{8}$. □

Fact 1.19 — Multiplication Rule. For two events:

$$\begin{aligned} P(A_1 \cap A_2) &= P(A_2)P(A_1|A_2) \\ &= P(A_1)P(A_2|A_1) \end{aligned}$$

For 3 events there are 3! possibilities:

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2) \\ &= P(A_2)P(A_3|A_2)P(A_1|A_2 \cap A_3) \\ &= \dots \end{aligned}$$

For 4 events we have 4! alternative formulas:

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3 \cap A_4) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_4|A_1 \cap A_2 \cap A_3) \\ &= P(A_4)P(A_2|A_4)P(A_3|A_4 \cap A_2)P(A_1|A_4 \cap A_2 \cap A_3) \\ &= \dots \end{aligned}$$

Example 40. An urn initially contains 5 white balls and 7 black balls. Each time a ball is selected, its color is noted and it is **replaced in the urn along with two other balls of the same color**. Compute the probability that the first two balls selected are black and the next two white.

Consider the following events to answer the question:

B_1 : event that the first ball chosen is black.

B_2 : event that the second ball chosen is black.

W_3 : event that the third ball chosen is white.

W_4 : event that the fourth ball chosen is white.

Solution: We are interested in calculating $P(B_1 \cap B_2 \cap W_3 \cap W_4)$. Using the Multiplication rule we get,

$$\begin{aligned} P(B_1 \cap B_2 \cap W_3 \cap W_4) &= P(B_1) \cdot P(B_2|B_1) \cdot P(W_3|B_1 \cap B_2) \cdot P(W_4|B_1 \cap B_2 \cap W_3) \\ &= \frac{7}{12} \times \frac{9}{14} \times \frac{5}{16} \times \frac{7}{18} \\ &= \frac{35}{768} \end{aligned}$$

□

Example 41. A deck of cards is randomly divided into four stacks of 13 cards each. Find the probability that each stack has exactly one ace. Hint:

Let A_1 be the event that $\mathbf{A}\spadesuit$ is in any one of the stacks.

Let A_2 be the event that $\mathbf{A}\spadesuit, \mathbf{A}\heartsuit$ are in different stacks.

Let A_3 be the event that $\mathbf{A}\spadesuit, \mathbf{A}\heartsuit, \mathbf{A}\diamonds$ are in different stacks.

Let A_4 be the event that $\mathbf{A}\spadesuit, \mathbf{A}\heartsuit, \mathbf{A}\diamonds, \mathbf{A}\clubsuit$ are in different stacks.

Solution: [using conditional probability.]

Therefore,

$$\begin{aligned} P(A_1 \cap A_2 \cap A_3 \cap A_4) &= P(A_1)P(A_2|A_1)P(A_3|A_1 \cap A_2)P(A_4|A_1 \cap A_2 \cap A_3) \\ &= 1 \times \frac{39}{51} \times \frac{26}{50} \times \frac{13}{49} = 0.105 \end{aligned}$$

□

Solution: [using combinatorics.]

$$|S| = \binom{52}{13} \binom{39}{13} \binom{26}{13} \binom{13}{13}$$

Let A be the event that each stack has exactly one ace. Therefore, each stack has one ace, and 12 non-ace cards. Therefore,

$$|A| = \binom{\binom{4}{1} \binom{48}{12}}{\binom{3}{1} \binom{36}{12}} \binom{\binom{2}{1} \binom{24}{12}}{\binom{1}{1} \binom{12}{12}}$$

Therefore,

$$\begin{aligned}
 P(A) &= \frac{|A|}{|S|} \\
 &= \frac{\binom{4}{1}\binom{48}{12} + \binom{3}{1}\binom{36}{12} + \binom{2}{1}\binom{24}{12} + \binom{1}{1}\binom{12}{12}}{\binom{52}{13}\binom{39}{13}\binom{26}{13}\binom{13}{13}} = 0.105
 \end{aligned}$$

□

Example 42. (Adopted from Meyer) A 52-card deck is thoroughly shuffled and you are dealt a hand of 13 cards.

1. If you have (at least) one ace, what is the probability that you have a second ace?
2. If you have (at least) the ace of spades, what is the probability that you have a second ace?

Hint: Define the events:

- A1: exactly one ace
 A2: exactly two aces
 A3: exactly three aces
 A4: exactly four aces

Solution: Define the events:

A1: exactly one ace, A2: exactly two aces, A3: exactly three aces, A4: exactly four aces.

Note that these events are disjoint (= Mutually exclusive). Remember that if A and B are two disjoint events, $P(A \cup B) = P(A) + P(B)$.

a) The problem is asking for the probability: $P(A2 \cup A3 \cup A4 | A1 \cup A2 \cup A3 \cup A4)$

Compute the probabilities of each of the four events:

$$P(A1) = \frac{\binom{4}{1}\binom{48}{12}}{\binom{52}{13}} = 0.438 \quad P(A2) = \frac{\binom{4}{2}\binom{48}{11}}{\binom{52}{13}} = 0.213 \quad P(A3) = \frac{\binom{4}{3}\binom{48}{10}}{\binom{52}{13}} = 0.0412 \quad P(A4) = \frac{\binom{4}{4}\binom{48}{9}}{\binom{52}{13}} = 0.00264$$

From the definition of conditional probability:

$$\begin{aligned}
 P(A2 \cup A3 \cup A4 | A1 \cup A2 \cup A3 \cup A4) &= \frac{P((A2 \cup A3 \cup A4) \cap (A1 \cup A2 \cup A3 \cup A4))}{P(A1 \cup A2 \cup A3 \cup A4)} \\
 &= \frac{P(A2 \cup A3 \cup A4)}{P(A1 \cup A2 \cup A3 \cup A4)} \\
 &= \frac{P(A2) + P(A3) + P(A4)}{P(A1) + P(A2) + P(A3) + P(A4)} \\
 &\approx \frac{0.257}{0.7} = 0.37
 \end{aligned}$$

b) Remarkably, the answer is different from part (a).

$$P(A1) = \frac{\binom{1}{1}\binom{48}{12}}{\binom{52}{13}} = 0.109 \quad P(A2) = \frac{\binom{1}{1}\binom{3}{1}\binom{48}{11}}{\binom{52}{13}} = 0.106 \quad P(A3) = \frac{\binom{1}{1}\binom{3}{2}\binom{48}{10}}{\binom{52}{13}} = 0.0309 \quad P(A4) =$$

$$\frac{\binom{1}{1}\binom{3}{3}\binom{48}{9}}{\binom{52}{13}} = 0.00264$$

$$P(A_2 \cup A_3 \cup A_4 | A_1 \cup A_2 \cup A_3 \cup A_4) \approx \frac{0.139}{0.249} = 0.56$$

□

Example 43. What is the probability that when a deck of cards is dealt in a game of bridge (each player gets 13 cards), the \heartsuit s will be dealt such that Alice gets 3, Bob gets 4, Charlie gets 2, David gets 4.

Solution: Let E_{Alice} be the event of Alice getting 3 \heartsuit s.

Let E_{Bob} be the event of Bob getting 4 \heartsuit s.

Let E_{Charlie} be the event of Charlie getting 2 \heartsuit s.

Let E_{David} be the event of David getting 4 \heartsuit s.

Therefore,

$$P(E_{\text{Alice}}) = \frac{\binom{13}{3}\binom{39}{10}}{\binom{52}{13}}$$

$$P(E_{\text{Bob}} | E_{\text{Alice}}) = \frac{\binom{10}{4}\binom{29}{9}}{\binom{39}{13}}$$

$$P(E_{\text{Charlie}} | E_{\text{Alice}} \cap E_{\text{Bob}}) = \frac{\binom{6}{2}\binom{20}{11}}{\binom{26}{13}}$$

$$P(E_{\text{David}} | E_{\text{Alice}} \cap E_{\text{Bob}} \cap E_{\text{Charlie}}) = \frac{\binom{4}{4}\binom{9}{9}}{\binom{13}{13}}$$

Therefore,

$$\begin{aligned} P(E_{\text{Alice}} \cap E_{\text{Bob}} \cap E_{\text{Charlie}} \cap E_{\text{David}}) &= P(E_{\text{Alice}}) \\ &\quad \times P(E_{\text{Bob}} | E_{\text{Alice}}) \\ &\quad \times P(E_{\text{Charlie}} | E_{\text{Alice}} \cap E_{\text{Bob}}) \\ &\quad \times P(E_{\text{David}} | E_{\text{Alice}} \cap E_{\text{Bob}} \cap E_{\text{Charlie}}) \\ &= \frac{\binom{13}{3}\binom{39}{10}}{\binom{52}{13}} \frac{\binom{10}{4}\binom{29}{9}}{\binom{39}{13}} \frac{\binom{6}{2}\binom{20}{11}}{\binom{26}{13}} \frac{\binom{4}{4}\binom{9}{9}}{\binom{13}{13}} \end{aligned}$$

□

1.7 Independent Events

Two events, A and B , are said to be independent if

$$P(A|B) = P(A)$$

→ Information about the occurrence of B does not affect the probability of A .

Fact 1.20 — The multiplication rule with independent events.

$$P(A \cap B) = P(A)P(B)$$

if and only if A and B are independent.

Fact 1.21 If A and B are independent, then so are A^c and B , A and B^c , A^c and B^c

Three events, A , B , and C , are said to be:

mutually independent if

$$P(A \cap B \cap C) = P(A)P(B)P(C)$$

and *pairwise* independent if

$$P(A \cap B) = P(A)P(B)$$

$$P(B \cap C) = P(B)P(C)$$

$$P(C \cap A) = P(C)P(A)$$



Pairwise independence does not imply mutual independence!

Example 44. two fair coins are tossed. Let

A : the first coin is H

B : the second coin is H

C : both coins match

a) are they pairwise independence?

b) are they mutually independent?

Solution:

$$S = \{(H, H), (H, T), (T, H), (T, T)\}$$

$$A = \{(H, H), (H, T)\} \quad B = \{(H, H), (T, H)\} \quad C = \{(H, H), (T, T)\}$$

$$A \cap B = \{(H, H)\} \quad A \cap C = \{(H, H)\} \quad B \cap C = \{(H, H)\}$$

$$A \cap B \cap C = \{(H, H)\}$$

a) they are pairwise independent because

$$P(A \cap B) = P(A)P(B) = 1/4$$

$$P(B \cap C) = P(B)P(C) = 1/4$$

$$P(C \cap A) = P(C)P(A) = 1/4$$

b) they are not mutually independent because

$$P(A \cap B \cap C) = 1/4 \neq P(A)P(B)P(C) = (1/2)^3$$

This makes sense: knowing that B and C occurred tells us that A also did.

□

Example 45. — **3 dice are rolled**, what is the probability that one of the dice results in 4?

Solution:[1] Let $F_i, i \in \{1, 2, 3\}$ be the event that the i th dice results in a 4. We are interested in $P(F_1 \cup F_2 \cup F_3)$. By the inclusion-exclusion formula we have $P(F_1 \cup F_2 \cup F_3) = P(F_1) + P(F_2) + P(F_3) - P(F_1 \cap F_2) - P(F_1 \cap F_3) - P(F_2 \cap F_3) + P(F_1 \cap F_2 \cap F_3)$. Since the events F_1, F_2, F_3 are mutually independent we can rewrite the above expression as

$$\begin{aligned} P(F_1 \cup F_2 \cup F_3) &= P(F_1) + P(F_2) + P(F_3) - P(F_1)P(F_2) - P(F_1)P(F_3) - P(F_2)P(F_3) \\ &\quad + P(F_1)P(F_2)P(F_3) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} - \left(\frac{1}{6} \times \frac{1}{6}\right) - \left(\frac{1}{6} \times \frac{1}{6}\right) - \left(\frac{1}{6} \times \frac{1}{6}\right) + \left(\frac{1}{6} \times \frac{1}{6} \times \frac{1}{6}\right) \\ &= \frac{91}{216} \end{aligned}$$

□

Solution:[2, using De Morgan]

$$P(F_1 \cup F_2 \cup F_3) = 1 - P(F_1^c \cap F_2^c \cap F_3^c) = 1 - (5/6)^3 = \frac{91}{216}$$

□

Example 46. A biased coin with $p =$ probability of coming up H, is tossed n times. What is the probability of having at least one H ?

Solution: Let $A =$ having at least one H out of the n tosses. Instead of enumerating all possible events containing at least one H and then compute the union of all those events, it is easier to note that $A^c =$ having all tosses come up T. Since $P(T) = 1 - p$ and all the tosses are independent, $P(A^c) = (1 - p)^n$ and the desired probability is

$$P(A) = 1 - (1 - p)^n \tag{1.4}$$

STOP! Note that combinatorics is not useful here because the coin is biased, which means that all outcomes in the relevant sample space are not equally likely.

If the coin were fair, i.e. $p = 1/2$, we could use combinatorics and get $|S| = 2^n, |A^c| = 1$ and $P(A^c) = 1/2^n$, which matches the more general results above that $P(A^c) = (1 - p)^n = (1 - 1/2)^n$.

□

Example 47. Suppose that A, B are mutually exclusive and $P(A) > 0$ and $P(B) > 0$. Are they independent?

Solution: No! Since $P(A \cap B) = 0$ for mutually exclusive events, knowing that one occurred means that the other cannot. Mathematically, they do not satisfy the condition for the independence

$$P(A \cap B) = P(A)P(B)$$

in this case (where $P(A) > 0$ and $P(B) > 0$).

□

Example 48. Suppose that $A \subset B$ and $P(A) > 0$ and $P(B) > 0$. Are two events A and B independent?

Solution: Since $A \subset B$,

$$P(A \cap B) = P(A)$$

The condition for the independence is

$$P(A \cap B) = P(A)P(B)$$

Hence, if $P(B) = 1$, A and B are independent but if $P(B) < 1$, A and B are not independent. \square

Example 49. If A and B are independent events, with $P(A) = \frac{1}{3}$ and $P(B) = \frac{1}{4}$, find the following:

- (a) $P(A^c \cap B^c)$
 (b) $P(A^c|B)$.

Solution: a) Since A and B are independent, A^c and B^c are independent. So $P(A^c \cap B^c) = P(A^c)P(B^c) = (1 - P(A))(1 - P(B)) = (1 - \frac{1}{3})(1 - \frac{1}{4}) = \frac{1}{2}$.

b) Since A and B are independent, A^c and B are also independent. So $P(A^c|B) = P(A^c) = 1 - P(A) = \frac{2}{3}$. \square

Example 50. Two fair dice are rolled.

Let A be the event that the sum of the results of the dice is 6.

Let B be the event that the result of the first dice is 4.

Let C be the event that the sum of the results of the dice is 7.

Which of the possible pairs of the events are independent?

Solution:

$$\begin{aligned} P(A) &= \frac{5}{36} & P(B) &= \frac{1}{6} & P(C) &= \frac{6}{36} \\ P(A \cap B) &= \frac{1}{36} & P(B \cap C) &= \frac{1}{36} & P(A \cap C) &= 0 \end{aligned}$$

Therefore,

$$P(A \cap B) \neq P(A)P(B)$$

$$P(B \cap C) = P(B)P(C)$$

$$P(C \cap A) \neq P(C)P(A)$$

Therefore, only B and C are independent. Why? Think about the meaning of $P(B|A)$ and $P(B|C)$.

\square

Example 51. Two cards are sequentially drawn (without replacement) from a well-shuffled deck of 52 cards. Let A be the event that the two cards drawn have the same value (e.g. both 4s)

and let B be the event that the first card drawn is an ace. Are these events independent?

Solution: To decide whether the two events are independent we need to check whether $P(A \cap B) = P(A)P(B)$.

$$\begin{aligned} P(A) &= \frac{52 \times 3}{52 \times 51} = \frac{1}{17} \\ P(B) &= \frac{4 \times 51}{52 \times 51} = \frac{1}{13} \\ P(A \cap B) &= \frac{4 \times 3}{52 \times 51} \\ &= \frac{1}{17} \times \frac{1}{13} \\ &= P(A)P(B) \end{aligned}$$

So yes, they are independent! This makes sense, all pairs have the same probability of being dealt. \square

1.8 Law of Total Probability

Fact 1.22 — Law of Total Probability . Given a division A_1, \dots, A_n of the sample space S , and an event B in S ,

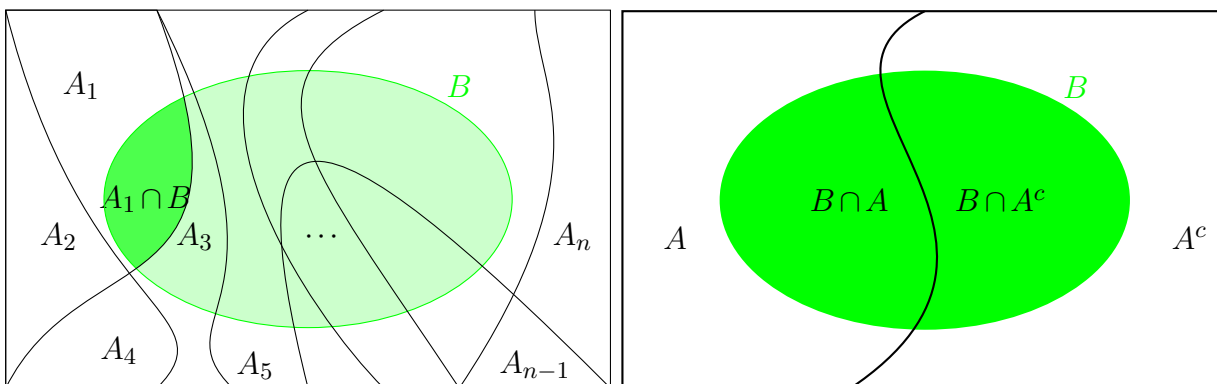
$$P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$$

For example, for $n = 3$:

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3).$$

For $n = 2$, we can say $A_1 = A$ and $A_2 = A^c$, so:

$$P(B) = P(B|A)P(A) + P(B|A^c)P(A^c). \quad (1.5)$$



Example 52. * A chocolate factory has three production lines. 50% of the production is milk chocolate, out of which 1% is defective. 30% of the production is dark chocolate, out of which 2% is defective.

20% of the production is white chocolate, out of which 0.5% is defective.

If a chocolate bar is picked randomly, what is the probability that it is defective?

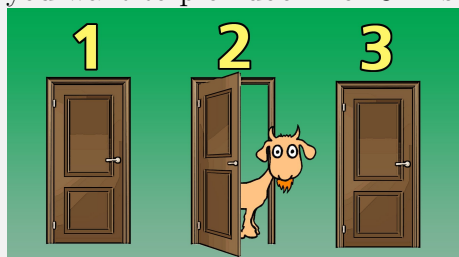
Hint: Let A_1 , A_2 , A_3 be the events that selected chocolate bar is made of milk, dark, white chocolate, respectively. Let B be the event that the selected chocolate bar is defective.

Solution: Therefore,

$$\begin{aligned} P(B) &= \sum_{i=1}^3 P(A_i)P(B|A_i) \\ &= ((0.5)(0.01)) + ((0.3)(0.02)) + ((0.2)(0.005)) \\ &= 0.005 + 0.006 + 0.001 \\ &= 0.012 \end{aligned}$$

□

Example 53. — Monty Hall Problem. You're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 2, which has a goat. He then says to you, "Do you want to pick door No. 3?" Is it to your advantage to switch your choice?



Hint: use the total probability rule with the events:

W: win by switching doors

E: car behind original door (before Monty Hall opens door No. 2 above)

Solution: Let:

W: win by switching doors

E: car behind original door (before Monty Hall opens door No. 2 above)

$$\begin{aligned} P(W) &= P(W|E)P(E) + P(W|E^c)P(E^c) \\ &= (0)(1/3) + (1)(2/3) \\ &= 2/3 \end{aligned}$$

□

Example 54. — Tornadoes** A structure is located in a region where tornado wind force must be considered in its design. Suppose that from the records of tornadoes for the past 200 years, it is estimated that during any 5-year period the probability of having 0, 1 and 2 tornadoes is 0.5, 0.3 and 0.2, respectively. If a tornado occurs, the structure will be damaged with probability $p = 5\%$.

- a) if two tornadoes occurred last year, what is the probability that there was damage to the structure?

b) what is the probability the a structure will be damaged in the next five years?
Hint: For part a) let

A_i be the events of having $i = 0, 1, 2$ tornadoes last year, and
 D the event that a particular structure was damaged last year.

Solution:

a) If two tornadoes occurred last year, what is the probability that there was damage to the structure? Let:

A_i = having $i = 0, 1, 2$ tornadoes last year, and

D = a structure was damaged last year.

For any particular structure, it is easier to calculate the probability of no damage given the two tornadoes ($= (1 - p)^2$), so

$$\begin{aligned} P(D|A_2) &= 1 - P(D^c|A_2) \\ &= 1 - (1 - p)^2 = 0.0975 \end{aligned}$$

b) what is the probability that the structure will be damaged in the next five years? Let

A_i = having $i = 0, 1, 2$ tornadoes in the next five years, and

D = the structure will be damaged in the next five years.

Since we don't know the number of tornadoes that will occur, we use the total probability rule:

$$\begin{aligned} P(D^c) &= \sum_{i=0}^2 P(D^c|A_i)P(A_i) \\ &= \sum_{i=0}^2 (1 - p)^i P(A_i) \\ &= (1 - p)^0 P(A_0) + (1 - p)^1 P(A_1) + (1 - p)^2 P(A_2) \\ &= ((1)(0.5)) + ((0.95)(0.3)) + ((0.95^2)(0.2)) = 0.9655 \end{aligned}$$

and the answer is $1 - 0.9655 = 0.0345$.

□

Example 55. A student in Monty Hall's probability course misses his exam and must take a makeup. Before the test, Prof. Hall invites him to choose one of five envelopes, two containing easy makeup exams, three containing hard ones, and the student takes an envelope.

"Before you start, you might enjoy looking at one of my hard makeup exams – just full of nasty probability problems," says the professor. From among the four remaining envelopes, he selects one at random that he knows to contain a hard exam and opens it.

"Excuse me," says the student, "but the envelope I picked looks a bit smudged. Could I swap it for one of the others?" And he does.

What is the probability that the student has an easy exam after making the swap?

Solution: Let:

Event W : easy exam after swap

Event E : first choice is an easy exam

Event H : first choice is a hard exam

$$P(W) = P(W|E)P(E) + P(W|H)P(H) = (1/3)(2/5) + (2/3)(3/5) = 8/15$$

□

Example 56. A box contains w white balls, b black balls and r red balls. A ball is chosen at random and if it is either black or red then it is replaced by a white ball and if it is white then it is replaced by a red ball. Now again draw a ball.

- What is the probability that the second ball drawn is red when the first ball drawn is red ?
- What is the probability that the second ball drawn is white?

Solution: Let W_i, B_i, R_i be the event that the i -th draw is a white, black and red ball respectively.

a)

$$P(R_2|R_1) = \frac{r-1}{w+b+r}.$$

b)

$$\begin{aligned} P(W_2) &= P(W_2|W_1)P(W_1) + P(W_2|B_1)P(B_1) + P(W_2|R_1)P(R_1) \\ &= \frac{w-1}{w+b+r} \frac{w}{w+b+r} + \frac{w+1}{w+b+r} \frac{b}{w+b+r} + \frac{w+1}{w+b+r} \frac{r}{w+b+r}. \end{aligned}$$

□

1.9 Bayes' Theorem

The power of Bayes' rule is that in many situations where we want to compute $P(A|B)$ it turns out that it is difficult to do so directly, yet we might have direct information about $P(B|A)$. Bayes' rule enables us to compute $P(A|B)$ in terms of $P(B|A)$.

Bayes rule with two events

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} && \text{(conditional probability)} \\ &= \frac{P(B|A)P(A)}{P(B)} && \text{(Bayes rule v1)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} && \text{(Bayes rule v2)} \end{aligned}$$

Bayes rule with a division of S . If we have a division A_1, \dots, A_n of the sample space S , then by

the Law of Total Probability $P(B) = \sum_{i=1}^n P(B|A_i)P(A_i)$ and:

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} && \text{(conditional probability)} \\ &= \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^n P(B|A_j)P(A_j)} && \text{(Bayes rule v3)} \end{aligned}$$

Example 57. In a bolt factory, machines 1,2 and 3 respectively produce 20 %, 30% and 50% of the total output. Of their output, 5%, 3% and 2% are defective. A bolt is selected at random.

a) What is the probability that it is defective?

b) Given that it is defective, what is the probability that it was made by machine 1?

Hint: Let D be the event that the bolt is defective and M_1, M_2, M_3 be the events that the selected bolt comes from machines 1,2 and 3 respectively.

Solution: We have

$$P(M_1) = 0.2, P(M_2) = 0.3, P(M_3) = 0.5 \text{ and}$$

$$P(D|M_1) = 0.05, P(D|M_2) = 0.03, P(D|M_3) = 0.02.$$

a) From the law of total probability,

$$P(D) = P(D|M_1)P(M_1) + P(D|M_2)P(M_2) + P(D|M_3)P(M_3) = 0.029$$

b)

$$P(M_1|D) = \frac{P(D|M_1)P(M_1)}{P(D)} = 0.52$$

□

Example 58. 1/10 of men and 1/7 of women are color-blind. A person is chosen at random and that person is color-blind. What is the probability that the person is male. Assume males and females to be in equal numbers. Let M=male, F=female, C=color-blind.

Solution: Let M=male, F=female, C=color-blind. Then

$$\begin{aligned} P(M|C) &= \frac{P(M \cap C)}{P(C)} \\ &= \frac{P(C|M)P(M)}{P(C|M)P(M) + P(C|F)P(F)} \\ &= \frac{\frac{1}{10} \cdot \frac{1}{2}}{\frac{1}{10} \cdot \frac{1}{2} + \frac{1}{7} \cdot \frac{1}{2}}. \end{aligned}$$

□

Example 59. A transmitter sends binary bits, 80% 0's and 20% 1's. When a 0 is sent, the receiver will detect it correctly 80% of the time. When a 1 is sent, the receiver will detect it correctly 90% of the time.

(a) What is the probability that a 1 is sent and a 1 is received?

(b) If a 1 is received, what is the probability that a 1 was sent?

Solution: We will consider the following events.

S_0 : event that the transmitter sent a 0.

S_1 : event that the transmitter sent a 1.

R_1 : event that 1 was received.

(a) We are interested in finding $P(S_1 \cap R_1)$.

$$\begin{aligned} P(S_1 \cap R_1) &= P(R_1|S_1)P(S_1) \\ &= 0.2 \times 0.9 \\ &= 0.18 \end{aligned}$$

(b) We are interested in finding $P(S_1|R_1)$.

$$\begin{aligned} P(S_1|R_1) &= \frac{P(S_1 \cap R_1)}{P(R_1)} = \frac{P(S_1 \cap R_1)}{P(R_1 \cap S_1) + P(R_1 \cap S_0)} \\ &= \frac{P(S_1 \cap R_1)}{P(R_1|S_1)P(S_1) + P(R_1|S_0)P(S_0)} = \frac{0.18}{0.18 + P(R_1|S_0)P(S_0)} \\ &= \frac{0.18}{0.18 + 0.8 \times 0.2} = 0.5294 \end{aligned}$$

□

Example 60. An urn contains 5 white and 10 black balls. A fair die is rolled and that number of balls are chosen from the urn.

(a) What is the probability that all of the balls selected are white?

(b) What is the conditional probability that the die landed on 3 if all the balls selected are white?

Solution: We will consider the following events.

W : event that all of the balls chosen are white.

D_i : event that the die landed on i , $1 \leq i \leq 6$.

(a) We want to find $P(W)$. We can do this as follows.

$$\begin{aligned} P(W) &= \sum_{i=1}^6 P(W \cap D_i) \\ &= \sum_{i=1}^6 P(D_i)P(W|D_i) \\ &= \sum_{i=1}^6 \frac{1}{6} \frac{\binom{5}{i}}{\binom{15}{i}} \\ &= \frac{1}{6} \left(\frac{5}{10} + \frac{10}{105} + \frac{10}{455} + \frac{5}{1365} + \frac{1}{3003} \right) \\ &= 0.1035 \end{aligned}$$

(b) We want to find $P(D_3|W)$. This can be done as follows.

$$\begin{aligned}
 P(D_3|W) &= \frac{D_3 \cap W}{P(W)} \\
 &= \frac{P(D_3) \times P(W|D_3)}{P(W)} \\
 &= \frac{1/6 \times \binom{5}{3} \binom{15}{3}}{0.1035} \\
 &= \frac{1/6 \times 10/455}{0.1035} \\
 &= \frac{0.00366}{0.1035} \\
 &= 0.03539
 \end{aligned}$$

□

Example 61. In answering a question on a multiple choice test, a student either knows the answer or the student just guesses. Suppose that the probability that the student knows the answer is 0.75. Assuming that there are 5 choices for each multiple-choice question, what is the conditional probability that the student knew the answer to a question given that the student answered it correctly?

Hint: Let:

C = student answers the question correctly,
 K = student knows the answer.

Solution:

The probability that the student who guesses will be correct is $1/5 = 0.20 = P(C|K^c)$.

$$\begin{aligned}
 P(K|C) &= \frac{P(K \cap C)}{P(C)} \\
 &= \frac{P(C|K)P(K)}{P(C|K)P(K) + P(C|K^c)P(K^c)} \\
 &= \frac{1 \cdot 0.75}{1 \cdot 0.75 + 0.20 \cdot 0.25} \\
 &= 0.9375
 \end{aligned}$$

□

Example 62. Stores A, B and C have 50, 75 and 100 employees respectively. and 50%, 60% and 70% of the employees are women. Resignations are equally likely among all employees, regardless of sex. One employee resigns and is a woman. What is the probability that she works in store A?

Solution: Let W be the event that a woman employee resigns from anywhere, and let A , B and C denote the event that a randomly selected employee works at the respective store. Then $P(A) = 50/225$, $P(B) = 75/225$ and $P(C) = 100/225$. Likewise the probabilities of resignation

of a woman from a store is given by the information to be $P(W|A) = 0.50$, $P(W|B) = 0.60$, and $P(W|C) = 0.70$. Then we can use Bayes Theorem (re-deriving it in the process of using it):

$$\begin{aligned} P(A|W) &= \frac{P(A \cap W)}{P(W)} \\ &= \frac{P(W|C)P(C)}{P(W|A)P(A) + P(W|B)P(B) + P(W|C)P(C)} \\ &= \frac{(0.50)(50/225)}{(0.50)(50/225) + (0.60)(75/225) + (0.70)(100/225)} \\ &\approx 0.17857 \end{aligned}$$

□

Example 63. (Adopted from Meyer) Outside of their hum-drum duties as CS-20 Teaching Assistants, Nick is trying to learn to levitate using only intense concentration and Keenan is trying to become the world champion flaming torch juggler. Suppose that Nick's probability of success is $1/6$, Keenan's chance of success is $1/4$, and these two events are independent.

- If at least one of them succeeds, what is the probability that Nick learns to levitate?
- If at most one of them succeeds, what is the probability that Keenan becomes the world flaming torch juggler champion?
- If exactly one of them succeeds, what is the probability that it is Nick?

Solution: Define the events:

N: Nick succeeds K: Keenan succeeds L: at Least one succeeds

a)

$$\begin{aligned} P(N|L) &= \frac{P(L|N)P(N)}{P(L)} \\ &= \frac{1 \times \frac{1}{6}}{P(L)} \\ &= \frac{1 \times \frac{1}{6}}{P(NK^c) + P(N^cK) + P(NK)} \\ &= \frac{1 \times \frac{1}{6}}{\frac{3}{24} + \frac{5}{24} + \frac{1}{24}} \\ &= \frac{4}{9} \end{aligned}$$

- $\frac{5}{23}$
- $\frac{3}{8}$

□

1.9.1 Updating probability estimates

It is useful to use Bayes rule in terms of updating our belief about a hypothesis A in the **light of new evidence B**. We say that our **posterior** belief $P(A|B)$ is calculated by multiplying our **prior** belief $P(A)$ by the likelihood $P(B|A)$ that B will occur if A is true.

The formulas are the same, e.g.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

but now we interpret everything on the right-hand side of this equation with the information **prior** to knowing the new evidence B .

Example 64. — Criminal investigation (adapted from Ross). In a certain stage of a criminal investigation, the inspector in charge is 60% convinced of the guilt of a certain suspect. Suppose that a new piece of DNA evidence uncovered that the criminal has diabetes. This prompted the inspector to test the suspect for diabetes, and the test came out positive, indicating that the suspect does have diabetes. If 10% of the population has diabetes how certain should that inspector be of the guilt of the suspect?

Let G be the event that the suspect is guilty, and D the event that the suspect has diabetes.

Solution:

We have $P(G) = 0.6$, $P(D|G) = 1$ because we know that the criminals has diabetes, $P(D|G^c) = 0.1$ because **before the evidence came to light** this is our best guess. Therefore,

$$\begin{aligned} P(G|D) &= \frac{P(D|G)P(G)}{P(D)} \\ &= \frac{P(D|G)P(G)}{P(D|G)P(G) + P(D|G^c)P(G^c)} \\ &= \frac{(1)(0.6)}{(1)(0.6) + (0.1)(0.4)} = 0.9375 \end{aligned}$$

□

Example 65. — A medical tests of rare diseases* Medical tests for a certain condition are not perfect. Consider a test that, when performed on an affected person, it comes up positive 95% of the times and yields a “false negative” 5% of the times. When the test is performed on a healthy person the test comes up negative in 99% of the cases and yields a “false positive” in 1% of the cases. If 0.5% of the population actually have the condition, what is the probability that Alice has the condition given that her test came up positive?

We will consider the following events to answer the question.

A : event that Alice has the medical condition.

B : event that Alice tested positive.

Solution: We are interested in $P(A|B)$. From the definition of conditional probability and the total probability theorem we get

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{0.95 \times 0.005}{0.95 \times 0.005 + 0.01 \times 0.995} \\ &= 0.323 \end{aligned}$$

This result means that 32.3% of the people who are tested positive actually suffer from the condition!

□

Example 66. — Quality of concrete material* In order to ensure the quality of concrete material used in a reinforced concrete construction, concrete cylinders are collected at random from concrete mixes delivered to the construction site by a mixing plant. Past records of concrete from the same plant show that **80% of concrete mixes are good** or of satisfactory quality. To further ensure that the concrete delivered on site is of good quality, the engineer requires that one cylinder among those collected each day be tested for minimum compressive strength. The test method is not perfect—**its reliability is only 95%**, meaning the probability that a good-quality concrete cylinder will pass the test is 0.95, or that a poor-quality cylinder can pass the test is 0.05. (10 points)

(a) If a concrete cylinder passes the test, find the probability that it is a good-quality concrete delivered on site. (10 points)

(b) Now, suppose the engineer is not satisfied with just testing one cylinder, and requires that a second cylinder be tested. If the second cylinder tested also gave a positive result, find the probability that the concrete is of good quality.

(c) If the third cylinder tested didn't pass the test, find the probability that the concrete is of good quality. (10 points)

Solution: Answer: (a) 0.987 (b) 0.999 (c) 0.981

(a) Define the following events:

G = good quality concrete

T = a concrete cylinder passes the test According to the available information, we have:

$$\begin{aligned} P(G) &= 0.8 \\ P(T|G) &= 0.95 \\ P(T|G^c) &= 0.05 \end{aligned}$$

Then, if a concrete cylinder passes the test, the probability that it is a good-quality concrete delivered on site is updated as follows:

$$\begin{aligned} P(G) &= \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|G^c)P(G^c)} \\ &= \frac{0.95 \times 0.80}{0.95 \times 0.80 + 0.05 \times 0.20} \\ &= 0.987 \end{aligned}$$

(b) Here, the second cylinder is tested positive, so now:

$$\begin{aligned} P(G) &= 0.987 \\ P(G|T) &= \frac{P(T|G)P(G)}{P(T|G)P(G) + P(T|G^c)P(G^c)} \\ &= \frac{0.95 \times 0.987}{0.95 \times 0.987 + 0.05 \times (1 - 0.987)} \\ &= 0.999 \end{aligned}$$

(c) The third cylinder is tested negative, so

$$\begin{aligned} P(G) &= 0.999 \\ P(G|T^c) &= \frac{P(T^c|G)P(G)}{P(T^c|G)P(G) + P(T|G)P(G^c)} \\ &= \frac{0.05 \times 0.999}{0.05 \times 0.999 + 0.95 \times (1 - 0.999)} = 0.981 \end{aligned}$$

□

1.10 More Problems

Example 67. Suppose the travel time between two major cities A and B by air is 7 or 8 hr if the flight is nonstop; however, if there is one stop, the travel time would be 10, 11, or 12 hr. A nonstop flight between A and B would cost \$1000, whereas with one stop the cost is only \$650. Then, between cities B and C, all flights are nonstop requiring 2 or 3 hours at a cost of \$250. (There is no flight from A to C)

For a passenger wishing to travel from city A to city C,

- What is the possibility space or sample space of his travel times from A to B? From A to C?
- What is the sample space of his travel cost from A to B?
- If T =travel time from city A to city C, and S =cost of travel from A to C, what is the sample space of T and S ?

Solution: (a) Sample space of travel time from A to B = {7, 8, 10, 11, 12}

Sample space of travel time from A to C = {9, 10, 11, 12, 13, 14, 15}

(b) Sample space of travel cost from A to B = {650, 1000}

(c) Sample space of T = {9, 10, 11, 12, 13, 14, 15}

Sample space of S = {900, 1250}

Sample space of T and S = {{9, 1250}, {10, 1250}, {11, 1250}, {12, 900}, {13, 900}, {14, 900}, {15, 900}}

□

Example 68. Two construction companies a and b are bidding for projects. Define A as the event that Company a wins a bid, and B as the event that Company b wins a bid.

Sketch the Venn diagrams for the sample spaces and their subsets of the following:

- Company a is submitting a bid for one project, and Company b is submitting its own bid for another project. (In this case, it is possible for both companies to win their respective bids)
- Companies a and b are submitting bids for the same project, and there are also other bidders for the project.
- Companies a and b are the only companies submitting competing bids for a single project.

Solution: From Ang and Tang textbook Example 2.9 and Example 2.10 on Page 38-39. In class.

□

Example 69. In a process that manufactures aluminum cans, the probability that a can has a flaw on its side is 0.03, the probability that a can has a flaw on its top is 0.05, and the probability that a can has a flaw on both the side and the top is 0.01. What is the probability that it has no flaw?

Solution: $P(\text{on side or on top}) = P(\text{on side}) + P(\text{on top}) - P(\text{on side and on top}) = 0.07$
 $P(\text{no flaw}) = 1 - P(\text{on side or on top}) = 0.93$

□

Example 70. (a) A die is rolled once. What is the probability that the result is even, if it is known that the result is higher than 3?

(b) Two dice are rolled once. what is the sample space of this experiment? what is the probability that the sum is greater than 7?

Solution:

Answer: (a) $2/3$ (b) $15/36$ (0.417)

A: Result is even

B: Result is higher than 3

C: Sum is greater than 7

$$(a) P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(\{2,4,6\} \cap \{4,5,6\})}{P(\{4,5,6\})} = \frac{P(\{4,6\})}{P(\{4,5,6\})} = \frac{\frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

(b) Sample space = $\{\{1, 1\}, \{1, 2\}, \{1, 3\}, \{1, 4\}, \{1, 5\}, \{1, 6\}, \{2, 1\}, \{2, 2\}, \{2, 3\}, \{2, 4\}, \{2, 5\}, \{2, 6\}, \{3, 1\}, \{3, 2\}, \{3, 3\}, \{3, 4\}, \{3, 5\}, \{3, 6\}, \{4, 1\}, \{4, 2\}, \{4, 3\}, \{4, 4\}, \{4, 5\}, \{4, 6\}, \{5, 1\}, \{5, 2\}, \{5, 3\}, \{5, 4\}, \{5, 5\}, \{5, 6\}, \{6, 1\}, \{6, 2\}, \{6, 3\}, \{6, 4\}, \{6, 5\}, \{6, 6\}\}$

Sample space of the sum = $\{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$

$$P(C) = \frac{15}{36} = 0.417$$

□

Example 71. A team of two engineers, A and B, was assigned to check a set of computations. The two work simultaneously but separately and independently. The probability of engineer A spotting a given error is 0.7, whereas that for B is 0.8.

(a) Suppose there is only one error in the computation. What is the probability that this error will be spotted by this team?

(b) If the error in part (a) was identified, what is the probability that it was discovered by A alone?

Solution:

Answer: (a) 0.94 (b) 0.149

A: A spots the error

B: B spots the error

C: Error spotted

$$(a) P(C) = 1 - P(C^c) = 1 - P(A^c)P(B^c) = 1 - (1 - 0.7) \times (1 - 0.8) = 0.94$$

$$(b) P(AB^c|C) = \frac{P(C|AB^c)}{P(C|AB^c) \cdot P(AB^c) + P(C|A^cB) \cdot P(A^cB) + P(C|AB) \cdot P(AB) + P(C|A^cB^c) \cdot P(A^cB^c)} = 0.149$$

□

Example 72. The proportion of people in a given community who have a certain disease is 0.01. A test is available to diagnose the disease. If a person has the disease, the probability that the test will produce a positive signal is 0.98. If a person does not have the disease, the probability that the test will produce a positive signal is 0.02. If a person tests positive, what is the probability that the person actually has the disease?

Solution:

Answer: 0.331

D: The person actually has the disease

+: The tests gives a positive signal

Using Bayes' rule:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|D^c)P(D^c)} = \frac{(0.98)(0.01)}{(0.98)(0.01)+(0.02)(1-0.01)} = 0.331$$

□

Example 73. — Tornadoes** 100 structures are located in a region where tornado wind force must be considered in its design. Suppose that from the records of tornadoes for the past 200 years, it is estimated that during any 5-year period the probability of having 0, 1 and 2 tornadoes is 0.5, 0.3 and 0.2, respectively. If a tornado occurs, a structure will be damaged with probability $p = 5\%$.

- if two tornadoes occurred last year, how many structures do you expect to have been damaged?
- what is the probability the a structure will be damaged in the next five years?
- how many structures do you expect to be damaged in the next five years?

Solution:

- If two tornadoes occurred last year, how many structures do you expect to have been damaged?

Let

A_i be the events of having $i = 0, 1, 2$ tornadoes last year, and

D the event that a particular structure was damaged last year.

For any particular structure, it is easier to calculate the probability of no damage given the two tornadoes ($= (1 - p)^2$), so

$$\begin{aligned} P(D|A_2) &= 1 - P(D^c|A_2) \\ &= 1 - (1 - p)^2 = 0.0975 \end{aligned}$$

and the answer would be $100P(D|A_2) = 9.75 \rightarrow 10$ structure.

- what is the probability the a structure will be damaged in the next five years? Let A_i be the events of having $i = 0, 1, 2$ tornadoes in the next five years, and

D the event that a structure will be damaged in the next five years.

Since we don't know the number of tornadoes that will occur, we use the total probability rule:

$$\begin{aligned} P(D^c) &= \sum_{i=0}^2 P(D^c|A_i)P(A_i) \\ &= \sum_{i=0}^2 (1 - p)^i P(A_i) \\ &= (1 - p)^0 P(A_0) + (1 - p)^1 P(A_1) + (1 - p)^2 P(A_2) \\ &= ((1)(0.5)) + ((0.95)(0.3)) + ((0.95^2)(0.2)) = 0.9655 \end{aligned}$$

and the answer is $1 - 0.9655 = 0.0345$.

- how many structures do you expect to be damaged in the next five years? $3.45 \rightarrow$ between 3 and 4 structures.

<https://eli.thegreenplace.net/2018/conditional-probability-and-bayes-theorem/>

□

2. Random Variables

Often in engineering or the natural sciences, outcomes of random experiments are numbers associated with some physical quantities. Such outcomes, called random variables, will be denoted by **capital letters**, e.g.

X = “time between Tech Trolleys at the CRC stop” or
 Y = “number of customers per day an Uber driver serves”,

and a particular realizations of a random variable by **lowercase letters**, e.g.

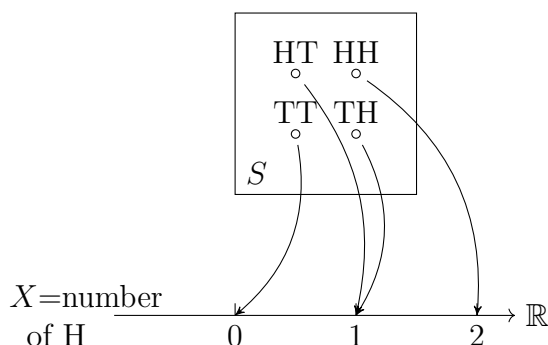
$x = 11.4$ min, or $y = 6$ customers.

Random variable: the results of an experiment expressed as a number. Mathematically, a function $X : S \rightarrow \mathbb{R}$ that maps points from the sample space to the real line is called a random variable.

There are two types of random variables:

A **discrete** random variable is a rv which takes a finite or countable number of values.

A **continuous** random variable is a rv which takes values in (an interval of) the real line.



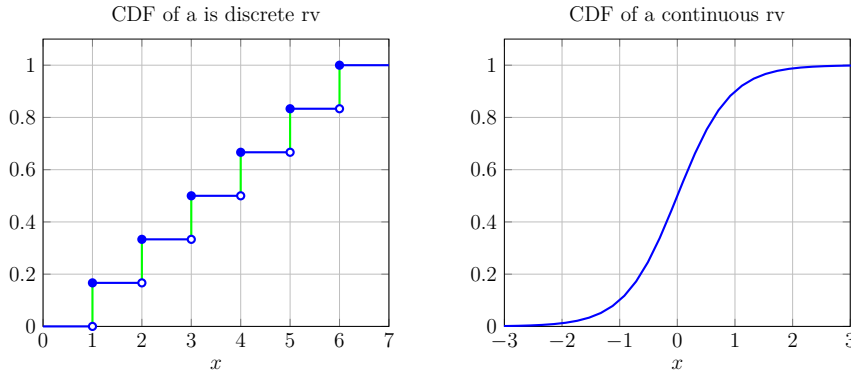
Events are statements of the type “ $X \leq x$ ”, “ $X > x$ ”, or “ $a < X < b$ ”. In general, an event is “ $X \in A$ ” where $A \subseteq \mathbb{R}$.

2.1 Probability distribution function

The function

$$F_X(x) = P(X \leq x), \quad x \in \mathbb{R},$$

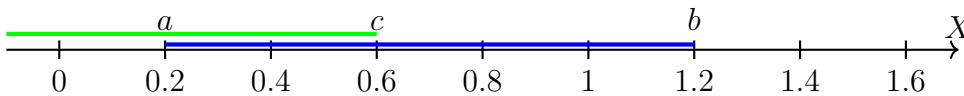
is called the **probability distribution, cumulative distribution function**, or **CDF** for short.



The probability of any statement about the X is computable when $F_X(x)$ is known, e.g.:

- The complement rule: $P(A^c) = 1 - P(A)$ becomes $P(X > a) = 1 - P(X \leq a) = 1 - F_X(a)$
- Probability of X falling on the interval $(a, b]$: $P(a < X \leq b) = F_X(b) - F_X(a)$
- Conditional probability: $P(A|B) = \frac{P(A \cap B)}{P(B)}$ becomes

$$\begin{aligned} P(a < X \leq b \mid X \leq c) &= \frac{P(a < X \leq b \cap X \leq c)}{P(X \leq c)} \\ &= \frac{P(a < X \leq c)}{P(X \leq c)} \quad (\text{assuming } a \leq c \leq b) \\ &= \frac{F_X(c) - F_X(a)}{F_X(c)} \end{aligned}$$



Fact 2.1 Let F_X be the cumulative distribution function of a random variable X . Then,

1. F_X is a non-decreasing function.
2. $F_X(\infty) = 1$.
3. $F_X(-\infty) = 0$.
4. F_X is right continuous, i.e., the function is equal to its right hand limit.

2.2 Quantiles (aka percentiles)

The **quantile** x_α , $0 \leq \alpha \leq 1$, for a random variable X is given by:

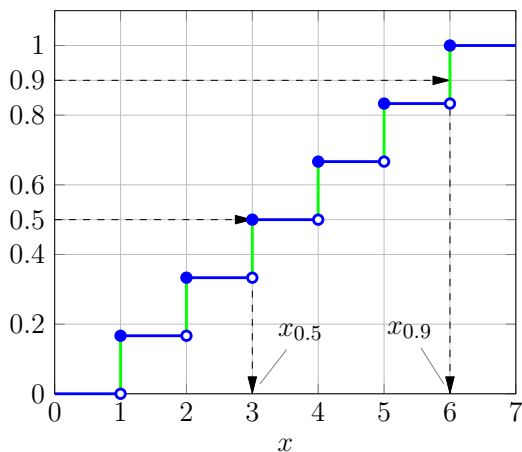
$$P(X \leq x_\alpha) = \alpha, \quad \rightarrow \quad x_\alpha = F_X^{-1}(\alpha).$$

The inverse function $F_X^{-1}(\alpha)$ is called the “quantile function”.

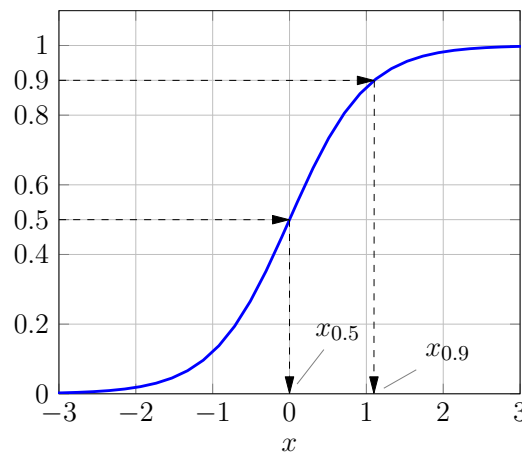
Important quantiles are:

- median (50th-percentile): $x_{0.5}$
- upper quartile (75th-percentile): $x_{0.75}$
- lower quartile (25th-percentile): $x_{0.25}$
- quintiles: $x_{0.2}$, $x_{0.4}$, $x_{0.6}$, $x_{0.8}$
- deciles: $x_{0.1}$, $x_{0.2}$, ...

CDF of a discrete rv



CDF of a continuous rv



Example 74. Derive the quantile function for the continuous CDF in the above figure, where:

$$F_X(x) = \frac{1}{e^{-2x} + 1}$$

Solution: Solving for x_α in $\alpha = \frac{1}{e^{-2x_\alpha} + 1}$ gives

$$x_\alpha = \frac{1}{2} \log \left(\frac{\alpha}{1 - \alpha} \right)$$

□

2.3 Discrete Random Variables

We saw that if X takes a finite or countable number of values it is called **discrete** random variables and the distribution function $F_X(x)$ is a “stair” looking function that is constant except the possible jumps. The size of a jump at point x is equal to the probability $P(X = x)$, denoted by $p_X(x)$, and called the **probability-mass function**.

Probability mass function (PMF). Gives the probability of a discrete random variable X having value x is called the probability mass function of X . It is denoted as

$$p_X(x) = P(X = x)$$

Observe that

$$F_X(a) = \sum_{x \leq a} p_X(x)$$

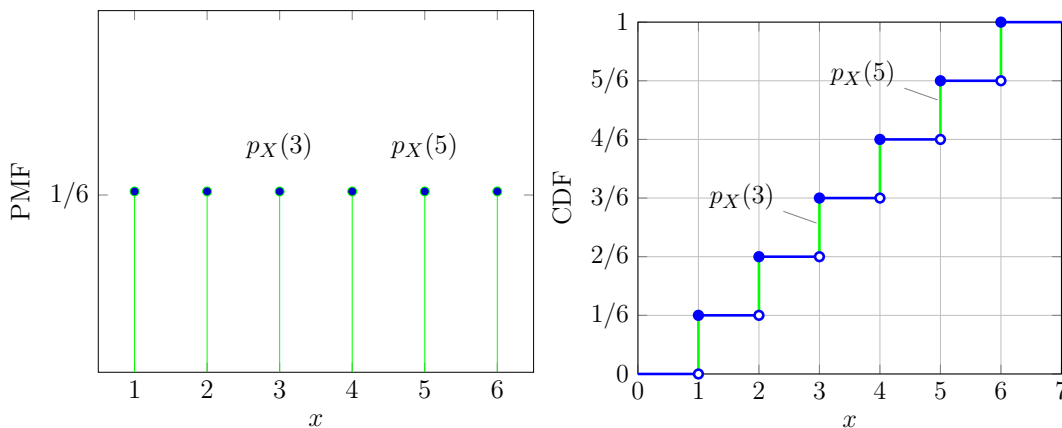
and

$$\sum_{x \in S} p_X(x) = 1$$

This is because the events $X = x$ are disjoint and hence form a *division* of the sample space S . These probabilities $p_X(x)$ are often referred to as *probability masses* since they represent discrete (point) masses of probability at specific locations along the real axis, just as point charges in electrostatics are used to represent discrete charges distributed along a line. These probabilities, written or plotted as a function of x , is called the *probability mass function (PMF)*.

Example 75. Let X be the result of a dice roll.

- Plot the CDF and PMF of X .
- Determine $P(X > 3)$
- Determine $P(2 < X \leq 5)$
- Determine $P(2 < X \leq 5 \mid X \leq 3)$
- Determine the median, upper quartile and lower quartile



STOP! This is called the discrete uniform distribution in $(1, 6)$.

- Plot the CDF and PMF of X .
- $P(X > 3) = 1/2$
- $P(2 < X \leq 5) = 1/2$
- $P(2 < X \leq 5 \mid X \leq 3) = 1/3$
- the median, upper quartile and lower quartile: 3, 5 and 2

Example 76. Consider the experiment of tossing **three fair coins**. Let X be the random variable that denotes the **number of heads** that result.

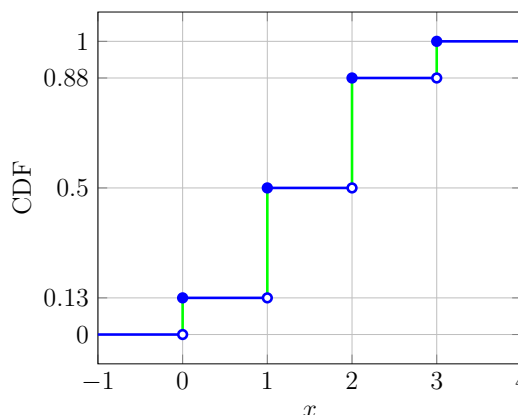
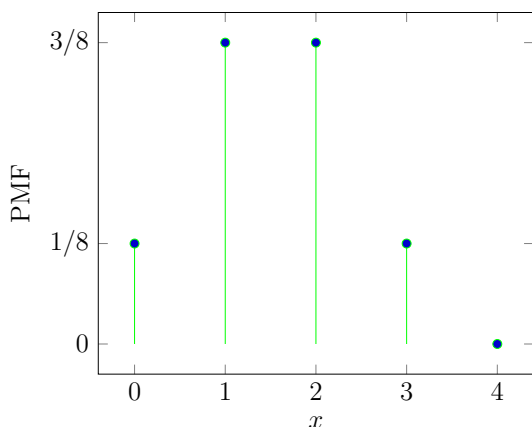
- Plot the CDF and PMF of X .
- Determine $P(X > 2)$
- Determine $P(0 < X \leq 3)$
- Determine $P(0 < X \leq 3 \mid X \leq 2)$
- Determine the median, upper quartile and lower quartile

Solution: The sample space is given by

H H H
H H T
H T H
H T T
T H H
T H T
T T H
T T T

Therefore, the PMF of X is given by

$$p_X(x) = \begin{cases} 1/8 & \text{if } x = 0 \text{ or } x = 3 \\ 3/8 & \text{otherwise} \end{cases}$$



- a) Plot the CDF and PMF of X .
- b) $P(X > 2) = 0.125$
- c) $P(0 < X \leq 3) = 0.875$
- d) $P(0 < X \leq 3 \mid X \leq 2) = 0.857$
- e) the median, upper quartile and lower quartile=1, 2, 1

□

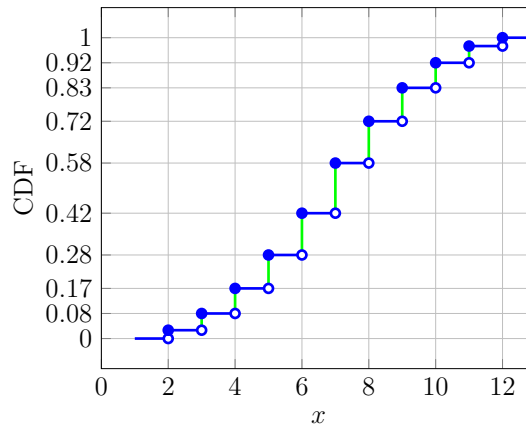
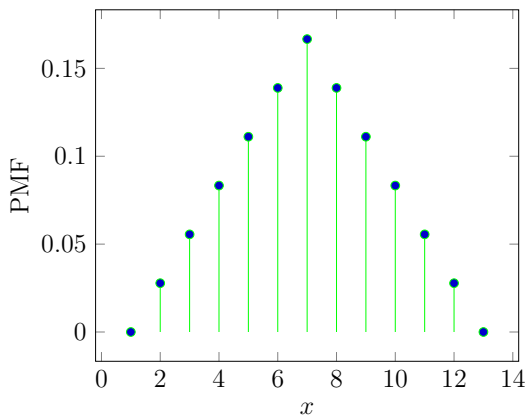
Example 77. — Sum of two dice* Let X be the sum of rolling 2 dice.

- a) Plot the CDF and PMF of X .
- b) Determine $P(X \geq 3)$
- c) Determine $P(5 < X \leq 8)$
- d) Determine $P(5 < X \leq 8 \mid X \leq 10)$
- e) Determine the median, upper quartile and lower quartile

Solution The PMF for X is given by

$$p_X(x) = \begin{cases} 1/36, & x = 2, 12 \\ 2/36, & x = 3, 11 \\ 3/36, & x = 4, 10 \\ 4/36, & x = 5, 9 \\ 5/36, & x = 6, 8 \\ 6/36, & x = 7 \end{cases}$$

Die 1/Die 2						
	2	3	4	5	6	7
	3	4	5	6	7	8
	4	5	6	7	8	9
	5	6	7	8	9	10
	6	7	8	9	10	11
	7	8	9	10	11	12



This is called the discrete triangular distribution

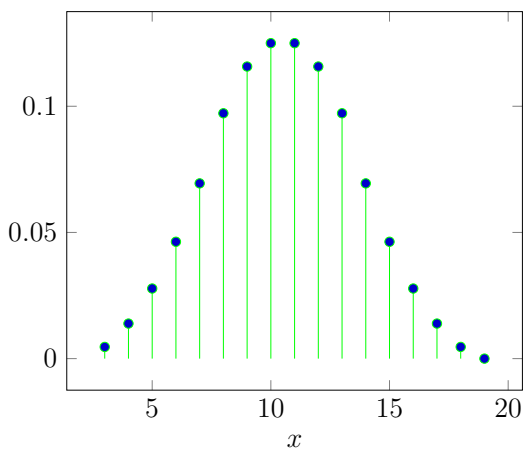
- Plot the CDF and PMF of X .
- $P(X \geq 3) = 0.972$
- $P(5 < X \leq 8) = 0.44$
- $P(5 < X \leq 8 \mid X \leq 10) = .485$
- the median, upper quartile and lower quartile = 7, 9, 5

Example 78. — **Sum of three dice** Let X be the sum of rolling 3 dice.

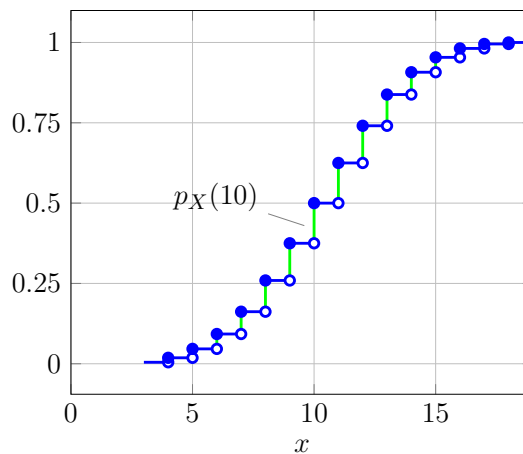
- Plot the CDF and PMF of X .
- Determine $P(X > 5)$
- Determine $P(5 \leq X \leq 15)$
- Determine $P(5 \leq X \leq 15 \mid X \leq 10)$
- Determine the median, upper quartile and lower quartile = 7, 9 and 5

Solution The PMF for X will be derived in future chapters. In the meantime, observe the bell shape that arises (thanks to the *Central Limit theorem*).

PMF of discrete rv: $p_X(x)$



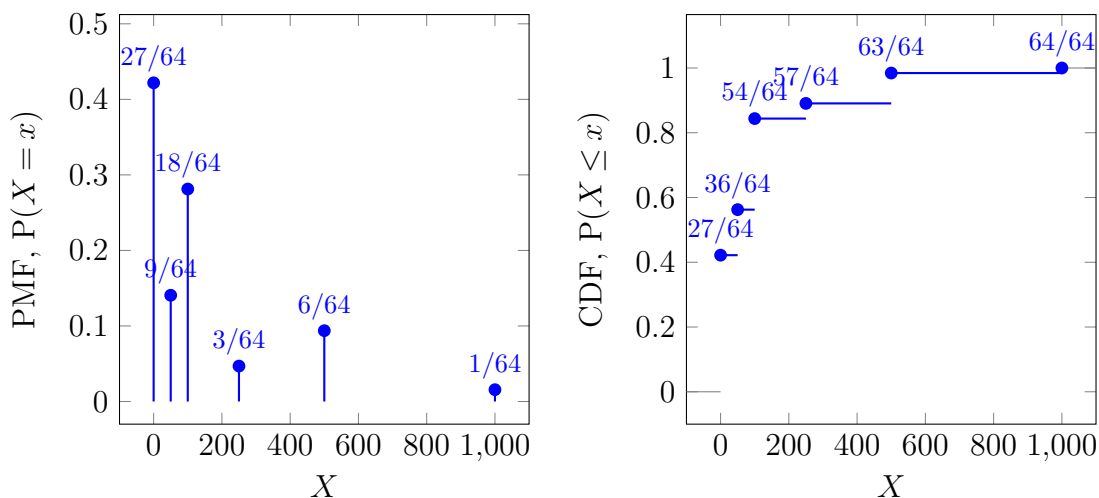
CDF of a discrete rv: $F_X(x)$



This PMF is very close to the *normal distribution*; this resemblance is a statement of the **Central Limit theorem**.

- Plot the CDF and PMF of X .
- $P(X > 5) = 0.953$
- $P(5 \leq X \leq 15) = 0.935$
- $P(5 \leq X \leq 15 \mid X \leq 10) = 0.963$
- the median, upper quartile and lower quartile = 10, 13, 8

STOP! Notice that the sample points of a discrete random variable are not necessarily evenly spaced:



Example 79. You have a coin with probability p of getting ‘Heads’. You flip this coin twice. For each flip, if the result is ‘Heads’, you win \$30. If the result is ‘Tails’, you lose \$20. Let X be your profit in the game.

- What is the sample space?
- Describe the probability mass function of X .
- Describe the cumulative distribution function of X .

Solution:

a)

$$\begin{aligned} S &= \{(30 + 30), (30 - 20), (-20 - 20)\} \\ &= \{60, 10, -40\} \end{aligned}$$

b)

$$P(X = x) = \begin{cases} p^2 & ; \quad x = 60 \\ 2p(1-p) & ; \quad x = 10 \\ (1-p)^2 & ; \quad x = -40 \end{cases}$$

c)

$$F_X(x) = \begin{cases} 0 & ; \quad x < -40 \\ (1-p)^2 & ; \quad -40 \leq x < 10 \\ (1-p)^2 + 2p(1-p) & ; \quad 10 \leq x < 60 \\ (1-p)^2 + 2p(1-p) + p^2 & ; \quad x \geq 60 \end{cases}$$

□

2.4 Expectation

The expectation (also known as **mean**) is probably the most important measure of central tendency of a rv. The others are mode and median.

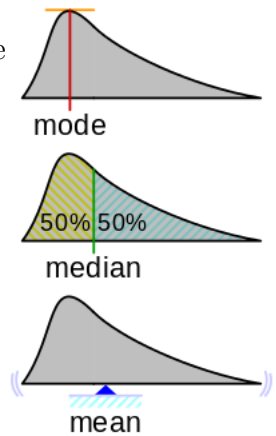
Expectation of a discrete rv Let X be a discrete rv with a set of possible values S and PMF $p_X(x)$. The **expected** or **mean** value of X is

$$E(X) = \mu_X = \sum_{x \in S} x \cdot p_X(x).$$

The symbol μ_X will be used interchangeably with $E(X)$.

In example 76, the expectation of the number of heads is given by

$$E(X) = 0 \times \frac{1}{8} + 3 \times \frac{1}{8} + 1 \times \frac{3}{8} + 2 \times \frac{3}{8} = \frac{3}{2}$$



STOP! As seen from the example, the expectation of a random variable may not be on its sample space.

Example 80. Find $E(X)$ where X is the outcome of rolling a fair dice.

Solution: Let X be the random variable that denotes the result of a single roll of dice. The PMF for X is given by

$$p_X(x) = \frac{1}{6}, \quad x = 1, 2, 3, 4, 5, 6.$$

The expectation of X is given by

$$E(X) = \sum_{x=1}^6 p_X(x) \cdot x = \frac{1}{6} (1 + 2 + 3 + 4 + 5 + 6) = 3.5$$

□

Example 81. When we roll two dice what is the expected value of their sum?

Solution Let X be the random variable denoting the sum. From example 77 above, we know that the PMF for X is given by

$$p_X(x) = \begin{cases} 1/36, & x = 2, 12 \\ 2/36, & x = 3, 11 \\ 3/36, & x = 4, 10 \\ 4/36, & x = 5, 9 \\ 5/36, & x = 6, 8 \\ 6/36, & x = 7 \end{cases}$$

The expectation of X is given by

$$\begin{aligned} E(X) &= \sum_{x=2}^{12} p_X(x) \cdot x \\ &= \frac{1}{36} \times 2 + \frac{2}{36} \times 3 + \frac{3}{36} \times 4 + \frac{4}{36} \times 5 + \frac{5}{36} \times 6 + \frac{6}{36} \times 7 + \\ &\quad \frac{5}{36} \times 8 + \frac{4}{36} \times 9 + \frac{3}{36} \times 10 + \frac{2}{36} \times 11 + \frac{1}{36} \times 12 \\ &= \frac{252}{36} = 7 \end{aligned}$$

Example 82. A class of 120 students is driven in 3 buses to a jazz concert. There are 36 students in the first bus, 40 in the second, and 44 in the third bus. When the buses arrive, a student is randomly chosen. Let X denote the number of students on the bus of the chosen student. Find $E(X)$.

Solution:

$$\begin{aligned} P(X = 36) &= \frac{36}{120} \\ P(X = 40) &= \frac{40}{120} \\ P(X = 44) &= \frac{44}{120} \end{aligned}$$

Therefore,

$$\begin{aligned} E(X) &= \sum_x x \cdot p_X(x) \\ &= 36 \left(\frac{36}{120} \right) + 40 \left(\frac{40}{120} \right) + 44 \left(\frac{44}{120} \right) \\ &= 40.2667 \end{aligned}$$

□

Example 83. — surge suppressor The owner of a small firm has just purchased a personal computer, which she expects will serve her for the next two years. The owner has been told that she "must" buy a surge suppressor to provide protection for her new hardware against possible surges or variations in the electrical current, which have the capacity to damage the computer. The amount of damage to the computer depends on the strength of the surge. It has been estimated that there is a 2% chance of incurring 350 dollar damage, 4% chance of incurring 300 dollar damage, and 11% chance of incurring 100 dollar damage from a surge within the next two years. An inexpensive suppressor, which would provide protection for only one surge, can be purchased.

How much should the owner be willing to pay if she makes decisions on the basis of expected value?

Solution: Let the random variable D be the amount of damage to the computer (in dollars) caused by a surge within the next two years. Then

$$\begin{aligned} E(D) &= 350P(D = 350) + 300P(D = 300) + 100P(D = 100) \\ &= 350 \cdot 0.02 + 300 \cdot 0.04 + 100 \cdot 0.11 = 30 \end{aligned}$$

Hence the owner expects her computer to incur \$30 of damage from a surge in the next two years and so should be willing to pay \$30 for a surge protector. \square

2.4.1 Expectation of a function of X

Many times were interested on a function of a random variable X for which the distribution $p_X(x)$ is known.

$$Y = g(X)$$

If all we want is the expected value $E(Y)$, there are **two options**:

1. compute the PMF of Y , $p_Y(y)$ (chapter 5), the sample space of Y , S_Y and use the definition of expectation

$$E(Y) = \mu_Y = \sum_{Y \in S_Y} y \cdot p_Y(y).$$

2. use Fact 2.2 below:

Fact 2.2 — Expectation of a function of X . For any function $g(X)$,

$$E(g(x)) = \sum_{x \in S} g(x) \cdot p_X(x)$$

Example 84. X has the following distribution.

$$P(X = -1) = 0.2$$

$$P(X = 0) = 0.5$$

$$P(X = 1) = 0.3$$

Find $E[X^2]$.

Solution:[1] Using

$$E(g(x)) = \sum_{x \in S} g(x) \cdot p_X(x)$$

we have:

$$\begin{aligned} E(X^2) &= (-1)^2(0.2) + (0^2)(0.5) + (1^2)(0.3) \\ &= 0.5 \end{aligned}$$

\square

Solution:[2] Let

$$Y = X^2$$

Using option one above, we have

$$\begin{aligned} P(Y = 0) &= P(X^2 = 0) \\ &= P(X = 0) \\ &= 0.5 \end{aligned}$$

$$\begin{aligned} P(Y = 1) &= P(X^2 = 1) \\ &= P(X = -1) + P(X = 1) \\ &= 0.5 \end{aligned}$$

Therefore,

$$\begin{aligned} E(Y) &= E[X^2] \\ &= (0)(0.5) + (1)(0.5) \\ &= 0.5 \end{aligned}$$

□

Example 85. Let the probability distribution of X be

X	-2	-1	0	1	2
$P(X = x)$	0.25	0.1	0.2	0.2	0.25

Calculate $E(|X|)$.

Solution: Using fact 2.2:

$$\begin{aligned} E(Y) = E(|X|) &= \sum_{x \in S_X} |x| \times P(X = x) \\ &= 2 * 0.25 + 1 * 0.1 + 0 * 0.2 + 1 * 0.2 + 2 * 0.25 \\ &= 1.3 \end{aligned}$$

□

Solution: Using option one above, define $Y = |X|$, and we have to calculate $E(Y)$.

$$\begin{aligned} S_Y &= \{0, 1, 2\} \\ P(Y = 0) &= P(|X| = 0) = P(X = 0) = 0.2 \\ P(Y = 1) &= P(X = 1) + P(X = -1) = 0.2 + 0.1 = 0.3 \\ P(Y = 2) &= P(X = 2) + P(X = -2) = 0.25 + 0.25 = 0.5 \end{aligned}$$

Hence

$$E(Y) = 0 \times 0.2 + 1 \times 0.3 + 2 \times 0.5 = 1.3$$

□

Example 86. Show that for any constants a, b and random variable X ,

$$E(a + bX) = a + b E(X)$$

Solution:

$$\begin{aligned} E(a + bX) &= \sum_x (a + bx)P(X = x) \\ &= a \sum_x P(X = x) + b \sum_x x P(X = x) \\ &= a + b E(X) \end{aligned}$$

□

2.4.2 Variance

We are interested in calculating how much a random variable deviates from its mean, some measure of $X - E(X)$. But we do not want the positive and the negative deviations to cancel out each other. This suggests that we should use the absolute value $|X - E(X)|$. But working with absolute values is messy. It turns out that squaring of $X - E(X)$ is more useful. This leads to the following definition.

Variance The variance of a random variable X is defined to be

$$V(X) = E[(X - E(X))^2]$$

The symbol σ_X^2 will be used interchangeably with $V(X)$.

Shortcut formula for the Variance

$$V(X) = E(X^2) - E(X)^2$$

Proof.

$$\begin{aligned} E((X - E(X))^2) &= E(X^2 - 2XE(X) + E(X)^2) \\ &= E(X^2) - 2E(XE(X)) + E(X)^2 \\ &= E(X^2) - 2E(X)^2 + E(X)^2 \\ &= E(X^2) - E(X)^2 \end{aligned}$$

In step 2 we used the linearity of expectation and the fact that $E(X)$ is a constant.

The standard deviation of a random variable X is

$$\sigma_X = \sqrt{V(X)}$$

The advantage is that it has the same units as X , so it is easier to interpret compared to the variance.

Coefficient of variation The coefficient of variation of a random variable X is defined as

$$\delta_X = \frac{\sigma_X}{|\mu_X|}$$

provided $\mu_X \neq 0$. The big advantage here is that the coefficient of variation is dimensionless! As a rule of thumb, when

$$\delta_X < 0.3$$

the random variable has moderate uncertainty.

Fact 2.3

$$V(aX + b) = a^2 V(X)$$

Proof.

$$\begin{aligned}
 V(aX + b) &= E[(aX + b - E(aX + b))^2] \\
 &= E[(aX + b - aE(X) - b)^2] \\
 &= E[a^2(X - E(X))^2] \\
 &= a^2 E[(X - E(X))^2] \\
 &= a^2 V(X)
 \end{aligned}$$

■

Example 87. Calculate $V(X)$ where X represents the outcome of rolling a fair dice.

Solution:

$$\begin{aligned}
 E(X) &= \sum_{x=1}^6 \frac{1}{6}x \\
 &= \frac{7}{2} \\
 E[X^2] &= \sum_{x=1}^6 \frac{1}{6}x^2 \\
 &= \frac{91}{6}
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 V(X) &= E[X^2] - E(X)^2 \\
 &= \frac{91}{6} - \frac{49}{4} \\
 &= \frac{35}{12}
 \end{aligned}$$

Notice that the coefficient of variation of a dice is about 0.5:

$$\delta_X = \frac{\sigma_X}{|\mu_X|} = \frac{\sqrt{35/12}}{7/2} \approx 0.49$$

□

Example 88. * Recall example 79. You have a coin with probability p of getting ‘Heads’. You flip this coin twice. For each flip, if the result is ‘Heads’, you win \$30. If the result is ‘Tails’, you lose \$20.

Let X be your profit in the game.

- What is the expected value of X ?
- What is the value of p upto which you would agree to participate in the game?
- What is $V(X)$?
- What is σ_X ?

Solution:

a)

$$\begin{aligned} E[X] &= (-40)(1-p)^2 + (10)2p(1-p) + (60)p^2 \\ &= 100p - 40 \end{aligned}$$

b) $100p - 40 > 0 \rightarrow p > 2/5$ c) Let's use $V(X) = E[X^2] - E(X)^2$:

$$\begin{aligned} E[X^2] &= (-40)^2(1-p)^2 + (10)^2 2p(1-p) + (60)^2 p^2 \\ &= 5000p^2 - 3000p + 1600 \end{aligned}$$

$$\begin{aligned} E(X)^2 &= (100p - 40)^2 \\ &= 10000p^2 - 8000p + 1600 \end{aligned}$$

Therefore,

$$\begin{aligned} V(X) &= E[X^2] - E(X)^2 \\ &= 5000p(1-p) \end{aligned}$$

d)

$$\begin{aligned} \sigma_X &= \sqrt{V(X)} \\ &= \sqrt{5000p(1-p)} \end{aligned}$$

□

Example 89. Consider three random variables X, Y, Z measured in the same units. Their probability mass distribution is as follows.

$$P(X = x) = \begin{cases} 1/2, & x = -2 \\ 1/2, & x = 2 \end{cases}$$

$$P(Y = y) = \begin{cases} 0.001, & y = -10 \\ 0.998, & y = 0 \\ 0.001, & y = 10 \end{cases}$$

$$P(Z = z) = \begin{cases} 1/3, & z = -10 \\ 1/3, & z = 0 \\ 1/3, & z = 10 \end{cases}$$

Which of the above random variables is more “spread out”?

Solution: It is easy to see that $E(X) = E(Y) = E(Z) = 0$.

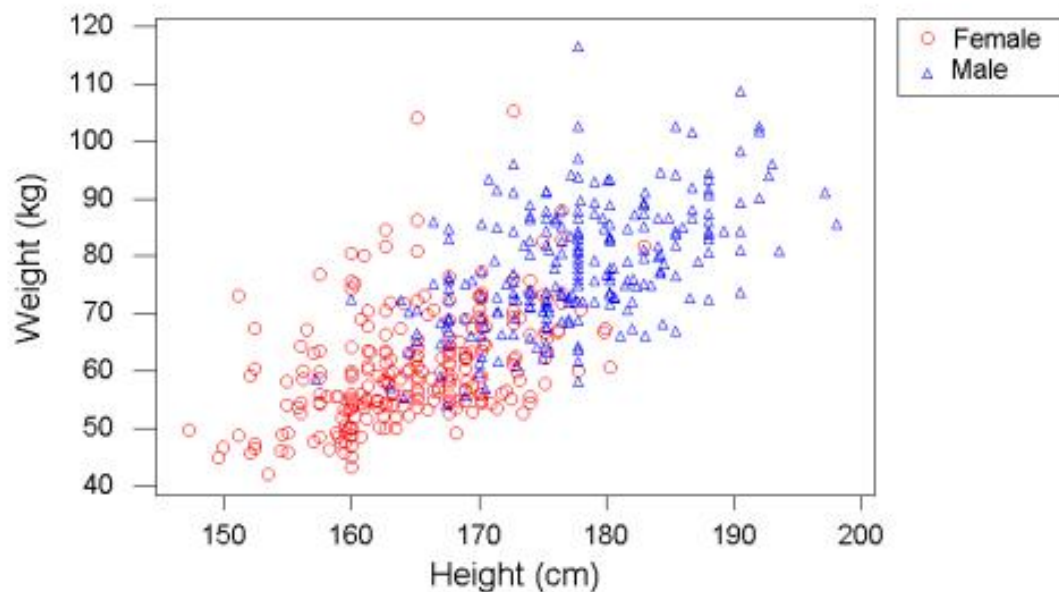
$$\begin{aligned}
 V(X) &= E(X^2) \\
 &= 0.5 \cdot (-2)^2 + 0.5 \cdot (2)^2 \\
 &= 4 \\
 V(Y) &= E(Y^2) \\
 &= 0.001 \cdot (-10)^2 + 0.998 \cdot 0^2 + 0.001 \cdot (10)^2 \\
 &= 0.2 \\
 V(Z) &= E(Z^2) \\
 &= (1/3) \cdot (-5)^2 + (1/3) \cdot 0^2 + (1/3) \cdot (5)^2 \\
 &= 16.67
 \end{aligned}$$

Thus Z is the most spread out and Y is the most concentrated. □

2.5 Jointly Distributed Discrete Random Variables

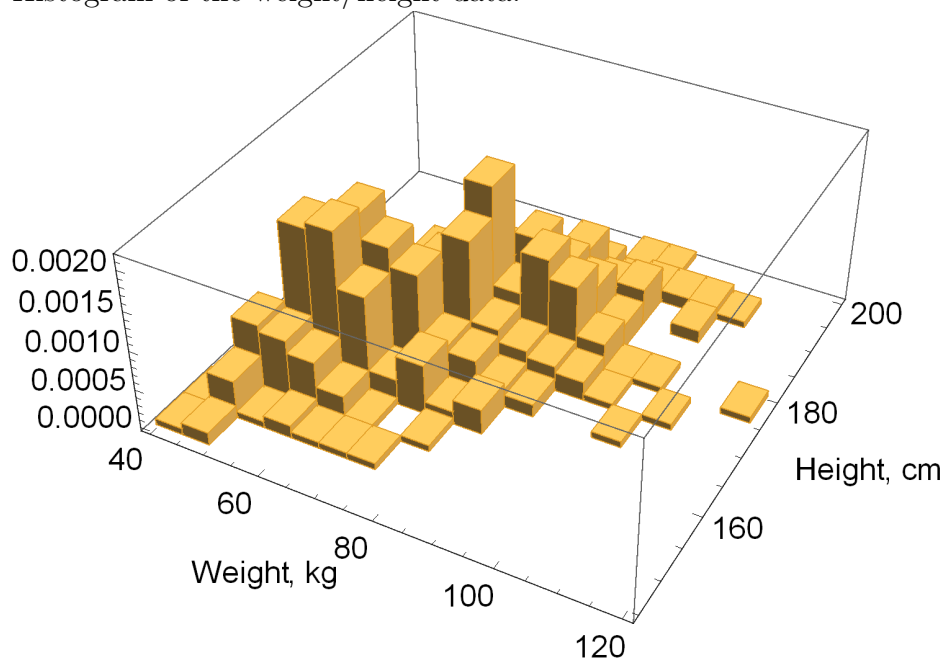
There are many practical situations where we need to deal with more than two measurement simultaneously. The Joint Probability Distribution of (X, Y) describes the *joint* random behavior of X and Y .

For instance we might be interested in the relationship between heights and weights of a population. Let (X_i, Y_i) denote the (weight, height) of person i , then (X_i, Y_i) are related since we can expect that if X_i is large/small then the associated Y_i tends to be large/small. Thus X_i and Y_i are **correlated** and we should describe the behavior of (X_i, Y_i) jointly (together).



→ [data source](#)

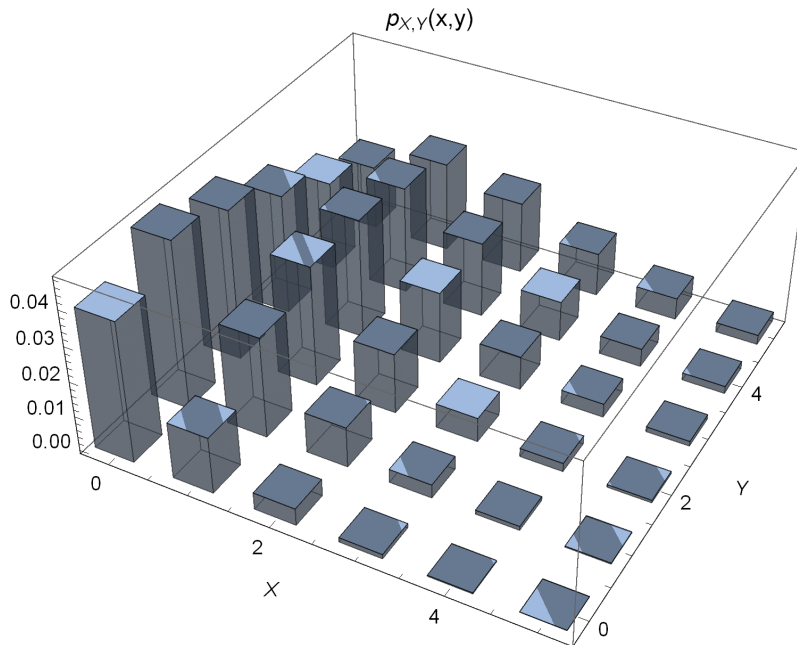
Histogram of the weight/height data:



The data for the histogram is:

$W \setminus H$	147 - 157	157 - 167	167 - 177	177 - 187	187 - 197	197 - 207
42 - 52	3	18	6	0	0	0
52 - 62	0	21	82	21	0	0
62 - 72	0	7	57	50	6	0
72 - 82	0	2	23	61	28	1
82 - 92	0	0	9	38	35	5
92 - 102	0	0	0	10	10	5
102 - 112	0	0	1	2	2	3
112 - 122	0	0	0	1	0	0

Dividing this table entries by the sum (507 individuals) would give the joint PMF. In general, a joint distribution looks like this:



2.5.1 Chapter 1 results in PMF notation

Recall from chapter 1 For two events A and B in sample space S :

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (\text{conditional probability})$$

$$P(A \cap B) = P(A|B)P(B) \quad (\text{multiplication rule})$$

$$P(A \cap B) = P(A)P(B) \quad (\text{independence})$$

$$P(B) = \sum_{i=1}^n P(B \cap A_i) \quad (\text{total probability, where } S \text{ is partitioned into } A_1, \dots, A_n)$$

$$= \sum_{i=1}^n P(B|A_i)P(A_i)$$

These definitions we developed for events in chapter 1 extend to random variables by defining events A and B and their intersection as

$$A = (X = x) \quad \text{and} \quad B = (Y = y) \quad \text{and} \quad A \cap B = (X = x, Y = y)$$

And now we express the results from chapter 1 in PMF notation, that is, in terms of probability mass functions. We start with $P(A \cap B)$, which we now call the joint PMF:

Joint PMF For two random variables X and Y the joint PMF is

$$p_{X,Y}(x,y) = P(X = x, Y = y)$$

Note:

$$\sum_X \sum_Y p_{X,Y}(x,y) = 1$$

The total probability rule we now call marginal PMF:

Marginal PMF The PMF of a single random variable is called marginal PMF:

$$p_X(x) = \sum_Y p_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = \sum_X p_{X,Y}(x,y)$$

The conditional probability rule we now call conditional PMF:

Conditional PMF Let X and Y be two discrete random variables. Then the conditional probability mass function (conditional pmf) of Y given $X = x$ is defined as,

$$p_{Y|X=x}(y) = P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{p_{X,Y}(x,y)}{p_X(x)}$$

for $y \in S_Y$ and $x \in S_X$. Please note that in the term $p_{Y|X=x}(y)$ we are thinking as if x is fixed and y is the variable, but obviously both can vary.

The conditional pmf of X given $Y = y$ for some $y \in S_Y$ is similarly defined as,

$$p_{X|Y=y}(x) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$



The PMF version of the Multiplication Rule:

$$\begin{aligned} p_{X,Y}(x,y) &= p_Y(y)p_{X|Y}(x) \\ &= p_X(x)p_{Y|X}(y) \end{aligned}$$

can be combined with the definition of marginal PMF:

$$p_X(x) = \sum_Y p_Y(y)p_{X|Y}(x)$$

The condition for independence $P(X = x, Y = y) = P(X = x) \cdot P(Y = y)$ in PMF notation becomes

Independence of Two Random Variables. Random variables X and Y are independent iff:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

for all (x,y) .

Alternative we could use $p_{X|Y}(x) = p_X(x)$ or $p_{Y|X}(y) = p_Y(y)$ to check for independence, but the above definition is the most common.

Calculating probabilities For any two random variables X and Y having joint PMF X :

$$P((X,Y) \in A) = \sum_{(x,y) \in A} p_{X,Y}(x,y)$$

Finally, let's define the joint CDF although for some reason it is rarely used in this chapter.

Joint Cumulative Probability Distribution Function For any two random variables X and Y , the joint cumulative probability distribution function of X and Y is defined to be

$$F_{X,Y}(a,b) = P(X \leq a, Y \leq b) = \sum_{x \leq a, y \leq b} p_{X,Y}(x,y)$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}$.

Example 90. 3 balls are randomly selected from an urn containing 3 red, 4 white, and 5 blue balls.

Let X be the number of red balls chosen.

Let Y be the number of white balls chosen.

Find the joint probability mass function of X and Y .

Solution:

$$P(X = x, Y = y) = \frac{\binom{3}{x} \binom{4}{y} \binom{5}{3-x-y}}{\binom{12}{3}}$$

Therefore,

$X \backslash Y$	0	1	2	3	TOT = $p_X(x)$
0	10/220	40/220	30/220	4/220	84/220
1	30/220	60/220	18/220	0	108/220
2	15/220	12/220	0	0	37/220
3	1/220	0	0	0	1/220
TOT = $p_Y(y)$	56/220	112/220	48/220	4/220	1

□

Example 91. Given the joint distribution of (X, Y) :

$p_{X,Y}(x,y)$		X		
		0	1	2
	0	3/28	9/28	3/28
Y	1	3/14	3/14	0
	2	1/28	0	0/28

are X and Y statistically independent?

Solution: Recall that random variables X and Y are independent iff:

$$p_{X,Y}(x,y) = p_X(x)p_Y(y)$$

for all (x,y) , and that the marginal distributions are defined as

$$p_X(x) = \sum_Y p_{X,Y}(x,y) \quad \text{and} \quad p_Y(y) = \sum_X p_{X,Y}(x,y)$$

which correspond to the row totals and column totals:

		X				
	$p_{X,Y}(x,y)$	0	1	2	$p_Y(y)$	
	0	3/28	9/28	3/28	15/28	
	Y	1	3/14	3/14	0	3/7
	2	1/28	0	0	1/28	
	$p_X(x)$	5/14	15/28	3/28	1	

Since $p_{X,Y}(0,0) \neq p_X(0) \cdot p_Y(0)$, X and Y are not independent. □

Example 92. In a class there are four freshman boys, six freshman girls, and six sophomore boys. How many sophomore girls must be present if gender and class are to be independent when a student is selected at random?

Solution: In class. □

Example 93. * The joint distribution of X and Y is given in the following table

$p_{XY}(x,y)$	X=0	X=1	X=3	TOT
Y=-1	0.11	0.03	0	0.14
Y=2.5	0.03	0.09	0.16	0.28
Y=3	0.15	0.15	0.06	0.36
Y=4.7	0.04	0.16	0.02	0.22
TOT	0.33	0.43	0.24	1

Find a) $P(Y - X \leq 2)$, b) $P(2 \leq Y \leq 4 | X = 1)$.

Solution: a) The best way to do this is to calculate the values of the required function, $Y - X$ in this case, and highlight the cells that are favorable to our condition, being less than or equal to two in this case:

Y-X	X=0	X=1	X=3
Y=-1	-1	-2	-4
Y=2.5	2.5	1.5	-0.5
Y=3	3	2	0
Y=4.7	4.7	3.7	1.7

Then, we add up all the joint probabilities corresponding to these highlighted cells:

$p_{XY}(x,y)$	X=0	X=1	X=3
Y=-1	0.11	0.03	0
Y=2.5	0.03	0.09	0.16
Y=3	0.15	0.15	0.06
Y=4.7	0.04	0.16	0.02

which gives 0.62.

b) The conditional distribution of $Y | X = 1$ is :

$p_{Y X=1}(y)$	X=1
Y=-1	0.03/0.43
Y=2.5	0.09/0.43
Y=3	0.15/0.43
Y=4.7	0.16/0.43
TOT	1

and the desired probability is $0.09/0.43+0.15/0.43=0.558$.

□

Example 94. The joint distribution of X and Y is given in the following table

$p_{XY}(x,y)$	X=-1	X=-2	X=2	X=3	TOT
Y=-3	0.14	0.14	0.01	0.05	0.34
Y=-1	0.15	0.06	0.06	0.04	0.31
Y=1	0.03	0.1	0.11	0.11	0.35
TOT	0.32	0.3	0.18	0.2	1

Find $P(Y + X \leq 0)$, $P(-2 \leq X \leq 2|Y = -1)$.

Solution: a)0.68 b)0.387

□

Example 95. Roll a balanced dice twice. Define random variables:

- X = number of 4's
- Y = number of 5's

- (a) Find the joint distribution of X and Y , $p_{X,Y}(x,y)$.
 (b) Find $P((X,Y) \in A)$ where $A = \{2x + y < 3\}$

Solution:

Possible values of X and Y : $x = 0, 1, 2$, $y = 0, 1, 2$, $x + y \leq 2$.

Sample space

		Roll 2					
		1	2	3	4	5	6
Roll 1	1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
	2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
	3	(3,1)	(3,2)	(3,3)	(3,4)	(3,5)	(3,6)
	4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
	5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
	6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

(36 equally likely sample points)

$p_{X,Y}(x,y)$		y		
		0	1	2
	0	16/36	8/36	1/36
x	1	8/36	2/36	0
	2	1/36	0	0

(Check if $\sum_x \sum_y p_{X,Y}(x,y) = 1$?)

(b) Find $P((X, Y) \in A)$ where $A = \{2x + y < 3\}$

$$P((X, Y) \in A) = p_{X,Y}(0,0) + p_{X,Y}(0,1) + p_{X,Y}(0,2) + p_{X,Y}(1,0) = 33/36 \quad \square$$

Example 96. Toss a balanced coin 3 times. Define random variables

- X = number of heads
- Y = (number of heads) – (number of tails)

Find the joint distribution of X and Y .

Solution:

Possible values of X and Y :

$$\begin{aligned} x &= 0, 1, 2, 3 \\ y &= -3, -1, 1, 3 \\ 2x - y &= 3 \quad (\text{why?}) \end{aligned}$$

Sample space

$$S = \{HHH, HHT, HTH, THH, HTT, THT, TTH, TTT\}$$

(8 equally likely sample points)

$p_{X,Y}(x, y)$	-3	-1	1	3
x	0	1	2	3
0	$1/8$	0	0	0
1	0	$3/8$	0	0
2	0	0	$3/8$	0
3	0	0	0	$1/8$

(Check if $\sum_x \sum_y p_{X,Y}(x, y) = 1$ —YES!) □

Example 97. — The Titanic data :

Passenger Status	Survivors	Fatalities	TOTAL
First Class	203	122	325
Second Class	118	167	285
Third Class	178	528	706
Crew	212	673	885
Total	711	1490	2201

There are two random variables in play here,

$$\begin{aligned} X &= 0 && \text{if passenger survived} \\ &= 1 && \text{if passenger died} \end{aligned}$$

and

$$\begin{aligned} Y &= 1 && \text{if passenger was in first class} \\ &= 2 && \text{if passenger was in second class} \\ &= 3 && \text{if passenger was in third class} \\ &= 4 && \text{if passenger was a crew member} \end{aligned}$$

So we approximate the joint PMF of X and Y as,

	X=0	X=1
Y=1	0.09	0.06
Y=2	0.05	0.08
Y=3	0.08	0.24
Y=4	0.10	0.30

For example $p_{X,Y}(0,1) = \frac{203}{2201} = 0.09$.

Find $P(X+Y \leq 2)$.

Solution:

$$\begin{aligned} P(X+Y \leq 2) &= P(X=0, Y=1) + P(X=0, Y=2) + P(X=1, Y=1) \\ &= 0.09 + 0.05 + 0.06 \\ &= 0.2 \end{aligned}$$

□

Example 98. Let's see how we can apply conditional probabilities in the titanic example,

	X=0	X=1
Y=1	0.09	0.06
Y=2	0.05	0.08
Y=3	0.08	0.24
Y=4	0.10	0.30

Determine the probability of being a survivor given the class.

Solution:

$$\begin{aligned}
P(\text{Survivor}|\text{First Class}) &= p_{X|Y=1}(0) = \frac{p_{X,Y}(0,1)}{p_Y(1)} = \frac{0.09}{0.09+0.06} = 0.6 \\
P(\text{Survivor}|\text{Second Class}) &= p_{X|Y=2}(0) = \frac{p_{X,Y}(0,2)}{p_Y(2)} = \frac{0.05}{0.05+0.08} = 0.3846 \\
P(\text{Survivor}|\text{Third Class}) &= p_{X|Y=3}(0) = \frac{p_{X,Y}(0,3)}{p_Y(3)} = \frac{0.08}{0.08+0.24} = 0.25 \\
P(\text{Survivor}|\text{Crew}) &= p_{X|Y=4}(0) = \frac{p_{X,Y}(0,4)}{p_Y(4)} = \frac{0.1}{0.1+0.3} = 0.25
\end{aligned}$$

So we see that probability of survival across passenger class is decreasing, although we should be careful while making remarks like this and consider other factors present. \square

2.5.2 Expectation with two random variables

In direct analogy with a single discrete random variable, the expected value of a function of random variables X and Y is defined as follows:

Expected value rule for multiple random variables

$$E(g(X, Y)) = \sum_X \sum_Y g(x, y) p_{X,Y}(x, y)$$

Conditional Expectation

$$E(g(X, Y)|Y = y) = \sum_X g(x, y) p_{X|Y}(x)$$

Linearity of Expectation

One of the most important properties of expectation that simplifies its computation is the *linearity of expectation*. By this property, the expectation of the sum of random variables equals the sum of their expectations. This is given formally in the following theorem.

Fact 2.4 — Linearity of Expectation. For any finite collection of random variables X_1, X_2, \dots, X_n ,

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$$

Proof. We will prove the statement for two random variables X and Y . The general claim can be proven using induction.

$$\begin{aligned}
E(X + Y) &= \sum_x \sum_y (x + y)p_{X,Y}(x, y) \\
&= \sum_x \sum_y (xp_{X,Y}(x, y) + yp_{X,Y}(x, y)) \\
&= \sum_x \sum_y xp_{X,Y}(x, y) + \sum_x \sum_y yp_{X,Y}(x, y) \\
&= \sum_x x \sum_y p_{X,Y}(x, y) + \sum_y y \sum_x p_{X,Y}(x, y) \\
&= \sum_x xp_X(x) + \sum_y yp_Y(y) \\
&= E(X) + E(Y)
\end{aligned}$$

■

It is important to note that no assumptions have been made about the random variables while proving the above theorem. For example, the random variables do not have to be independent for linearity of expectation to be true.

Example 99. Using linearity of expectation calculate the expected value of the sum of the numbers obtained when two dice are rolled.

Solution: Let X_1 and X_2 denote the random variables that denote the result when dice 1 and dice 2 are rolled respectively. We want to calculate $E(X_1 + X_2)$. By linearity of expectation

$$\begin{aligned}
E(X_1 + X_2) &= E(X_1) + E(X_2) \\
&= 3.5 + 3.5 \\
&= 7
\end{aligned}$$

□

Example 100. — Indicator (aka Bernoulli) Random Variables. Let X be the random variable such that

$$X = \begin{cases} 1 & \text{if event } A \text{ occurs (with probability } p = P(A)) \\ 0 & \text{otherwise} \end{cases}$$

show that

$$\begin{aligned}
E(X) &= p \\
V(X) &= p(1 - p)
\end{aligned}$$

Solution:

$$\begin{aligned}
E(X) &= 1 \times p + 0 \times (1 - p) = p \\
E(X^2) &= 1^2 \times p + 0^2 \times (1 - p) = p \\
V(X) &= E(X^2) - E(X)^2 = p - p^2 = p(1 - p)
\end{aligned}$$

□

Example 101. — **The hat check problem.** Suppose that n people leave their hats at the hat check. If the hats are randomly returned what is the expected number of people that get their own hat back?

Solution: Let X be the random variable that denotes the number of people who get their own hat back. Let $X_i, 1 \leq i \leq n$, be the random variable such that

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th person gets his/her own hat back (with probability } p = 1/n) \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$E(X_i) = 1 \times \frac{1}{n} + 0 \times \left(1 - \frac{1}{n}\right) = \frac{1}{n}$$

Now clearly,

$$X = \sum_{i=1}^n X_i = X_1 + X_2 + X_3 + \dots + X_n$$

By linearity of expectation we get

$$E(X) = \sum_{i=1}^n E(X_i) = \sum_{i=1}^n \frac{1}{n} = n \times \frac{1}{n} = 1$$

So on average, only one person will get his/her hat back! □

Example 102. Suppose we throw n balls into n bins with the probability of a ball landing in each of the n bins being equal. What is the expected number of empty bins?

Solution: Let X be the random variable denoting the number of empty bins. Let X_i be a random variable that is 1 if the i -th bin is empty and is 0 otherwise:

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th bin is empty, with probability } p = (1 - 1/n)^n \\ 0 & \text{otherwise} \end{cases}$$

Clearly

$$X = \sum_{i=1}^n X_i$$

By linearity of expectation, we have

$$\begin{aligned} E(X) &= \sum_{i=1}^n E(X_i) \\ &= \sum_{i=1}^n P(X_i = 1) \\ &= \sum_{i=1}^n \left(1 - \frac{1}{n}\right)^n \\ &= n \left(1 - \frac{1}{n}\right)^n \end{aligned}$$

As $n \rightarrow \infty$, $\left(1 - \frac{1}{n}\right)^n \rightarrow \frac{1}{e}$. Hence, for large enough values of n we have

$$E(X) \approx n/e$$

This means that on average about one third of the bins will be empty. □

Fact 2.5 If X and Y are independent, then, for any function g and h ,

$$E[g(X)h(Y)] = E[g(X)] E[h(Y)]$$

Proof.

$$\begin{aligned} E(g(X)h(Y)) &= \sum_x \sum_y g(x)h(y)p_{X,Y}(x,y) \\ &= \sum_x \sum_y g(x)h(y)p_X(x)p_Y(y) \\ &= \sum_x g(x)p_X(x) \sum_y h(y)p_Y(y) \\ &= E[g(X)] E[h(Y)] \end{aligned}$$

Fact 2.6 — $E(XY) = E(X)E(Y)$ when X, Y are independent. This follows from the previous result with

$$g(X) = X, \quad h(Y) = Y$$

2.6 Covariance

Covariance Let X and Y be random variables. Then

$$\text{Cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

is defined to be the covariance of X and Y .

Fact 2.7 — **Shortcut formula for covariance.**

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

Proof.

$$\begin{aligned} \text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E[XY - E(X)Y - XE(Y) + E(X)E(Y)] \\ &= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

■

Fact 2.8 If X and Y are independent, then

$$\text{Cov}(X, Y) = 0$$

However, the converse is not true.

Proof.

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y)$$

As X and Y are independent $E(XY) = E(X)E(Y)$, so

$$\begin{aligned}\text{Cov}(X, Y) &= E(X)E(Y) - E(X)E(Y) \\ &= 0\end{aligned}$$

■

Uncorrelated random variables X and Y are said to uncorrelated if and only if

$$\text{Cov}(X, Y) = 0$$

The units of covariance are the [units of X] \times [units of Y], which is problematic because we don't know how to interpret them. For this reason it is better to use the correlation coefficient:

(Pearson's) correlation coefficient

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}$$

It can be shown that

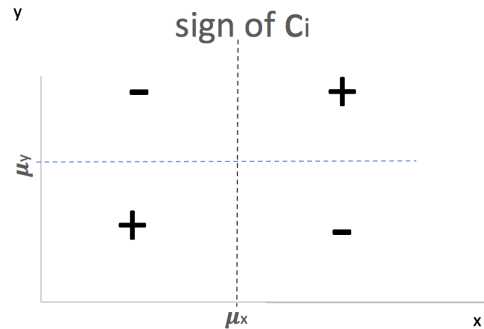
$$-1 \leq \rho_{X,Y} \leq +1$$

ρ is unitless, therefore it can be used to compare across different ramp variables.

2.6.1 Interpretation of Covariance and Correlation

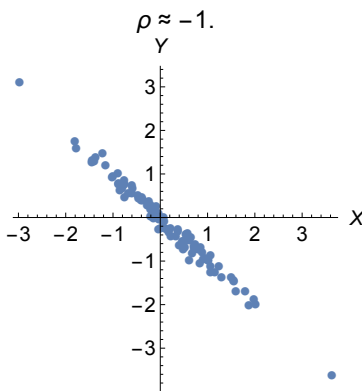
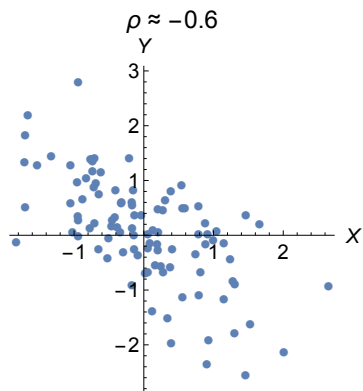
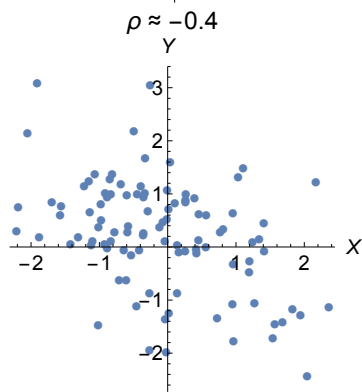
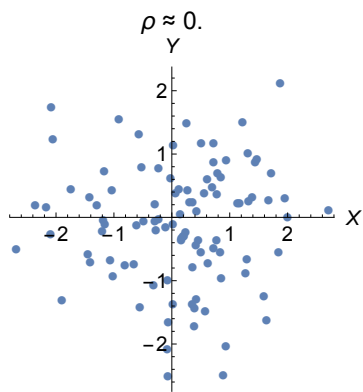
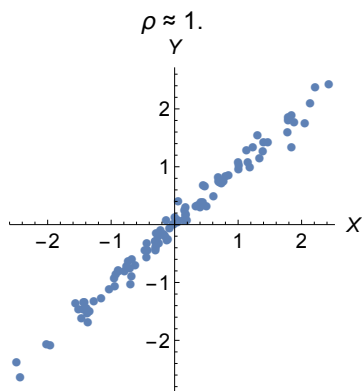
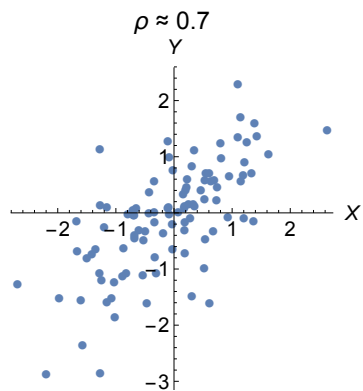
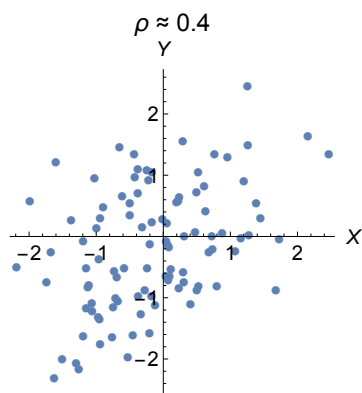
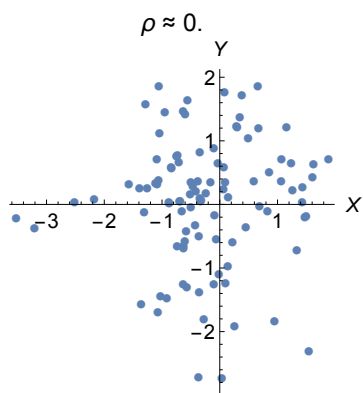
Let the following approximation of $\text{Cov}(X, Y)$ based on a sample $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$:

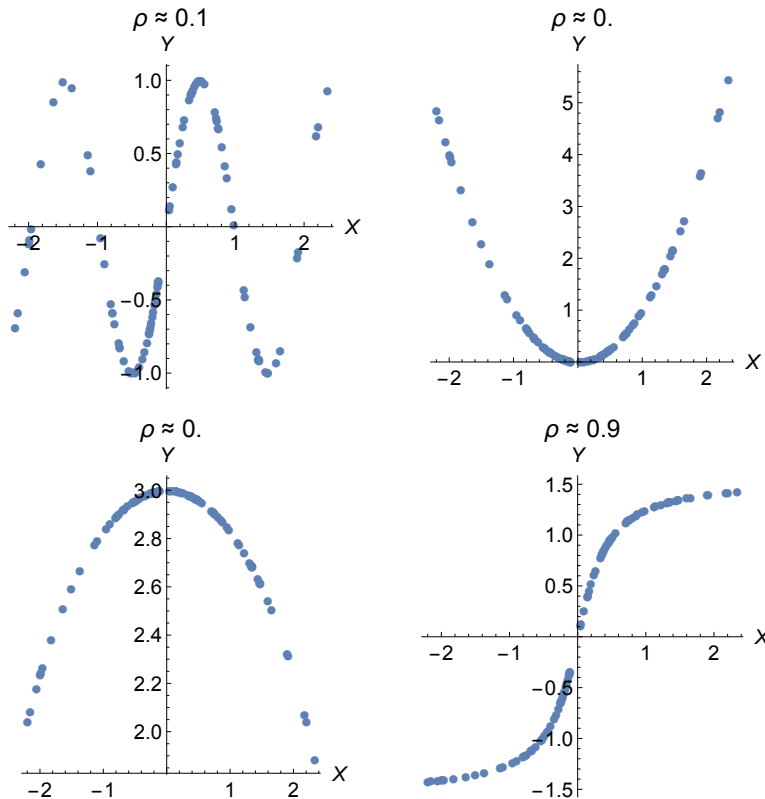
$$\begin{aligned}\text{Cov}(X, Y) &= E\left[(X - \mu_X)(Y - \mu_Y)\right] \\ &= \sum_{x,y} (x - \mu_X)(y - \mu_Y)p_{X,Y}(x, y) \\ &\approx \frac{1}{n} \sum_{i=1}^n \underbrace{(x_i - \mu_X)(y_i - \mu_Y)}_{c_i}\end{aligned}$$



The term $c_i = (x_i - \mu_X)(y_i - \mu_Y)$ will be located in the positive or negative quadrant on the following figure depending upon the sign of each term. If x_i and y_i are both greater or both less than their means, then c_i is positive. If one is positive and one is negative then c_i is negative:

The following figures illustrate scatter plots for typical values for the correlation coefficient.





STOP! Correlation does not necessarily mean causation. For example:

1. The yield of oranges and apples are highly correlated in the Monterey Valley. Therefore, to produce more apples one should produce more oranges?
2. There is a high correlation between the number of police officers and the number of crimes on a given city. Therefore, to reduce crime rates one should reduce the police force?
3. → [More examples from Wikipedia...](#)

The use of a **controlled experiment** is the most effective way of establishing causality between variables. In a controlled study, the sample or population is split in two, with both groups being comparable in almost every way. The two groups then receive different treatments, and the outcomes of each group are assessed.

2.6.2 Covariance of linear combinations

In this section we consider multiple random variables

$$\mathbf{X} = (X_1, X_2, \dots, X_n)^T$$

which we call the random vector \mathbf{X} , with means, variances and covariances (and correlations) given:

$$\begin{aligned} E(X_i) &= \mu_i \\ V(X_i) &= \sigma_i^2 \\ \text{Cov}(X_i, X_j) &= \sigma_{ij} \\ &= \rho_{ij} \sigma_i \sigma_j \end{aligned}$$

Covariance matrix The covariance matrix of the random vector \mathbf{X} with vector mean $\boldsymbol{\mu} = \{\mu_1, \mu_2, \dots, \mu_n\}$ is the $n \times n$ matrix whose $(i, j)^{th}$ element is $\text{Cov}(X_i, X_j)$:

$$\Sigma_{\mathbf{X}} = \text{E} \left((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T \right) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

Its two main properties are:

1. it is symmetric since

$$\text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i)$$

2. its diagonal elements give the variance, since

$$\text{Cov}(X_i, X_i) = \sigma_i^2$$

Fact 2.9 — Covariance of linear combinations. Let the following linear combinations:

$$U = \sum_{i=1}^n a_i X_i = \mathbf{a}^T \mathbf{X}$$

$$V = \sum_{j=1}^n b_j X_j = \mathbf{b}^T \mathbf{X}$$

with $\mathbf{a} = (a_1, \dots, a_n)^T$ and $\mathbf{b} = (b_1, \dots, b_n)^T$. Then,

$$\begin{aligned} \text{Cov}(U, V) &= \sum_{i=1}^n \sum_{j=1}^n a_i b_j \sigma_{ij} \\ &= \mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{b} \end{aligned}$$

Proof. From the properties of the expectation we have that:

$$\text{E}(U) = \sum_{i=1}^n a_i \mu_i = \mathbf{a}^T \boldsymbol{\mu}, \quad \text{E}(V) = \sum_{j=1}^n b_j \mu_j = \mathbf{b}^T \boldsymbol{\mu}$$

therefore, by definition, the covariance of U and V is,

$$\begin{aligned} \text{Cov}(U, V) &= \text{E} \left[(U - \text{E}(U)) (V - \text{E}(V)) \right] \\ &= \text{E} \left((\mathbf{a}^T \mathbf{X} - \mathbf{a}^T \boldsymbol{\mu}) (\mathbf{b}^T \mathbf{X} - \mathbf{b}^T \boldsymbol{\mu}) \right) \\ &= \text{E} \left(\mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu}) \mathbf{b}^T (\mathbf{X} - \boldsymbol{\mu}) \right) \\ &= \text{E} \left(\mathbf{a}^T (\mathbf{X} - \boldsymbol{\mu}) (\mathbf{X} - \boldsymbol{\mu})^T \mathbf{b} \right) \\ &= \mathbf{a}^T \text{E} \left((\mathbf{X} - \boldsymbol{\mu}) (\mathbf{X} - \boldsymbol{\mu})^T \right) \mathbf{b} \\ &= \mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{b} \end{aligned}$$

Without matrix notation, the proof goes like this:

$$\begin{aligned}
\text{Cov}(U, V) &= \text{E} \left[(U - \text{E}(U)) (V - \text{E}(V)) \right] \\
&= \text{E} \left[\left(\sum_{i=1}^n a_i X_i - \sum_{i=1}^n a_i \mu_i \right) \left(\sum_{j=1}^m b_j X_j - \sum_{j=1}^m b_j \mu_j \right) \right] \\
&= \text{E} \left[\sum_{i=1}^n a_i (X_i - \mu_i) \sum_{j=1}^m b_j (X_j - \mu_j) \right] \\
&= \text{E} \left[\sum_{i=1}^n \sum_{j=1}^m a_i b_j (X_i - \mu_i) (X_j - \mu_j) \right] \\
&= \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, X_j)
\end{aligned}$$

■

Important corollaries of theorem 2.9:

Variance of linear combination

$$V(\mathbf{a}^T \mathbf{X}) = \mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{a}$$

or in standard notation:

$$V \left(\sum_{i=1}^n a_i X_i \right) = \sum_{i=1}^n a_i^2 \sigma_i^2 + \underbrace{2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \sigma_{i,j}}_{0 \text{ if } X_i \text{'s are independent}} \quad (2.1)$$

This formula illustrates the concept of **propagation of errors**, where the errors (variances and covariances) in the X_i 's produce errors in the function $U = \sum_{i=1}^n a_i X_i$.

Proof. Let $U = \sum_{i=1}^n a_i X_i$

$$\begin{aligned}
V(U) &= \text{Cov}(U, U) && \text{(by definition)} \\
&= \sum_{i=1}^n \sum_{j=1}^n a_i a_j \text{Cov}(X_i, X_j) && \text{(theorem 2.9)} \\
&= \sum_{i=1}^n a_i^2 V(X_i) + \sum_{i=1}^n \sum_{j \neq i}^n a_i a_j \text{Cov}(X_i, X_j) && (i = j \text{ terms first)} \\
&= \sum_{i=1}^n a_i^2 V(X_i) + 2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \text{Cov}(X_i, X_j) && \text{(symmetry of } \Sigma_{\mathbf{X}} \text{)}
\end{aligned}$$

■

Important results for two random variables:

$$V(X + Y) = V(X) + V(Y) + 2\text{Cov}(X, Y)$$

$$V(X - Y) = V(X) + V(Y) - 2\text{Cov}(X, Y)$$

$$\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$$

$$\rho_{aX+b, cY+d} = \rho_{X, Y}$$

Example 103. The random variables X and Y have joint probability distribution specified by the following table:

	$y=1$	$y=2$	$y=3$
$x=1$	0.30	0.05	0.00
$x=2$	0.05	0.20	0.05
$x=3$	0.00	0.05	0.30

- (a) Find the expectation of XY .
 (b) Find the covariance $\text{Cov}(X, Y)$ between X and Y .
 (c) What is the correlation between X and Y ?
 (d) Suppose the random variables X and Y above are connected to random variables U and V by the relations

$$X = 2U + 5$$

$$Y = 4V + 5$$

What is the covariance $\text{Cov}(U, V)$?

- (e) What is the correlation between U and V ?

Solution: Part a) The mass function of XY is tabulated below:

xy	1	2	3	4	6	9
$P(XY = xy)$	0.3	0.1	0.0	0.2	0.1	0.3

Therefore

$$E(XY) = 0.3 + 0.2 + 0 + 0.8 + 0.6 + 2.7 = 4.6$$

Part b)

We require $E(X)$, $E(Y)$, $V(X)$, and $V(Y)$; these are as follows

$$\begin{aligned}
 E(X) &= 0.35 + 0.6 + 1.05 \\
 &= 2.0 \\
 &= E(Y), \\
 E(X^2) &= 0.35 + 2^2 \times 0.3 + 3^2 \times 0.35 \\
 &= 4.7 \\
 &= E(Y^2), \\
 \text{Var}(X) &= 4.7 - 2^2 \\
 &= 0.7 \\
 &= \text{Var}(Y).
 \end{aligned}$$

Hence

$$\begin{aligned}
 \text{Cov}(X, Y) &= E(XY) - E(X)E(Y) \\
 &= 4.6 - 4 \\
 &= 0.6
 \end{aligned}$$

Part c) For the correlation we require

$$\begin{aligned}
 \rho(X, Y) &= \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \\
 &= \frac{0.6}{\sqrt{0.7^2}} \\
 &= \frac{6}{7}
 \end{aligned}$$

Part d)

$$\begin{aligned}
 U &= X/2 - 5/2 \\
 V &= Y/4 - 5/4
 \end{aligned}$$

Since $\text{Cov}(aX + b, cY + d) = ac \text{Cov}(X, Y)$ and so $\text{Cov}(U, V) = \text{Cov}(X, Y)/(2 \cdot 4) = 0.6/(2 \cdot 4)$.

Part e) The correlation between U and V can be written

$$\begin{aligned}
 \rho_{U, V} &= \frac{\text{Cov}(U, V)}{\sqrt{\text{V}(U)}\sqrt{\text{V}(V)}} \\
 &= \frac{\text{Cov}(X, Y)/(2 \cdot 4)}{\sqrt{(\text{V}(X)/2^2)(\text{V}(V)/4^2)}} \\
 &= \rho_{X, Y} \\
 &= \frac{6}{7}
 \end{aligned}$$

□

Example 104. — **2 Dice** There is a blue and yellow dice. Compute the correlation between X , the number on the blue dice, and S , the total of the two dice.

Solution:

Write

$$S = X + Y$$

where Y is the number on the yellow dice. Here $n = 2$, $a_1 = 1$, $a_2 = 0$, and $b_1 = 1$, $b_2 = 1$. Therefore,

$$\begin{aligned} \text{Cov}(X, S) &= \text{Cov}(X, X + Y) \\ &= \text{Cov}(X, X) + \text{Cov}(X, Y) \\ &= V(X) + 0 \end{aligned}$$

Also,

$$V(S) = V(X + Y) = V(X) + V(Y)$$

But $V(Y) = V(X)$. So the correlation between X and S is

$$\rho_{X,S} = \frac{V(X)}{\sqrt{V(X)}\sqrt{2V(X)}} = 0.707$$

Since correlation is at most 1 in absolute value, 0.707 is considered a fairly high correlation. Of course, we did expect X and S to be highly correlated.

What is surprising, though, is that the correlation here is independent of the actual variance of X and Y . So, for instance, if these are unfair dice, but with identical weightings, we still would get a correlation of 0.707. \square

Example 105. Find the covariance matrix of $\{U, V\}$ where

$$U = X_1 + X_2$$

and

$$V = X_1 - X_2$$

Solution: Recall from Fact 2.9:

$$\begin{aligned} \text{Cov}(U, V) &= \mathbf{a}^T \Sigma_{\mathbf{X}} \mathbf{b} \\ &= \sum_{i=1}^n \sum_{j=1}^n a_i b_j \sigma_{ij} \end{aligned}$$

Here $n = 2$, $a_1 = 1$, $a_2 = 1$ and $b_1 = 1$, $b_2 = -1$. Therefore,

$$\begin{aligned} \text{Cov}(U, V) &= \text{Cov}(X_1, X_1) - \text{Cov}(X_1, X_2) + \text{Cov}(X_2, X_1) + \text{Cov}(X_2, X_2) \\ &= \text{Cov}(X_1, X_1) - \text{Cov}(X_2, X_2) \\ &= \sigma_1^2 + \sigma_2^2 \end{aligned}$$

Also,

$$\begin{aligned} V(U) &= \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} \\ V(V) &= \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} \end{aligned}$$

and so the covariance matrix is:

$$\begin{pmatrix} \sigma_1^2 + \sigma_2^2 + 2\sigma_{12} & \sigma_1^2 + \sigma_2^2 \\ \sigma_1^2 + \sigma_2^2 & \sigma_1^2 + \sigma_2^2 - 2\sigma_{12} \end{pmatrix}$$

□

Example 106. *

The joint PMF of precipitation, X (in.) and runoff, Y (cfs) (discretized here for simplicity) due to storms at a given location is as follows:

	$X=1$	$X=2$	$X=3$
$Y=10$	0.0	0.25	0.10
$Y=20$	0.10	0.0	0.10
$Y=30$	0.05	0.15	0.25

- (a) What is the probability that the next storm will bring a precipitation of 2 in. and a runoff of more than 20 cfs?
 (b) After a storm, the rain gauge indicates a precipitation of 2 in. What is the probability that the runoff in this storm is 20 cfs or more?
 (c) Are X and Y statistically independent? Substantiate your answer.
 (d) Determine and plot the marginal PMF of runoff.
 (e) Determine and plot the PMF of runoff for a storm whose participation is 2 in.
 (f) Determine the correlation coefficient between precipitation and runoff.

Solution: Answer: (a) 0.15 (b) 0.375 (c) Not independent (f) 0.1103

(a)

$$\begin{aligned} P(X = 2, Y > 20) &= P(X = 2, Y = 30) \\ &= 0.15 \end{aligned}$$

(b)

$$\begin{aligned} P(Y \geq 20 | X = 2) &= \frac{P(X = 2, Y = 20) + P(X = 2, Y = 30)}{P(X = 2)} \\ &= \frac{0 + 0.15}{0.25 + 0 + 0.15} \\ &= 0.375 \end{aligned}$$

(c)

$$\begin{aligned} P(X = 1) &= 0 + 0.1 + 0.05 \\ &= 0.15 \end{aligned}$$

$$\begin{aligned} P(Y = 10) &= 0 + 0.25 + 0.1 \\ &= 0.35 \end{aligned}$$

$$\begin{aligned} P(X = 1, Y = 10) &= 0 \\ &\neq P(X = 1) \cdot P(Y = 10) \end{aligned}$$

So they are not independent.

(d)

$$\begin{aligned} P(Y = 10) &= 0 + 0.25 + 0.1 \\ &= 0.35 \end{aligned}$$

$$\begin{aligned} P(Y = 20) &= 0.1 + 0 + 0.1 \\ &= 0.2 \end{aligned}$$

$$\begin{aligned} P(Y = 30) &= 0.05 + 0.15 + 0.25 \\ &= 0.45 \end{aligned}$$

(e)

$$\begin{aligned} P(Y = 10|X = 2) &= \frac{0.25}{0.25 + 0 + 0.15} \\ &= 0.625 \end{aligned}$$

$$\begin{aligned} P(Y = 20|X = 2) &= \frac{0}{0.25 + 0 + 0.15} \\ &= 0 \end{aligned}$$

$$\begin{aligned} P(Y = 30|X = 2) &= \frac{0.15}{0.25 + 0 + 0.15} \\ &= 0.375 \end{aligned}$$

(f)

$$\begin{aligned} E(X) &= 0.15 \times 1 + 0.4 \times 2 + 0.45 \times 3 \\ &= 2.3 \end{aligned}$$

$$\begin{aligned} E(Y) &= 0.35 \times 10 + 0.2 \times 20 + 0.45 \times 30 \\ &= 21 \end{aligned}$$

$$\begin{aligned} E(X^2) &= 0.15 \times 1^2 + 0.4 \times 2^2 + 0.45 \times 3^2 \\ &= 5.8 \end{aligned}$$

$$\begin{aligned} E(Y^2) &= 0.35 \times 10^2 + 0.2 \times 20^2 + 0.45 \times 30^2 \\ &= 520 \end{aligned}$$

$$\begin{aligned} E(XY) &= 0 \times 10 + 0.25 \times 20 + 0.10 \times 30 + 0.1 \times 20 + 0 \times 40 + 0.1 \times 60 + 0.05 \times 30 + 0.15 \times 60 + 0.25 \times 90 \\ &= 49 \end{aligned}$$

$$\begin{aligned} \rho &= \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E(X)^2} \sqrt{E(Y^2) - E(Y)^2}} \\ &= \frac{49 - 2.3 \times 21}{\sqrt{5.8 - 2.3^2} \sqrt{520 - 21^2}} \\ &= 0.1103 \end{aligned}$$

□

Example 107. — speeding tickets** A study investigates a group of people who have received speeding tickets in the past year. The joint PMF is summarized in the following table. Let X be the number of tickets received by a person in the past year and Y be the age of that person.

	$X=1$	$X=2$
$Y=30$	c	$c + 1/8$
$Y=40$	c	$c - 1/8$

- (a) Determine the value of c . (5 points)
 (b) Determine the marginal PMF of X , and the conditional PMF of X . (10 points)
 (c) If someone is 30 years old, find the probability that this person gets exact one ticket. (5 points)
 (d) Are X and Y statistically independent? Substantiate your answer. (5 points)
 (e) Determine the correlation coefficient between X and Y . (10 points)

Solution: (a)

$$1 = c + (c + \frac{1}{8}) + c + (c - \frac{1}{8})$$

$$c = 0.25$$

(b) Marginal:

$$p_{X=1} = c + c$$

$$= 0.5$$

$$p_{X=2} = (c + \frac{1}{8}) + (c - \frac{1}{8})$$

$$= 0.5$$

Conditional:

$$p_{X=1|Y=30} = \frac{c}{c + (c + \frac{1}{8})}$$

$$= 0.4$$

$$p_{X=2|Y=30} = \frac{c + \frac{1}{8}}{c + (c + \frac{1}{8})}$$

$$= 0.6$$

$$p_{X=1|Y=40} = \frac{c}{c + (c - \frac{1}{8})}$$

$$= 0.67$$

$$p_{X=2|Y=40} = \frac{c - \frac{1}{8}}{c + (c - \frac{1}{8})}$$

$$= 0.33$$

(c)

$$p_{X=1|Y=30} = \frac{c}{c + (c + \frac{1}{8})}$$

$$= 0.4$$

(d) Not independent. Because $p_{X=1|Y=30} \neq p_{X=1}$

(e)

$$\begin{aligned}
 E(X) &= 1.5 \\
 E(Y) &= \frac{270}{8} \\
 E(X^2) &= 2.5 \\
 E(Y^2) &= 1162.5 \\
 E(XY) &= c \times 1 \times 30 + c \times 1 \times 40 + \left(c + \frac{1}{8}\right) \times 2 \times 30 + \left(c - \frac{1}{8}\right) \times 2 \times 40 \\
 &= 50 \\
 \rho &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \\
 &= \frac{E(XY) - E(X)E(Y)}{\sqrt{E(X^2) - E^2(X)} \sqrt{E(Y^2) - E^2(Y)}} \\
 &= -0.258
 \end{aligned}$$

□

Example 108. * Given the joint distribution of (X, Y) :

$p_{X,Y}(x, y)$		X			Σ
		-1	0	1	
Y	-1	1/16	3/16	1/16	5/16
	0	3/16	0	3/16	3/8
	1	1/16	3/16	1/16	5/16
Σ		5/16	3/8	5/16	1

Calculate the covariance of X and Y . Are they statistically independent?

Solution: The covariance is zero but since $p_{X,Y}(0,0) \neq p_X(0) \cdot p_Y(0)$, X and Y are not independent.

□

Example 109. — **Tornadoes, take 2** 100 structures are located in a region where tornado wind force must be considered in its design. Suppose that from the records of tornadoes for the past 200 years, it is estimated that (i) during any given year the probability of having 0, 1 and 2 tornadoes is 0.5, 0.3 and 0.2, respectively, (ii) the number of tornadoes in different years are independent, and (iii) if a tornado occurs, a structure will be damaged with probability $\mathbf{p=5\%}$.

- if two tornadoes occurred last year, how many structures do you expect to have been damaged?
- what is the probability the a structure will be damaged in the next five years?
- calculate the mean and variance of the number of structures damaged in the next five years?
- If you're a contractor in charge of rehabilitating the structures in the region after a tornado damage, compute the mean and variance of your yearly income, U , if you charge c dollars per rehabilitation work.
- calculate the coefficient of variation of your yearly income, and comment.

Solution: Let

X = number of tornadoes on a given year, $S_X = \{0, 1, 2\}$ tornadoes.

Y_i number of times structure $i = 1, 2, \dots, 100$ is damaged due to tornadoes on a given year.

The event $(Y_i > 0 | X = x)$ is similar to obtaining at least one Head out of x tosses of a coin with $P(\text{Head}) = p$, therefore:

$$P(Y_i > 0 | X = x) = 1 - (1 - p)^x$$

- a) If two tornadoes occurred last year, how many structures do you expect to have been damaged?

Let Z = number of structures damaged last year. Let

$$Z_i = \begin{cases} 1 & \text{if structure } i \text{ was damaged last years} \\ 0 & \text{otherwise} \end{cases}$$

We are interested in the expected value of $Z = \sum_{i=1}^{100} Z_i$. By linearity of expectation we get

$$E(Z) = \sum_{i=1}^{100} E(Z_i) = \sum_{i=1}^{100} P(Z_i = 1) = \sum_{i=1}^{100} P(Y_i > 0 | \mathbf{X}=2) = 100(1 - (1 - p)^2) = 9.75$$

→ 10 structure.

- b) what is the probability the a structure will be damaged in the next five years?

Since the number of tornadoes are independent from year to year, we can focus on a single year, calculate the probability of $D = \mathbf{damage\ in\ one\ year\ for\ one\ structure}$, and then “flip a coin” five times with

$$P(D) = P(Y_i > 0) = 1 - P(Y_i = 0)$$

Since we don't know the number of tornadoes that will occur, we use the total prob. rule:

$$\begin{aligned} P(Y_i = 0) &= \sum_{x=0}^2 P(Y = 0 | X = x)P(X = x) \\ &= \sum_{x=0}^2 (1 - p)^x P(X = x) \\ &= (1 - p)^0 P(X = 0) + (1 - p)^1 P(X = 1) + (1 - p)^2 P(X = 2) \\ &= ((1)(0.5)) + ((0.95)(0.3)) + ((0.95^2)(0.2)) = 0.9655 \end{aligned}$$

then $P(D) = 1 - 0.9655 = 0.0345$ for one year. The desired probability is $1 - (1 - P(D))^5 = 0.16$.

- c) calculate the mean and variance of the number of structures damaged in the next five years?

Let

$$Z_i = \begin{cases} 1 & \text{if structure } i \text{ is damaged in five years} \\ 0 & \text{otherwise} \end{cases}$$

Thus, $P(Z_i = 1) = 0.16$ and $E(Z_i) = 0.16$, $V(Z_i) = 0.16(1 - 0.16) = 0.13$. We are interested in

$$Z = \sum_{i=1}^{100} Z_i$$

By linearity of expectation we get

$$E(Z) = \sum_{i=1}^{100} E(Z_i) = \sum_{i=1}^{100} 0.16 = 16$$

and by result (2.1) for the variance of a sum we have,

$$V(Z) = \sum_{i=1}^{100} V(Z_i) = \sum_{i=1}^{100} 0.13 = 13$$

- d) If you're a contractor in charge of rehabilitating the structures in the region after a tornado damage, compute the mean and variance of your yearly income, U , if you charge c dollars per rehabilitation work.

Let Y_i be the number of times structure i is damaged due to tornadoes on a given year. We are interested in $U = \sum_{i=1}^{100} cY_i$. By linearity of expectation we get

$$E(U) = \sum_{i=1}^{100} cE(Y_i) = 100cE(Y_i)$$

and by result (2.1) for the variance of a sum we have,

$$V(U) = \sum_{i=1}^{100} c^2V(Y_i) = 100c^2V(Y_i)$$

To compute the mean and variance of Y_i we need its PMF. From the problem statement we have p_X , and we can determine $p_{Y_i|X}$ as we did on part a), and then we use the multiplication rule,

$$p_{X,Y_i}(x,y) = p_X(x)p_{Y_i|X}(y).$$

marginal distribution of X :

x	0	1	2
$p_X(x)$	0.5	0.3	0.2

conditional distribution of $Y_i|X$: $P(Y_i = y|X = x) = \binom{x}{y}p^y(1-p)^{x-y}$:

y	0	1	2	Σ
$p_{Y_i X=0}(y)$	1	0	0	1
$p_{Y_i X=1}(y)$	$1-p = 0.95$	$p = 0.05$	0	1
$p_{Y_i X=2}(y)$	$(1-p)^2 = 0.9025$	$2p(1-p) = 0.095$	$p^2 = 0.0025$	1

and we can calculate the joint distribution of (X, Y_i) :

$p_{X,Y_i}(x,y)$	0	$\boxed{Y_i}$ 1	2	Σ
0	0.5	0	0	0.5
\boxed{X} 1	0.285	0.015	0	0.3
2	0.1805	0.019	0.0005	0.2
Σ	0.9655	0.034	0.0005	1

and finally we have the marginal distribution of Y_i :

y	0	1	2
$p_Y(y)$	0.9655	0.034	0.0005

and we obtain $E(Y_i) = 0.035$, $V(Y_i) = 0.0347$. Therefore,

$$E(U) = 3.5c$$

$$V(U) = 3.47c^2$$

e) $\delta_U = V(U)^{1/2} / E(U) = 0.53$

Since this coefficient of variation is greater than 30 %, the yearly income has a large variability. □

Example 110. — Stock Prices, simple portfolio Model Let

X_i = rate of return for stock i

μ_i = expected rate of return (historically 10% annually)

σ_i = price volatility (standard deviation of X_i , historically 15% monthly)

The value of a portfolio of the stocks

$$\mathbf{X} = (X_1, X_2, \dots, X_n)^T$$

is:

$$U = \sum_{i=1}^n a_i X_i = \mathbf{a}^T \mathbf{X}$$

$$E(U) = \sum_{i=1}^n a_i \mu_i = \mathbf{a}^T \boldsymbol{\mu}$$

where the a_i 's represent the weight of each stock in the portfolio, with:

$$0 \leq a_i \leq 1$$

$$\sum_{i=1}^n a_i = 1$$

The risk of the portfolio is given by its variance:

$$V(U) = \mathbf{a}^T \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{a} \tag{2.2}$$

$$= \sum_{i=1}^n a_i^2 \sigma_i^2 + \underbrace{2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \sigma_{i,j}}_{0 \text{ if } X_i \text{'s are independent}} \tag{2.3}$$

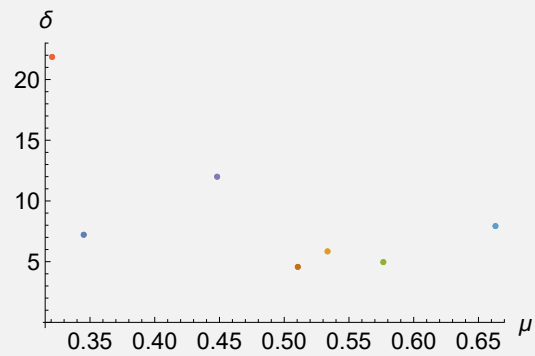
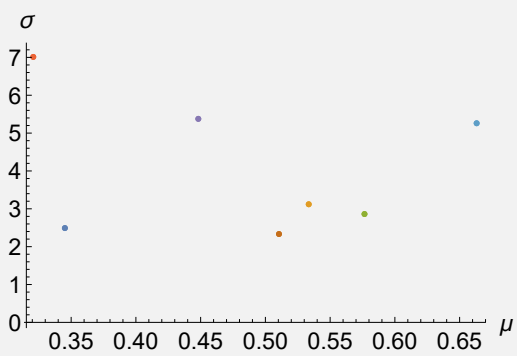
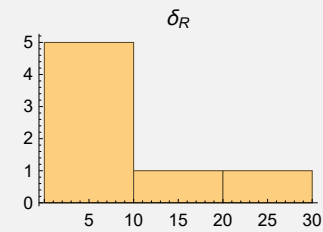
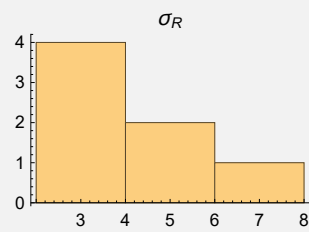
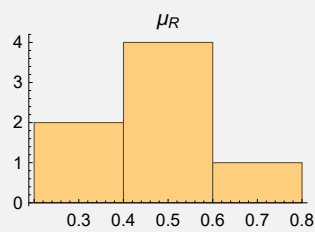
where:

$$\boldsymbol{\Sigma}_{\mathbf{X}} = E((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

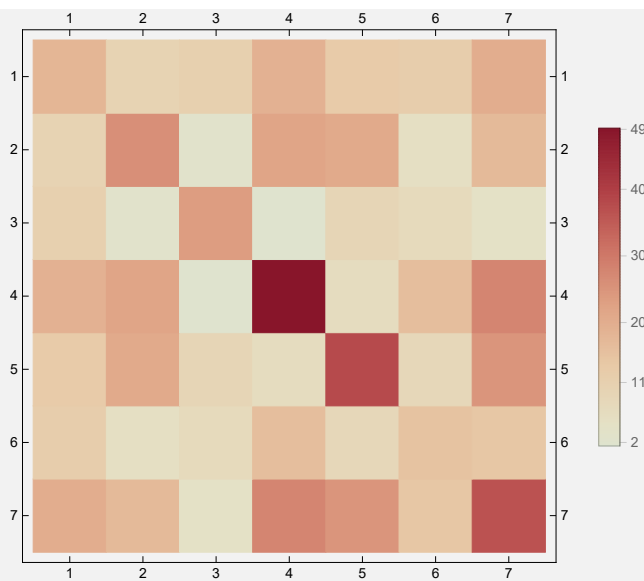
is the covariance matrix for the stock returns. The idea is to find that the weights a_i that minimizes the variance while maximizing the value of U .

Consider the following tech stocks weekly return data for 2016-2017:

	stock	μ_i	σ_i	δ_i
1	GOOGL	0.35	2.49	7.22
2	AAPL	0.53	3.12	5.85
3	FB	0.58	2.86	4.96
4	TWTR	0.32	7.01	21.86
5	TSLA	0.45	5.38	12.
6	MSFT	0.51	2.33	4.57
7	NFLX	0.66	5.26	7.93



$\Sigma_{\mathbf{X}} =$



$$\Sigma_{\mathbf{X}} =$$

6.2	4.	4.1	6.5	4.6	4.6	6.5
4.	9.7	1.4	6.8	6.6	3.2	6.
4.1	1.4	8.2	1.2	4.	3.7	3.2
6.5	6.8	1.2	49.2	3.7	5.8	11.6
4.6	6.6	4.	3.7	28.9	3.9	9.5
4.6	3.2	3.7	5.8	3.9	5.4	5.1
6.5	6.	3.2	11.6	9.5	5.1	27.7

- if your portfolio only has stocks from Netflix and Facebook, what are the weights that minimize the variance?
- what is the expected value, variance and coefficient of variation of the return of the portfolio in part a)?
- repeat a) and b) for Apple and Twitter?
- which portfolio would you recommend buying, why?

Solution:

- if your portfolio only has stocks from Netflix and Facebook, what are the weights that minimize the variance?

$$\begin{aligned}
 V(U) &= a_3^2 V(X_3) + a_7^2 V(X_7) + 2a_3 a_7 \text{Cov}(X_3, Y_7) \\
 &= a_3^2 \cdot 8.2 + (1 - a_3)^2 \cdot 27.7 + 2a_3(1 - a_3) \cdot 3.2 \\
 &= a_3(29.5186a_3 - 48.9958) + 27.666
 \end{aligned}$$

which is minimized at $a_3^* = 0.83$, which implies that $a_7^* = 1 - 0.83 = 0.17$.

- what is the expected value, variance and coefficient of variation of the return of the portfolio in part a)?

Evaluating the respective formulas with weights $a_3^* = 0.83$ and $a_7^* = 0.17$ gives

$$0.59, 2.7, 4.6,$$

respectively.

c) repeat a) and b) for Apple and Twitter?

$$V(U) = a_2(45.2135a_2 - 84.6517) + 49.1704$$

which is minimized at $a_2^* = 0.94$, which implies that $a_4^* = 1 - 0.94 = 0.06$.

The expected value, variance and coefficient of variation are

$$0.52, 9.55, 5.94,$$

resp.

d) which portfolio would you recommend buying, why?

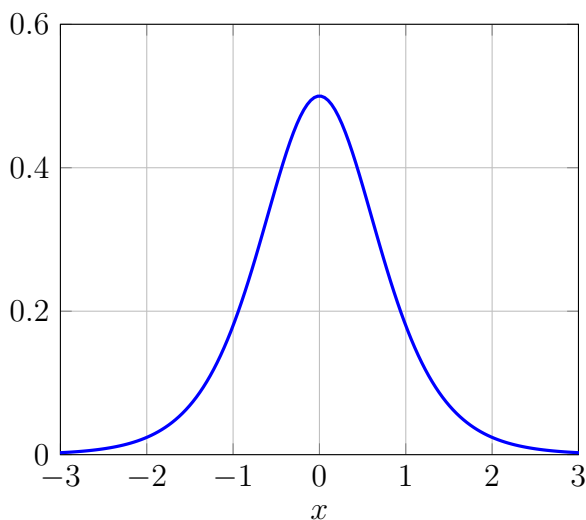
□

3. Continuous Random Variables

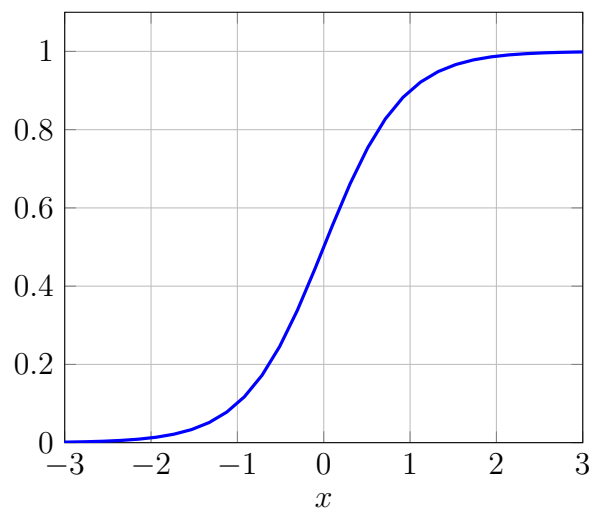
A **continuous** random variable X takes values in an interval of the real line or all of the real line. Therefore, $F_X(x)$ is continuous (with no jumps), which means that

$$P(X = x) = 0 \quad \text{for all } x \quad \text{and} \quad P(X \leq x) = P(X < x)$$

PDF of a continuous rv: $f_X(x)$



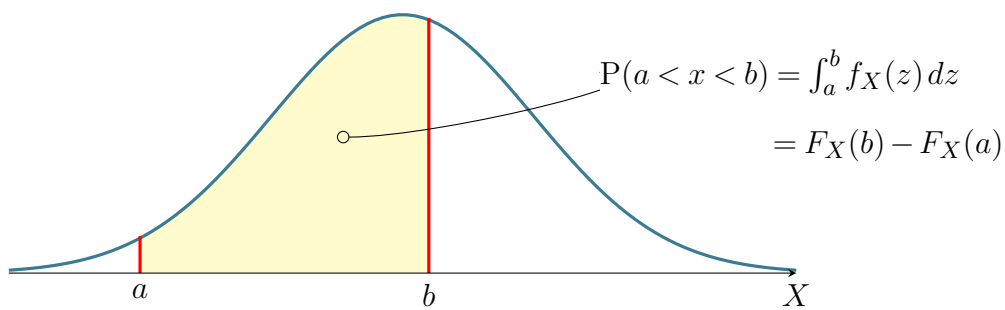
CDF of a continuous rv: $F_X(x)$



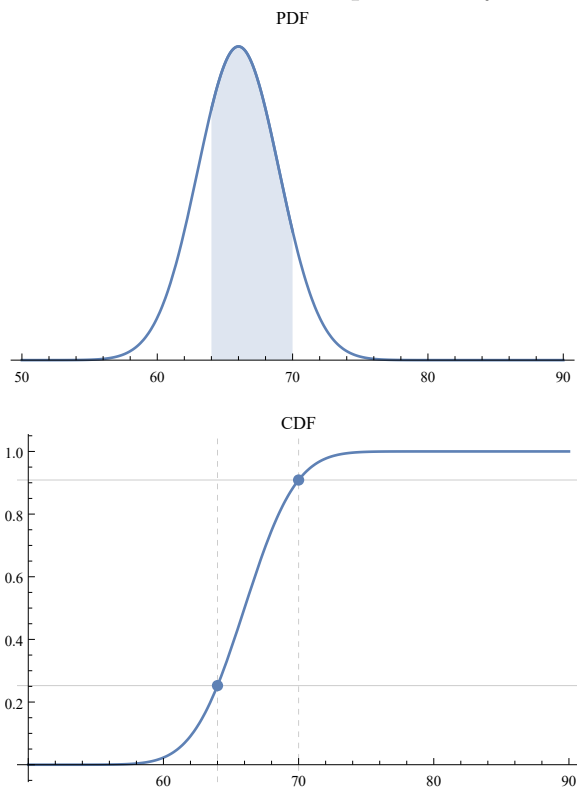
The CDF is still

$$F_X(x) = P(X \leq x) = \int_{-\infty}^x f_X(z) dz$$

The derivative $f_X(x) = F'_X(x)$ is called **probability density function** (PDF).



→ [GeoGebra](#) for interactive probability calculations for the most important distributions.



Note:

$$F_X(x) = \int_{-\infty}^x f_X(z) dz.$$

and

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

Hence any nonnegative function that integrates to one defines a cdf.

The PDF has units Unlike the PMF of a discrete random variable, the PDF has units:

$$\text{units of } f_X(x) = (\text{units of } X)^{-1}$$

This means that $f_X(x)$ it's hard to interpret because it depends on the units of measurement.

The term $f_X(x)dx$ is meaningful because it represents a probability:

$$f_X(x) dx \approx P(x < X < x + dx) \quad (3.1)$$

Expectation For a continuous random variable X the expectation is defined as

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$$

and for any real function g ,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Variance The variance formulas are the same as before:

$$V(X) = E[(X - E(X))^2] = E[X^2] - E(X)^2$$

but using the above definition of expectation.

Example 111. Let X be a continuous variable whose probability density function is

$$f_X(x) = \begin{cases} cx & ; \quad 0 < x < 1 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

- a) Find c .
- b) Find $E(X)$.

Solution:

- a) Find c .

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(x) dx \\ &= c \int_0^1 x dx = c/2 \end{aligned}$$

We get $c = 2$

- b) Find $E(X)$.

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^1 2x^2 dx = \frac{2}{3} \end{aligned}$$

□

Example 112. — * The PDF of the random variable X is,

$$f_X(x) = \begin{cases} k(1 - x^2 + \frac{1}{2}x^3), & 0 < x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

where k is a constant.

- (a) Determine the value of k
 (b) Determine the **mean value**, **variance** and **coefficient of variation** of X
 (c) Determine $P(0.3 < X < 0.9 \mid X > 0.6)$

Solution:

Answer: (a) $k = \frac{3}{4}$ (b) 0.9, 0.39, 0.69 (c) 0.247

(a)

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(x) dx \\ &= \int_0^2 k(1 - x^2 + \frac{1}{2}x^3) dx \end{aligned}$$

We get $k = \frac{3}{4}$

(b) Mean

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f_X(x) dx \\ &= \int_0^2 kx(1 - x^2 + \frac{1}{2}x^3) dx \\ &= \int_0^2 \frac{3}{4}x(1 - x^2 + \frac{1}{2}x^3) dx \\ &= 0.9 \end{aligned}$$

Variance

$$\begin{aligned} E(X^2) &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \\ &= \int_0^2 kx^2(1 - x^2 + \frac{1}{2}x^3) dx \\ &= \int_0^2 \frac{3}{4}x^2(1 - x^2 + \frac{1}{2}x^3) dx \\ &= 1.2 \end{aligned}$$

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= 1.2 - 0.9^2 \\ &= 0.39 \end{aligned}$$

Coefficient of Variation

$$\begin{aligned}\delta_x &= \frac{\sqrt{V(X)}}{E(X)} \\ &= 0.69\end{aligned}$$

(c)

$$\begin{aligned}P(0.3 < X < 0.9 | X > 0.6) &= \frac{P(0.6 < X < 0.9)}{P(X > 0.6)} \\ &= \frac{\int_{0.6}^{0.9} f_X(x) dx}{\int_{0.6}^{\infty} f_X(x) dx} \\ &= \frac{0.554 - 0.408}{1 - 0.408} \\ &= 0.247\end{aligned}$$

□

Example 113. Let X be a continuous variable whose probability density function is

$$f_X(x) = \begin{cases} c(4x - 2x^2) & ; 0 < x < 2 \\ 0 & ; \text{otherwise} \end{cases}$$

- a) Find c .
- b) Find $P(X > 1)$.

Solution:

1. As $f_X(x)$ is a probability density function,

$$\begin{aligned}1 &= \int_{-\infty}^{\infty} f_X(x) dx \\ &= \int_0^2 c(4x^2 - 2x^2) dx \\ &= \frac{8c}{3}\end{aligned}$$

Therefore,

$$c = \frac{3}{8}$$

- 2.

$$\begin{aligned}P(X > 1) &= \int_1^{\infty} f_X(x) dx \\ &= \int_1^2 \frac{3}{8}(4x - 2x^2) dx \\ &= \frac{1}{2}\end{aligned}$$

□

Example 114. — Cauchy distribution Let X be a continuous variable whose probability density function is

$$f_X(x) = \frac{c}{1+x^2} \quad -\infty < x < \infty$$

- Find c .
- Find $F_X(x)$.
- Find $V(X)$.

Solution:

- a) As $f_X(x)$ is a probability density function,

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(x) dx \\ &= c \int_{-\infty}^{\infty} \frac{1}{1+x^2} dx \\ &= c\pi \end{aligned}$$

Therefore,

$$c = \frac{1}{\pi}$$

- b) $F_X(x) = \int_{-\infty}^x \frac{1}{\pi(y^2+1)} dy = \pi \left(\tan^{-1}(x) + \frac{\pi}{2} \right)$
- c) $E(X) = \int_{-\infty}^{\infty} x \cdot \frac{1}{\pi(x^2+1)} dx$ does not converge, and nor does $E(X^2)$, or the variance $V(X)$: the Cauchy distribution is said to be “pathological”.

□

Example 115. — Exponential distribution. The amount of time in hours that a computer functions before breaking down has the distribution

$$f_X(x) = \begin{cases} \lambda e^{-\frac{x}{100}} & ; \quad x \geq 0 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

What is the probability that the computer functions for more than 50 but less than 150 hours?

Solution:

$$\begin{aligned} 1 &= \int_{-\infty}^{\infty} f_X(x) dx \\ &= \lambda \int_0^{\infty} e^{-\frac{x}{100}} dx \\ &= 100\lambda \end{aligned}$$

Therefore,

$$\lambda = \frac{1}{100}$$

Therefore,

$$\begin{aligned} P(50 < x < 150) &= \int_{50}^{150} \lambda e^{-\frac{x}{100}} dx \\ &= \int_{50}^{150} \frac{1}{100} e^{-\frac{x}{100}} dx \\ &\approx 0.384 \end{aligned}$$

□

Example 116. — Uniform distribution. The probability density function of X is given by the Uniform distribution in $(0, 1)$:

$$f_X(x) = \begin{cases} 1 & ; \quad 0 \leq x \leq 1 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

Find $E[e^X]$.

Solution:

$$\begin{aligned} E[e^X] &= \int_{-\infty}^{\infty} e^x f_X(x) dx \\ &= \int_0^1 e^x dx \\ &= e - 1 \end{aligned}$$

□

3.1 Joint Continuous Variables

We already covered the theory of jointly distributed random variables in chapter 2 for the discrete case. Here we simply Use probability density instead of probability mass function, but the equations are identical. The only difference is that we use integrals and not summations.

Joint CDF and PDF X and Y are said to be jointly continuous if there exists a function $f_{X,Y}(x,y)$ defined for all real x and y , such that

$$\begin{aligned} F_{X,Y}(x,y) &= P(-\infty \leq X \leq x, -\infty \leq Y \leq y) \\ &= \int_{-\infty}^y \int_{-\infty}^x f_{X,Y}(a,b) da db \end{aligned}$$

Therefore: $f_{X,Y}(x,y) = \frac{\partial^2}{\partial x \partial y} F(x,y)$.

Note:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$$

Probability of events. An event C is any subset (area or region) of the $X - Y$ plane, and

$$P((X,Y) \in C) = \iint_{(x,y) \in C} f_{X,Y}(x,y) dx dy \quad (3.2)$$

Marginal PDF The PDF of a single random variable is called marginal PDF:

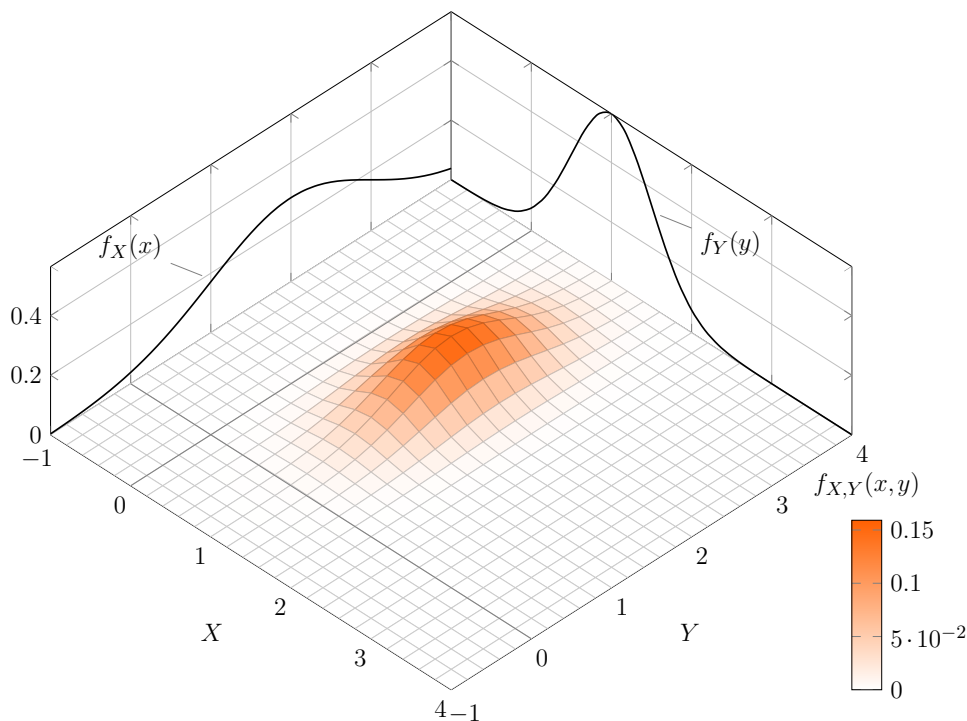
$$f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) dy \quad \text{and} \quad f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y) dx$$

Conditional probability density functions For two continuous random variables X and Y :

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

and the cumulative distribution function is

$$F_{X|Y}(x) = \int_{-\infty}^x f_{X|Y}(x) dx$$



Independence If X and Y are continuous, then

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all (x,y) , if and only if X and Y are independent.

Expectation

$$E(g(X,Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y) dx dy$$

$$E(g(X,Y) | Y = y) = \int_{-\infty}^{\infty} g(x,y) f_{X|Y}(x) dx$$

Notice that:

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f_{X,Y}(x,y) dx dy \\ &= \int_{-\infty}^{\infty} y \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y) dx \right) dy \\ &= \int_{-\infty}^{\infty} y f_Y(y) dy \end{aligned}$$

Covariance The formula for covariance is identical to the discrete case,

$$\begin{aligned} \text{Cov}(X,Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y) \end{aligned}$$

but using the definition of expectation above.

Example 117. — * **Two friends meet?** Two friends try to meet at a certain place between 5 pm and 6 pm. Each person arrives at a time uniformly distributed in the time - interval independently of each other and stays for **20 minutes**.

Find the probability that they meet.

Solution: We have:

$$X \sim U(0,60)$$

$$Y \sim U(0,60)$$

with the Uniform distribution in $(0, 60)$:

$$f_X(x) = \begin{cases} 1/60 & ; \quad 0 \leq x \leq 60 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

and therefore by independence

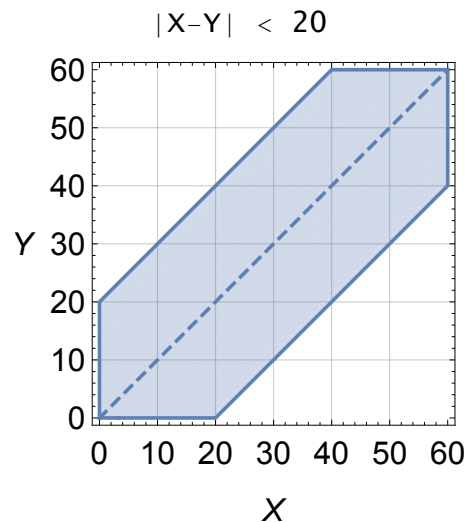
$$\begin{aligned} f_{X,Y}(x,y) &= f_X(x)f_Y(y) \\ &= \begin{cases} 1/60^2 & ; \quad 0 \leq x,y \leq 60 \\ 0 & ; \quad \text{otherwise} \end{cases} \end{aligned}$$

in this particular case.

The event of interest C is $|X - Y| \leq 20$. From the figure we can see that

$$\begin{aligned} P((X,Y) \in C) &= \iint_{(x,y) \in C} f_{X,Y}(x,y) dx dy \\ &= \iint_{(x,y) \in C} \frac{1}{60^2} dx dy \\ &= \text{area of } C / 60^2 \\ &= 0.55 \end{aligned}$$

Notice that the above integral is proportional to the area of C only because $f_{X,Y}(x,y)$ is constant in this particular case; this is not true in general.



□

Example 118. Let

$$X \sim U(0,1)$$

$$Y \sim U(0,1)$$

Calculate the probability density function of $X + Y$

Solution: Let's compute the CDF of $X + Y$ and then take derivatives to obtain the PDF. For the CDF, the event of interest C is $X + Y \leq t$. From the graphical representation of this event in the

X-Y plane we can see that

$$\begin{aligned}
 F_{X+Y}(t) &= P(X+Y \leq t) \\
 &= \iint_{(x,y) \in C} f_{X,Y}(x,y) dx dy \quad \text{but } f_{X,Y}(x,y) = 1, \text{ and from the figure seen in class:} \\
 &= \begin{cases} \frac{t^2}{2} & ; \quad 0 \leq t \leq 1 \\ 1 - \frac{(2-t)^2}{2} & ; \quad 1 \leq t \leq 2 \\ 0 & ; \quad \text{otherwise} \end{cases}
 \end{aligned}$$

Therefore, taking derivatives

$$f_{X+Y}(t) = \begin{cases} t & ; \quad 0 \leq t \leq 1 \\ 2-t & ; \quad 1 < t < 2 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

□

Example 119. — * **An accident** occurs at a point X that is uniformly distributed on a road of length L . At the time of the accident, an ambulance is at a location Y that is also uniformly distributed on the same road.

Assuming that X and Y are independent, find the expected distance $|X - Y|$ between the point of occurrence of the accident, and the position of the ambulance.

Solution:

$$f_X(x) = \begin{cases} \frac{1}{L} & ; \quad 0 < x < L \\ 0 & ; \quad \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{L} & ; \quad 0 < y < L \\ 0 & ; \quad \text{otherwise} \end{cases}$$

As the variables are independent,

$$\begin{aligned}
 f_{X,Y}(x,y) &= f_X(x)f_Y(y) \\
 &= \begin{cases} \frac{1}{L^2} & ; \quad 0 < x < L, 0 < y < L \\ 0 & ; \quad \text{otherwise} \end{cases}
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 E[|X - Y|] &= \frac{1}{L^2} \int_0^L \int_0^L |x - y| \, dy \, dx \\
 &= \frac{1}{L^2} \int_0^L \left(\int_0^x (x - y) \, dy + \int_x^L (y - x) \, dy \right) \, dx \\
 &= \frac{1}{L^2} \int_0^L \left(\frac{L^2}{2} + x^2 - xL \right) \, dx \\
 &= \frac{L}{3}
 \end{aligned}$$

□

This example shows how random variables can have a $\text{Cov}(X, Y) = 0$ but still be dependent:

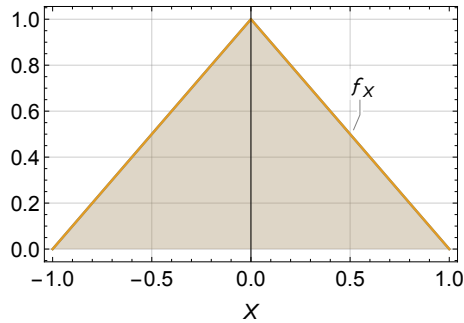
Example 120. Find $\text{Cov}(X, Y)$ for $Y = X^2$ where X has a Triangular Distribution in $(-1, 1)$.

Solution: Let

$$\begin{cases} X \sim \text{Triang}(-1, 1) \\ Y = X^2 \end{cases}$$

with

$$f_X(x) = \begin{cases} 1+x & -1 \leq x \leq 0 \\ 1-x & 0 < x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$



Therefore,

$$\begin{aligned}
 E(X) &= 0 \\
 E(Y) &= E(X^2) = \int_{-1}^1 x^2 f_X(x) dx = \int_{-1}^0 x^2(1+x) dx + \int_0^1 x^2(1-x) dx \\
 &= \int_{-1}^0 x^2 dx + \int_0^1 x^2 dx = 2/3 \\
 E(XY) &= E(X^3) = \int_{-1}^1 x^3 f_X(x) dx = \int_{-1}^0 x^3(1+x) dx + \int_0^1 x^3(1-x) dx \\
 &= \int_{-1}^0 x^3 dx + \int_0^1 x^3 dx = 0
 \end{aligned}$$

Therefore, $\text{Cov}(X, Y) = 0 - 0 = 0$

□

Example 121. The joint density function of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} 2e^{-x}e^{-2y} & ; \quad 0 < x < \infty, 0 < y < \infty \\ 0 & ; \quad \text{otherwise} \end{cases}$$

Compute

1. $P(X > 1, Y < 1)$
2. $P(X < Y)$
3. $P(X < a)$

Solution:

1.

$$\begin{aligned}
 P(X > 1, Y < 1) &= \int_{-\infty}^1 \int_1^{\infty} f_{X,Y}(x,y) dx dy \\
 &= \int_0^1 \int_1^{\infty} 2e^{-x}e^{-2y} dx dy \\
 &= \int_0^1 2e^{-2y} - e^{-x} \Big|_{x=1}^{x=\infty} dy \\
 &= e^{-1} \int_0^1 2e^{-2y} dy \\
 &= e^{-1} (1 - e^{-2})
 \end{aligned}$$

2.

$$\begin{aligned}
P(X < Y) &= \iint_{(x,y):x<y} f_{X,Y}(x,y) dx dy \\
&= \iint_{(x,y):x<y} 2e^{-x} e^{-2y} dx dy \\
&= \int_0^{\infty} \int_0^y 2e^{-x} e^{-2y} dx dy \\
&= \int_0^{\infty} 2e^{-2y} (1 - e^{-y}) dy \\
&= \int_0^{\infty} 2e^{-2y} dy - \int_0^{\infty} 2e^{-3y} dy \\
&= 1 - \frac{2}{3} \\
&= \frac{1}{3}
\end{aligned}$$

3.

$$\begin{aligned}
P(X < a) &= \int_{-\infty}^a \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy dx \\
&= \int_0^a \int_0^{\infty} 2e^{-2y} e^{-x} dy dx \\
&= \int_0^a e^{-x} dx \\
&= 1 - e^{-a}
\end{aligned}$$

□

Example 122. Consider the bivariate density function

$$f(x,y) = \frac{12}{7}(x^2 + xy), \quad 0 \leq x, y \leq 1.$$

Find the probability that $X > Y$.

Solution: The desired probability can be found by integrating f over the region $A = \{(x,y) | 0 \leq y \leq x \leq 1\}$. Note that A is not a rectangle, so we use (??):

$$P(X > Y) = \frac{12}{7} \int_0^1 \int_0^x (x^2 + xy) dy dx = \frac{9}{14}.$$

□

Example 123. — * The joint density function of X and Y is given by

$$f_{X,Y}(x,y) = \begin{cases} e^{-(x+y)} & ; \quad 0 < x < \infty, 0 < y < \infty \\ 0 & ; \quad \text{otherwise} \end{cases}$$

- Are X and Y independent? Explain why.
- Find the density function of the random variable $Z = X/Y$.
- Find the expected value of the function $g(X,Y) = XY$. *Hint:* $\int x e^{-x} = -(x+1)e^{-x}$.

Solution:

- Are X and Y independent? Explain why. Yes because the joint distribution is the product of the marginals:

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

$$f_X(x) = \int_0^{\infty} e^{-(x+y)} dy = e^{-x} \int_0^{\infty} e^{-y} dy = e^{-x}$$

So indeed,

$$e^{-(x+y)} = e^{-x}e^{-y}$$

- Find the density function of the random variable $Z = X/Y$.

$$\begin{aligned} F_{\frac{X}{Y}}(a) &= \mathbb{P}\left(\frac{X}{Y} \leq a\right) \\ &= \iint_{(x,y): \frac{x}{y} \leq a} f_{X,Y}(x,y) dx dy \\ &= \iint_{(x,y): \frac{x}{y} \leq a} e^{-(x+y)} dx dy \\ &= \int_0^{\infty} \int_0^{ay} e^{-(x+y)} dx dy \\ &= \int_0^{\infty} (1 - e^{-ay}) e^{-y} dy \\ &= -e^{-y} + \frac{e^{-(a+1)y}}{a+1} \Big|_0^{\infty} = 1 - \frac{1}{a+1} \end{aligned}$$

Therefore,

$$\begin{aligned} f_{\frac{X}{Y}}(a) &= \frac{dF_{\frac{X}{Y}}(a)}{da} \\ &= \frac{1}{(a+1)^2} \end{aligned}$$

- c) Find the expected value of the function $g(X, Y) = XY$. *Hint:* $\int xe^{-x} = -(x+1)e^{-x}$.
Since X, Y are *independent*:

$$E(XY) = E(X)E(Y) = 1 \times 1$$

□

Example 124. — * **Two random variables** X and Y have the following joint PDF:

$$f_{X,Y}(x,y) = \begin{cases} kx & ; \quad 0 < x < 1, 0 < y < x \\ 0 & ; \quad \text{otherwise} \end{cases}$$

Determine:

- (a) the value of k
 (b) the conditional and marginal PDF of X
 (c) the conditional and marginal PDF of Y
 (d) the standard deviation of X
 (e) the standard deviation of Y
 (f) the correlation coefficient between X and Y
 (g) the mean and variance of the function $g(X, Y) = |X - Y|$

Solution: Answer: (a) 3 (d) 0.193 (e) 0.244 (f) 0.397 (g) 0.375, 0.0594

(a)

$$\iint kx dx dy = 1$$

We get $k = 3$

(b) Marginal PDF of X :

$$\int_0^x 3x dy = 3x^2, 0 < x < 1$$

Marginal PDF of Y :

$$\int_y^1 3x dx = \frac{3}{2}(1 - y^2), 0 < y < 1$$

(c) Conditional PDF of X :

$$\begin{aligned} f_{X|Y} &= \frac{3x}{\frac{3}{2}(1 - y^2)} \\ &= \frac{2x}{1 - y^2}, 0 < x < 1 \end{aligned}$$

Conditional PDF of Y :

$$\begin{aligned} f_{Y|X} &= \frac{3x}{3x^2} \\ &= \frac{1}{x}, 0 < y < 1 \end{aligned}$$

(d)

$$\begin{aligned} E(X) &= \int_0^1 x \cdot 3x^2 dx \\ &= \frac{3}{4} \end{aligned}$$

$$\begin{aligned} E(X^2) &= \int_0^1 x^2 \cdot 3x^2 dx \\ &= \frac{3}{5} \end{aligned}$$

$$\begin{aligned} \sigma_X &= \sqrt{E(X^2) - E^2(X)} \\ &= 0.193 \end{aligned}$$

(e)

$$\begin{aligned} E(Y) &= \int_0^1 y(1-y^2) dy \\ &= \frac{3}{8} \end{aligned}$$

$$\begin{aligned} E(Y^2) &= \int_0^1 y^2(1-y^2) dy \\ &= \frac{1}{5} \end{aligned}$$

$$\begin{aligned} \sigma_Y &= \sqrt{E(Y^2) - E^2(Y)} \\ &= 0.244 \end{aligned}$$

(f)

$$\begin{aligned} E(XY) &= \iint xy \cdot 3x dx dy \\ &= \frac{3}{10} \end{aligned}$$

$$\begin{aligned} \rho &= \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \\ &= \frac{E(XY) - E(X)E(Y)}{\sigma_X \sigma_Y} \\ &= 0.397 \end{aligned}$$

(g) Because $0 < y < x$, $|X - Y| = x - y$

$$\begin{aligned} E[|X - Y|] &= E(X - Y) \\ &= \frac{3}{8} \\ E[|X - Y|^2] &= E[(X - Y)^2] \\ &= \frac{1}{5} \end{aligned}$$

$$\begin{aligned} V[|X - Y|] &= E[|X - Y|^2] - E^2[|X - Y|] \\ &= \frac{19}{320} \\ &= 0.0594 \end{aligned}$$

□

Example 125.

$$f_{X,Y}(x,y) = \begin{cases} \frac{e^{-\frac{x}{y}} e^{-y}}{y} & ; 0 < x < \infty, 0 < y < \infty \\ 0 & ; \text{otherwise} \end{cases}$$

Find $P(X > 1|Y = y)$.

Solution:

$$\begin{aligned} f_{X|Y}(x) &= \frac{f_{X,Y}(x,y)}{f_Y(y)} \\ &= \frac{\frac{e^{-\frac{x}{y}} e^{-y}}{y}}{e^{-y} \int_0^{\infty} \left(\frac{1}{y}\right) e^{-\frac{x}{y}} dx} \\ &= \frac{e^{-\frac{x}{y}}}{y} \end{aligned}$$

Therefore,

$$\begin{aligned} P(X > 1|Y = y) &= \int_1^{\infty} \frac{1}{y} e^{-\frac{x}{y}} dx \\ &= e^{-\frac{1}{y}} \end{aligned}$$

□

3.1.1 Exercises

- Let the rv's X and Y have the joint pdf given below: $f(x,y)=kxy^2$, for $0 \leq x \leq 2$, $x \leq y \leq 3$.
- Find the constant k .
- Find the marginal pdf's of X and Y .

4. Are X and Y independent?

Let the rv's X and Y have the joint pdf given below:

$$f(x, y) = \begin{cases} 2e^{-x-y} & 0 \leq x \leq y < \infty \\ 0 & \text{otherwise} \end{cases}$$

1. Find $P(X + Y \leq 3)$.
2. Find the marginal pdf's of Y and X .
3. Are X and Y independent? Justify your answer.

Let X be the force applied to a randomly selected beam, and Y the time to failure of the beam. Suppose that X is uniformly distributed between 500 and 600 pounds. Suppose also that the conditional pdf of Y given that a force $X = x$ is applied is zero for negative y and $f_{Y|X=x}(y) = \lambda(x)e^{-\lambda(x)y}$, for $y > 0$, where $\lambda(x) = 0.02x - 9.999$.

Find the joint distribution of (X, Y) .

Find the expected time to failure of a randomly selected beam when the force applied is $X = 580$. (*Hint:* Use the formula for the mean value of an exponential random variable.)

A type of steel has microscopic defects which are classified on continuous scale from 0 to 1, with 0 the least severe and 1 the most severe. This is called the defect index. Let X and Y be the static force at failure and the defect index for a particular type of structural member made of this steel. For a member selected at random, these are jointly distributed random variables with joint pdf

$$f(x, y) = \begin{cases} 24x & \text{if } 0 \leq y \leq 1 - 2x \text{ and } 0 \leq x \leq .5 \\ 0 & \text{otherwise} \end{cases}$$

1. Draw the support of this pdf, i.e. the region of (x, y) values where $f(x, y) > 0$.
2. Are X and Y independent? Answer this question without first computing the marginal pdfs of X and Y . Justify your answer.
3. Find each of the following: f_X , f_Y , $E(X)$, and $E(Y)$.
4. Find the conditional pdf of Y given $X = x$.
5. Use the conditional pdf found above to calculate $E(Y|X = 0.3)$.

John and his trainer Yvonne have agreed to meet between 6 A.M. and 8 A.M. for a workout but will aim for 6 A.M. Let $X = \#$ of hours that John is late, and $Y = \#$ of hours that Yvonne is late. Suppose that the joint distribution of X and Y is

$$f(x, y) = \begin{cases} \frac{1}{4} & 0 \leq x \leq 2, 0 \leq y \leq 2 \\ 0 & \text{elsewhere} \end{cases}$$

1. Determine the marginal probability density function of X . Do the same for Y . [If you can, guess the marginal pdf of Y without any additional calculations.]
2. Compute $E(X)$ and $\text{Var}(X)$. Do the same for Y . [If you can, guess the $E(Y)$ and $\text{Var}(Y)$ without any additional calculations.]
3. Are X and Y independent? Justify your answer.

4. Special Distributions

4.1 Uniform Random Variable

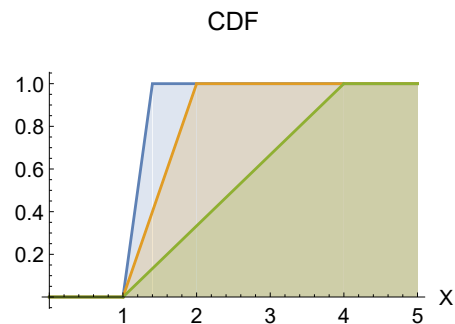
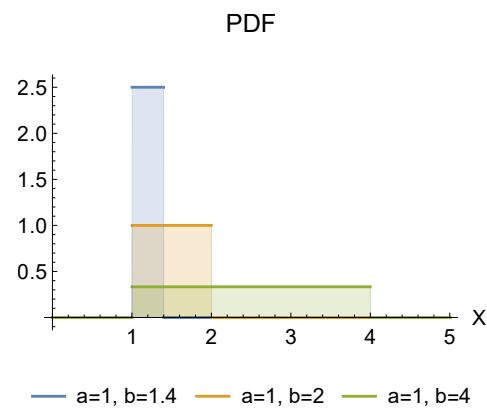
Uniform Random Variable over the interval (a,b) denoted as $X \sim U(a,b)$

$$f_X(x) = \begin{cases} \frac{1}{b-a} & ; a < x < b \\ 0 & ; \text{otherwise} \end{cases}$$

$$F_X(x) = \begin{cases} 0 & ; x < a \\ \frac{x-a}{b-a} & ; a \leq x \leq b \\ 1 & ; b < x \end{cases}$$

$$E[X] = \frac{a+b}{2}$$

$$V(X) = \frac{(b-a)^2}{12}$$



Example 126. Let

$$X \sim U(a,b)$$

Show that

$$E[X] = \frac{a+b}{2}$$

Solution:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f(x) dx \\ &= \int_a^b x \frac{1}{b-a} dx = \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{a+b}{2} \end{aligned}$$

□

Example 127. Let

$$X \sim U(a, b)$$

Show that

$$V(X) = \frac{(b-a)^2}{12}$$

Solution:

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f(x) dx = \int_a^b x^2 \frac{1}{b-a} dx = \frac{b^3 - a^3}{3(b-a)} \\ &= \frac{a^2 + ab + b^2}{3} \end{aligned}$$

Therefore,

$$\begin{aligned} V(X) &= E[X^2] - E[X]^2 \\ &= \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

□

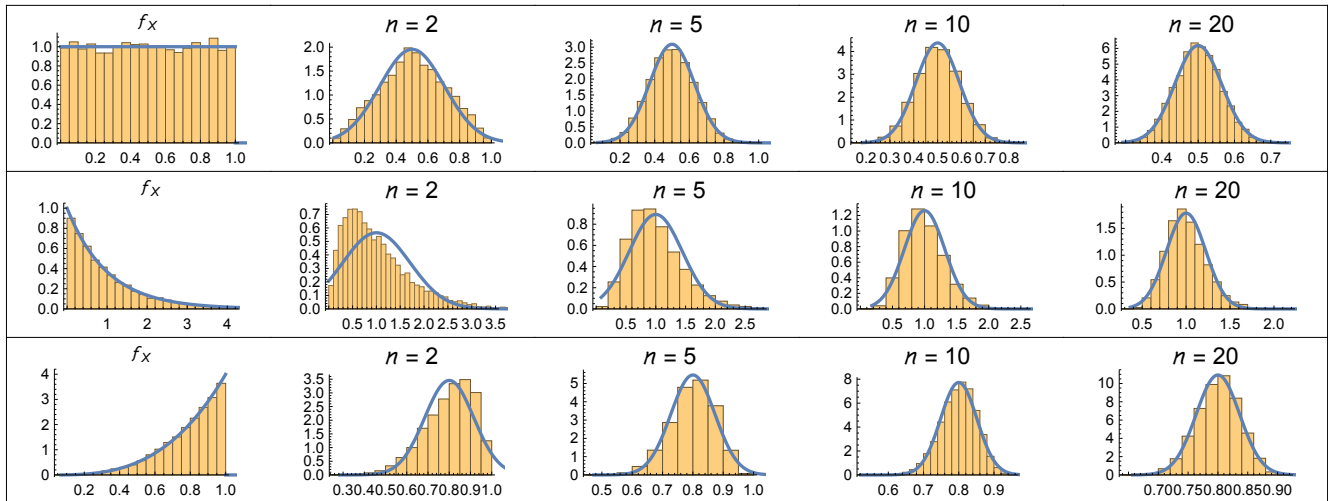
4.2 Normal Distribution

The normal distribution is arguably the most important distribution in statistics. It arises in nature all the time due to the **Central Limit Theorem**. For instance, the CLT implies that the average of random variables

$$U = \sum_{i=1}^n X_i$$

tends to the normal distribution regardless of the distribution of the X_i 's, as illustrated in the following figure.

The figure below shows the agreement of the CLT for the PDF of $U = \frac{1}{n} \sum_{i=1}^n X_i$ where the $X_i \sim f_X$.



It can be seen that regardless of the initial distribution f_X that CLT provides a good approximation for $n > 5$.

Normal Random Variable (aka **Gaussian** rv) A random variable X is said to be a normal random variable,

$$X \sim N(\mu, \sigma^2)$$

if its probability density function is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.1}$$

where μ and σ^2 are parameters, and

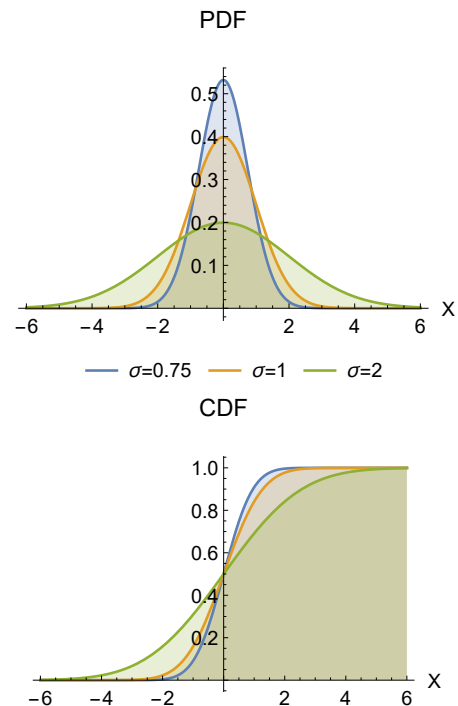
$$E(X) = \mu$$

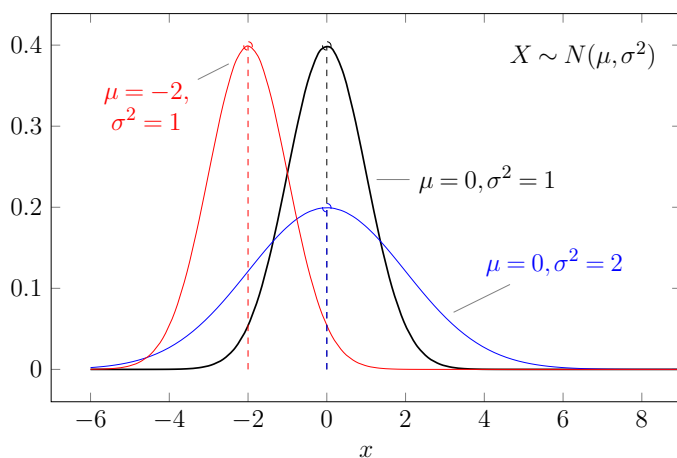
$$V(X) = \sigma^2$$

The CDF of a normal rv :

$$F_X(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}} da$$

does not have an analytical solution and has to be approximated numerically.





STOP! The parameter μ indicates the “location”, and the parameter σ is a “scale” parameter, determining how far it reaches from left to right.

Standard Normal Random Variable A random variable Z is said to be a standard normal random variable if $\mu = 0$ and $\sigma^2 = 1$: $Z \sim N(0, 1)$

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

By convention, the standard normal CDF is **denoted** $\Phi(z)$ and not F_Z :

$$\begin{aligned} \Phi(z) &= P(Z \leq z) \\ &= \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-a^2/2} da \end{aligned}$$

which does not have an analytical solution. It has been approximated with numerical integration and tabulated in **normal probability tables**.

→ [GeoGebra](#) for interactive probability calculations.

Some properties:

1. The normal curve is bell-shaped and is symmetric about the mean, so the mean, median, and mode are equal
2. The normal curve approaches, but never touches, the x -axis.

Fact 4.1 If $X \sim N(\mu, \sigma^2)$, then

$$Y = aX + b$$

is normally distributed with parameters $a\mu + b$ and $a^2\sigma^2$: $Y \sim N(a\mu, a^2\sigma^2)$.

Proof.

$$F_Y(y) = P(Y \leq y) = P(aX + b \leq y) = P\left(X \leq \frac{y-b}{a}\right) = F_X\left(\frac{y-b}{a}\right)$$

Therefore, differentiating,

$$\begin{aligned} f_Y(y) &= \frac{1}{a} f_X\left(\frac{y-b}{a}\right) = \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{\left(\frac{y-b}{a}-\mu\right)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}a\sigma} e^{-\frac{(y-(a\mu+b))^2}{2(a\sigma)^2}} \end{aligned}$$

which corresponds to the normal PDF (4.1) with parameters $a\mu + b$ and $a^2\sigma^2$. ■

Fact 4.2 — z-scores. If $X \sim N(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma}$$

is normally distributed with parameters 0 and 1: $Z \sim N(0, 1)$.

Note that quantiles are related by:

$$x_\alpha = \mu + z_\alpha \sigma \tag{4.2}$$

where $z_\alpha = \Phi^{-1}(\alpha)$ from the table.

Fact 4.3 Let Z be a standard normal random variable. Then,

$$\Phi(-z) = 1 - \Phi(z)$$

4.2.1 The Central Limit Theorem for sums

We can generalize the previous facts in the very important central limit theorem:

Fact 4.4 — The Central Limit Theorem (CLT). The linear combination

$$U = \sum_{i=1}^n a_i X_i = \mathbf{a}^T \mathbf{X}$$

with $\mathbf{a} = (a_1, \dots, a_n)^T$ **tends to the normal distribution as $n \rightarrow \infty$** with

$$E(U) = \sum_{i=1}^n a_i \mu_i = \mathbf{a}^T \boldsymbol{\mu},$$

$$V(U) = \mathbf{a}^T \boldsymbol{\Sigma}_X \mathbf{a}$$

$$= \sum_{i=1}^n a_i^2 \sigma_i^2 + \underbrace{2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \sigma_{i,j}}_{0 \text{ if } X_i \text{'s are independent}}$$

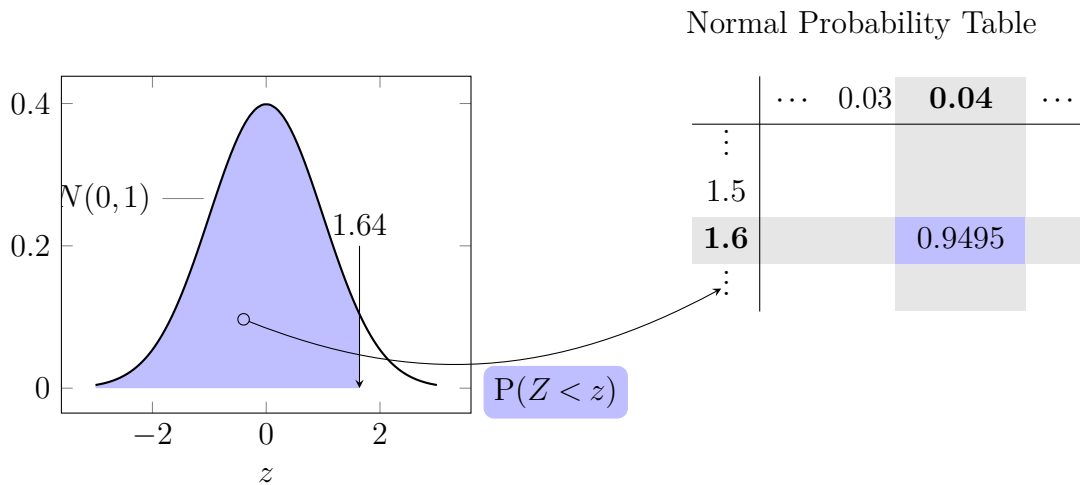
where

$$\Sigma_{\mathbf{X}} = E((\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})^T) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is the Covariance matrix.

4.2.2 How to read normal probability tables

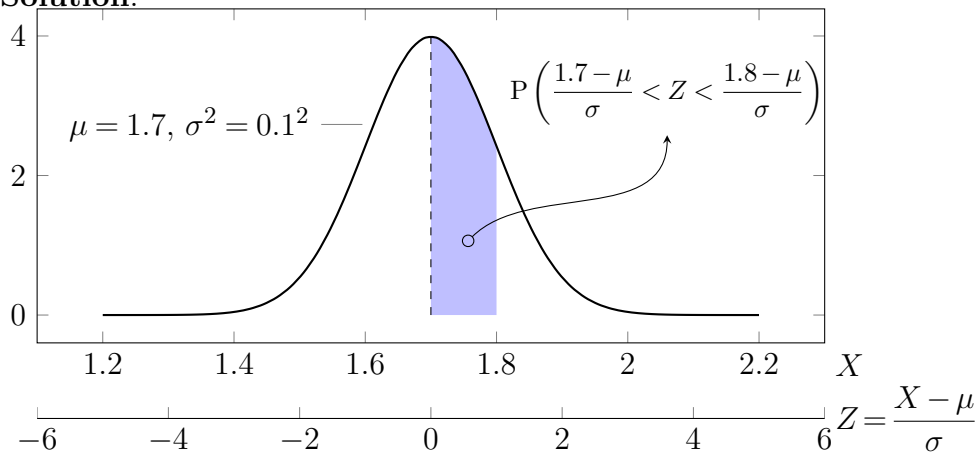
→ Download a [normal probability tables](#).



→ normal probability calculation with the TI 83/84.

Example 128. If $X \sim N(\mu, \sigma^2)$ with $\mu = 1.7$, $\sigma^2 = 0.1^2$, calculate $P(1.7 < X < 1.8)$

Solution:



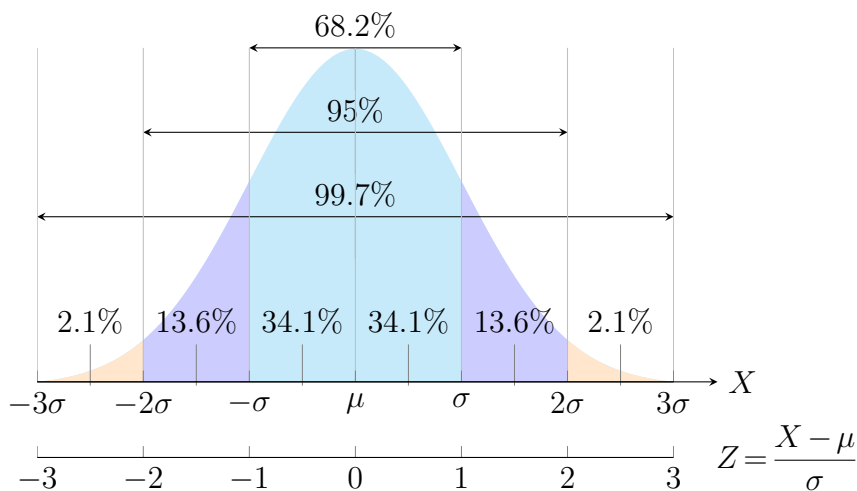
$$\begin{aligned}
 P(1.7 < X < 1.8) &= P\left(\frac{1.7 - \mu}{\sigma} < Z < \frac{1.8 - \mu}{\sigma}\right) \\
 &= P(0 < Z < 1) \\
 &= \Phi(1) - \Phi(0) \\
 &= 0.8413 - 0.5 \\
 &= 0.3413
 \end{aligned}$$

□

4.2.3 The “68-95-99.7 Rule”

For a normal random variable,

- Approximately 68% of the values lie within one standard deviation of the mean.
- Approximately 95% of the values lie within two standard deviations of the mean.
- Approximately ALL (99.7%) of the values lie within three standard deviations of the mean.



Example 129. With $\mu = 5$ and $\sigma = 1$, the Rule says that about 95% lie between $\mu - 2\sigma$ and $\mu + 2\sigma$, which is the interval from 3 to 7.

Example 130. The length of time required to complete a college test is found to be normally distributed with mean 50 minutes and standard deviation 12 minutes.

- When should the test be terminated if we wish to allow sufficient time for 90% of the students to complete the test?
- What proportion of students will finish the test between 30 and 60 minutes?

Solution: Let X be the length of time to complete the test. Then $Z = \frac{X - 50}{12} \sim N(0, 1)$.

- Need to find the 90th percentile, $x_{0.9} = \mu + z_{0.9}\sigma$, with $z_{0.9} = 1.28$ from the table. So at least $x_{0.9} = 65.36$ minutes should be given.
- $P(30 < X < 60) = P(-1.67 < Z < 0.83) = 0.75$.

□

Example 131. Let X be a normal random variable with parameters

$$\mu = 3$$

$$\sigma^2 = 9$$

Find

a) $P(2 < X < 5)$

b) $P(X > 3)$

Solution:

a) Let

$$Z = \frac{X - \mu}{\sigma}$$

Therefore, Z is a standard normal random variable.

Therefore,

$$\begin{aligned} P(2 < X < 5) &= P\left(\frac{2-3}{3} < \frac{X-3}{3} < \frac{5-3}{3}\right) \\ &= P\left(-\frac{1}{3} < Z < \frac{2}{3}\right) \\ &= \Phi\left(\frac{2}{3}\right) - \Phi\left(-\frac{1}{3}\right) \\ &= \Phi\left(\frac{2}{3}\right) - \left(1 - \Phi\left(\frac{1}{3}\right)\right) \\ &\approx 0.3779 \end{aligned}$$

b) Let

$$Z = \frac{X - \mu}{\sigma}$$

Therefore, Z is a standard normal random variable.

Therefore,

$$\begin{aligned} P(X > 3) &= P\left(\frac{X-3}{3} > \frac{3-3}{3}\right) \\ &= P(Z > 0) \\ &= 0.5 \end{aligned}$$

□

Example 132. A school wishes to accept 2000 students for their freshman class, and they expect 20,000 applications. In order to make their admissions decisions very easy, the only criterion they will use is SAT score. So, their goal is to accept a student if and only if their SAT score is in the top 10%. However, because their computer system is so old, the applications only come in one at a time, and they must decide whether to accept or reject before moving on to the next application. Assuming that SAT scores are normally distributed with a mean of 1000 and a standard deviation of 200, how should they set the score threshold to end up with as close to

2000 students as possible? Give your answer first symbolically (in terms of a pdf, cdf, etc), then use a normal distribution table to provide a numerical answer.

Solution: We want to find the SAT score x such that 90% of scores are below x and 10% of scores are above x .

We look through the table for the value closest to 0.90, and find that in a standard normal distribution, $P(X \leq 1.28) = 0.8997$ and $P(X \leq 1.29) = 0.9015$. We'll use the first value because its probability is closer to 0.9.

Hence, the cutoff should be placed 1.28 standard deviations above the mean. This is

$$1000 + 200(1.28) = 1256 \text{ points.}$$

□

Example 133. — * **Traffic congestion** occurs when the demand exceeds the capacity of the system. The current airplane traffic demand at an airport (number of takeoffs and landings per hr) during the peak hours of each day is a normal variate with a **mean** of 200 planes and a **standard deviation** of 50 planes.

(a) If the present runway capacity (for landings and take-offs) is 350 planes per hour, what is the current probability of traffic congestion at this airport? Assume that there is one peak hour per day

(b) If the mean traffic demand increases 10% each year, with the c.o.v. remaining constant, what would be the probability of congestion at the airport in 10 yrs?

(c) If the projected growth of traffic demand is correct, what airport capacity will be required in 10 yr to maintain the current probability of congestion?

Solution: Answer: (a) 0.00135 (b) 0.69146 (c) 700

(a)

$$\begin{aligned} P(\text{congestion}) &= 1 - \phi\left(\frac{350 - 200}{50}\right) \\ &= 0.00135 \end{aligned}$$

(b) Mean: $200 \times (1 + 10\% \times 10) = 400$

standard deviation: $\frac{50 \cdot 400}{200} = 100$

$$\begin{aligned} P(\text{congestion}) &= 1 - \phi\left(\frac{350 - 400}{100}\right) \\ &= 0.69146 \end{aligned}$$

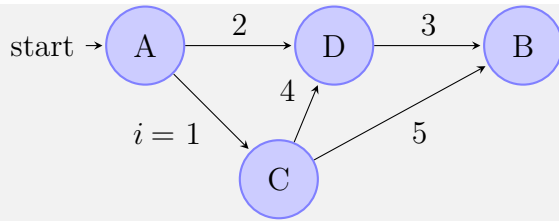
(c) New capacity:

$$\begin{aligned} \text{Capacity} &= 400 + 3 \times 100 \\ &= 700 \end{aligned}$$

□

Example 134. — **Travel time from point A and to point B** In the transportation network below, let

$$X_i = \text{the travel time on link } i = 1, 2, \dots, 5$$



From historical records we have good estimations of means, variances and co-variances:

i	μ_i	σ_i	δ_i
1	10.	2.	0.2
2	11.	3.3	0.3
3	11.	3.3	0.3
4	4.	0.8	0.2
5	10.	2.	0.2

$$\Sigma_{\mathbf{x}} = \begin{pmatrix} 4. & 0.66 & 0.66 & 0.16 & 0.4 \\ 0.66 & 10.89 & 7.62 & 0.26 & 0.66 \\ 0.66 & 7.62 & 10.89 & 1.06 & 0.66 \\ 0.16 & 0.26 & 1.06 & 0.64 & 0.16 \\ 0.4 & 0.66 & 0.66 & 0.16 & 4. \end{pmatrix}$$

- what is the fastest route from A to B?
- what is the probability that route A D B is faster than A C B?
- what is the probability that route A D B is faster than A C D B?

Solution:

- what is the fastest route from A to B?

Let:

Route 1: $A \rightarrow D \rightarrow B$

Route 2: $A \rightarrow C \rightarrow B$

Route 3: $A \rightarrow C \rightarrow D \rightarrow B$

T_i = Travel time on route $i = 1, 2, 3$.

Then,

$$T_1 = X_2 + X_3$$

$$T_2 = X_1 + X_5$$

$$T_3 = X_1 + X_3 + X_4$$

Since travel times are the sum of normal random variables,

$T_i \sim N(E(T_i), V(T_i))$ with:

$$E(T_1) = \mu_2 + \mu_3 = 22$$

$$E(T_2) = \mu_1 + \mu_5 = 20$$

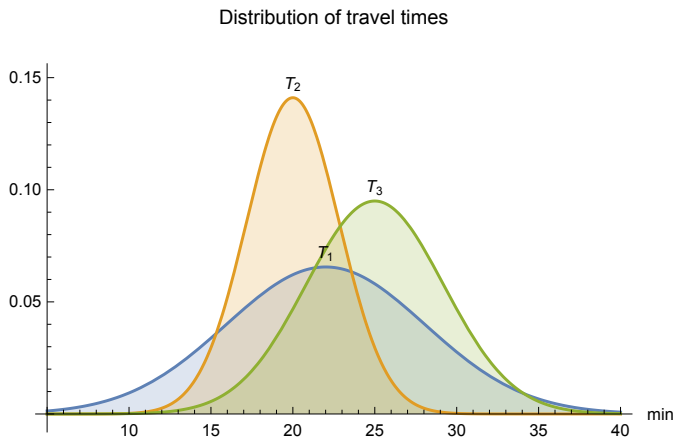
$$E(T_3) = \mu_1 + \mu_3 + \mu_4 = 25$$

and

$$V(T_1) = 2\sigma_{2,3} + \sigma_2^2 + \sigma_3^2 = 37.$$

$$V(T_2) = 2\sigma_{1,5} + \sigma_1^2 + \sigma_5^2 = 8.8$$

$$V(T_3) = 2\sigma_{1,4} + 2\sigma_{1,3} + 2\sigma_{4,3} + \sigma_1^2 + \sigma_4^2 + \sigma_3^2 = 19.28$$



Route 2 is the fastest on average, but for risk-taking people Route 1 could be beneficial.
 b) what is the probability that Route 1: A D B is faster than Route 2: A C B?

$$\begin{aligned}
 P(T_1 < T_2) &= P(T_1 - T_2 < 0), \quad \text{let } Y = T_1 - T_2 \\
 &= P(Y < 0) = 0.376726
 \end{aligned}$$

Since Y is a linear combination of $T_1 - T_2$, it is also normally distributed $Y \sim N(E(Y), V(Y))$ with:

$$\begin{aligned}
 E(Y) &= E(T_1) - E(T_2) &&= 2 \\
 V(Y) &= V(T_1) + V(T_2) - 2\text{Cov}(T_1, T_2) &&= 40.54
 \end{aligned}$$

make sure you understand the negative sign!

$$\text{Cov}(T_1, T_2) = \sigma_{1,2} + \sigma_{1,3} + \sigma_{2,5} + \sigma_{3,5} = 2.64$$

from the covariance matrix.

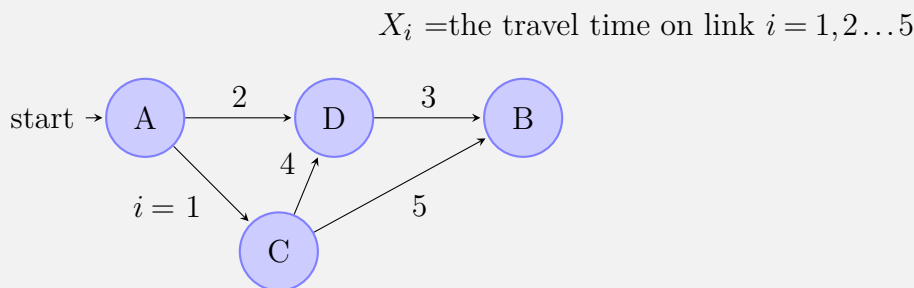
c) what is the probability that Route 1 is faster than Route 3: A C D B?
 Similar to the previous answer, but now:

$$\text{Cov}(T_1, T_3) = \sigma_{1,2} + \sigma_{1,3} + \sigma_{2,3} + \sigma_{2,4} + \sigma_{3,4} + \sigma_3^2 = 21.2$$

$$\text{and } P(T_1 - T_3 < 0) = 0.788644$$

□

Example 135. — Travel time from point A and to point B (again) In the transportation network below, let



From historical records we have good estimations of means, variances and co-variances:

i	μ_i	σ_i	δ_i
1	10.	1.	0.1
2	12.	3.6	0.3
3	13.	3.9	0.3
4	4.	0.4	0.1
5	10.	1.	0.1

$$\Sigma_{\mathbf{X}} = \begin{pmatrix} 1. & 0. & 0. & 0. & 0. \\ 0. & 12.96 & 9.83 & 0. & 0. \\ 0. & 9.83 & 15.21 & 0.62 & 0. \\ 0. & 0. & 0.62 & 0.16 & 0. \\ 0. & 0. & 0. & 0. & 1. \end{pmatrix}$$

- what is the fastest route from A to B?
- what is the probability that route A D B is faster than A C B?
- what is the probability that route A D B is faster than A C D B?

Solution:

- what is the fastest route from A to B?

Let:

Route 1: $A \rightarrow D \rightarrow B$

Route 2: $A \rightarrow C \rightarrow B$

Route 3: $A \rightarrow C \rightarrow D \rightarrow B$

$T_i =$ Travel time on route $i = 1, 2, 3$.

Then,

$$T_1 = X_2 + X_3$$

$$T_2 = X_1 + X_5$$

$$T_3 = X_1 + X_3 + X_4$$

Since travel times are the sum of normal random variables,

$T_i \sim N(E(T_i), V(T_i))$ with:

$$E(T_1) = \mu_2 + \mu_3 = 25$$

$$E(T_2) = \mu_1 + \mu_5 = 20$$

$$E(T_3) = \mu_1 + \mu_3 + \mu_4 = 27$$

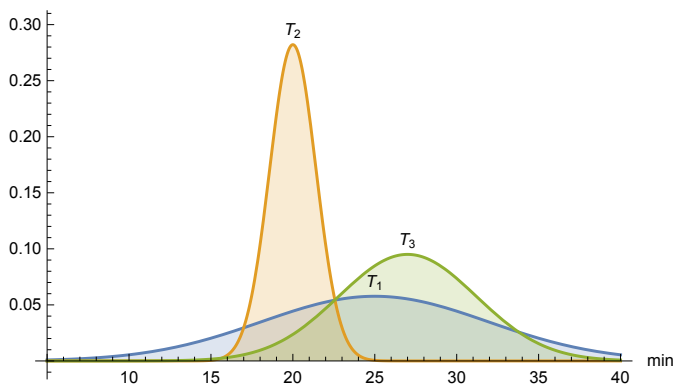
and

$$V(T_1) = 2\sigma_{2,3} + \sigma_2^2 + \sigma_3^2 = 47.8$$

$$V(T_2) = 2\sigma_{1,5} + \sigma_1^2 + \sigma_5^2 = 2$$

$$V(T_3) = 2\sigma_{1,4} + 2\sigma_{1,3} + 2\sigma_{4,3} + \sigma_1^2 + \sigma_4^2 + \sigma_3^2 = 17.6$$

Distribution of travel times



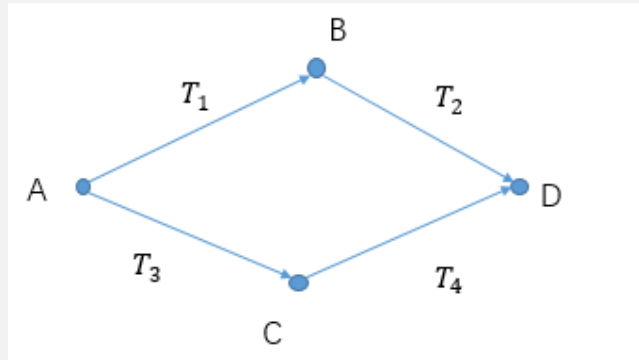
Route 2 is the fastest on average, but for risk-taking people Route 1 could be beneficial.

b) $P(T_1 - T_2 < 0) = 0.239367$

c) $P(T_1 - T_3 < 0) = 0.702722$

□

Example 136. — * **Bob and John** are traveling from city A to city D. Bob decides to take the upper route (through B), whereas John takes the lower route (through C) as shown in the following figure:



The travel times (in hours) between the cities indicated are normally distributed as follows:

$$T_1 \sim N(8, 4)$$

$$T_2 \sim N(5, 1)$$

$$T_3 \sim N(5, 4)$$

$$T_4 \sim N(7, 4)$$

Although the travel times can generally be assumed to be statistically independent, T_3 and T_4 are dependent with a correlation coefficient of 0.8.

(a) What is the probability that John will not arrive in city D within 12 hours?

(b) What is the probability that Bob will arrive in city D earlier than John by at least 1 hour?

(c) Which route (upper or lower) should be taken if one wishes to minimize the expected travel time from A to D? Justify.

Solution: (a)

Let T_J be John's travel time in hours:

$$T_J = T_3 + T_4$$

$$\mu_{T_J} = 5 + 7 = 12$$

$$\begin{aligned} \sigma_{T_J} &= \sqrt{4 + 4 + 2 \times 0.8 \times 2 \times 2} \\ &= 3.795 \end{aligned}$$

Hence

$$\begin{aligned} P(T_J < 12) &= \Phi\left(\frac{12 - 12}{3.795}\right) \\ &= \Phi(0) \\ &= 0.5 \end{aligned}$$

(b)

Let T_B be Bob's travel time in hours:

$$\begin{aligned} T_B &= T_1 + T_2 \\ \mu_{T_B} &= 8 + 5 = 13 \\ \sigma_{T_B} &= \sqrt{4 + 1} \\ &= \sqrt{5} \end{aligned}$$

Hence

$$P(T_J - T_B > 1) = P(T_B - T_J + 1 < 0)$$

Now let $R = T_B - T_J + 1$; R is normal with

$$\begin{aligned} \mu_R &= \mu_{T_B} - \mu_{T_J} + 1 = 2 \\ \sigma_R &= \sqrt{\sigma_{T_B}^2 + \sigma_{T_J}^2} \\ &= \sqrt{19.4} \end{aligned}$$

Hence

$$\begin{aligned} P(R < 0) &= \Phi\left(\frac{0 - 2}{\sqrt{19.4}}\right) \\ &= \Phi(-0.454) \\ &= 0.326 \end{aligned}$$

(c) Since the lower route (A-C-D) has a smaller expected travel time and variance one could take the lower route to minimize expected travel time from A to D. But a risk-seeking person might want to take the longer route with higher variance. \square

Example 137. — * **The daily revenue** X of a store is the sum of the amounts paid by each customer i, Y_i during one day. These amounts Y_i have a mean and variance of \$15 and $(\$15)^2$.
 (a) Write down an equation relating X and the amounts paid by each customer during one day. (5 points)
 (b) On a given day, 100 customers purchased items in the store. Approximate the probability that the daily revenue exceeded 1250. (15 points)

Solution: (a) Let: n : Total number of customers

Y_i : Amount paid by each customer

Equation relating X and the amounts paid by each customer during one day:

$$X = \sum_{i=1}^n Y_i$$

(b) According to the question, Y_i follows exponential distribution:

$$\begin{aligned} \mu_{Y_i} &= 15 \\ \sigma_{Y_i} &= 15 \end{aligned}$$

Because $X = \sum_{i=1}^{100} Y_i$, according to CTL, X follows normal distribution with:

$$\begin{aligned} \mu_X &= 15 \times 100 \\ &= 1500 \\ \sigma_X &= 15 \times \sqrt{100} \\ &= 150 \end{aligned}$$

The probability that the daily revenue exceeded 1250:

$$\begin{aligned} P(X > 1250) &= 1 - \Phi\left(\frac{1250 - \mu_X}{\sigma_X}\right) \\ &= 1 - \Phi(-1.67) \\ &= 0.953 \end{aligned}$$

□

4.3 Lognormal Distribution

$$X \sim \text{LogN}(\lambda, \xi^2), X > 0 \leftrightarrow \boxed{\log X \sim N(\lambda, \xi^2)}$$

$$\lambda = E(\log X) \quad , \quad \xi^2 = V(\log X)$$

The PDF is

$$f(x) = \frac{1}{\sqrt{2\pi\xi x}} e^{-(\log(x)-\lambda)^2/2\xi^2}, x > 0$$

The mean and variance are:

$$E(X) = e^{\lambda + \xi^2/2}$$

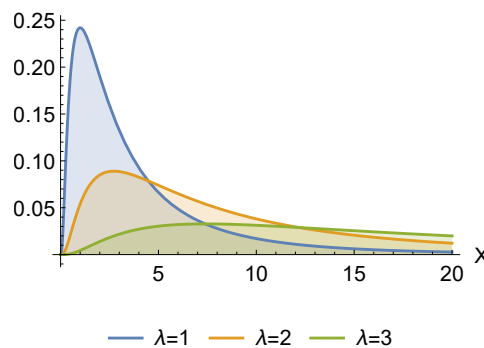
$$V(X) = e^{2\lambda + \xi^2} (e^{\xi^2} - 1)$$

and hence the coefficient of variation squared is

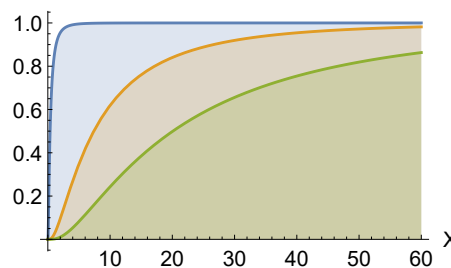
$$\begin{aligned} \delta_X^2 &= e^{\xi^2} - 1 \\ &\approx \xi^2 \quad \text{when } \xi^2 \text{ is small, say } \xi < 1/3. \end{aligned}$$

Note: $\lambda = \log x_{0.5}$

PDF, ($\xi = 1$)



CDF



STOP! The parameters λ and ξ are generally not given and need to be calculated first.

Typically, there are the following situations:

1. if we are given μ_X, σ_X^2 :

$$\delta_X^2 = \sigma_X^2 / \mu_X^2 \tag{4.3}$$

$$\xi^2 = \log(1 + \delta_X^2) \quad \text{or} \quad \xi^2 = \delta_X^2 \quad \text{if } \delta_X^2 \text{ is small, say } < 1/3. \tag{4.4}$$

$$\lambda = \log \mu_X - \xi^2/2 \tag{4.5}$$

2. if we are given $x_{0.5}, \delta_X$:

$$\xi^2 = \log(1 + \delta_X^2) \quad \text{or} \quad \xi^2 = \delta_X^2 \quad \text{if } \delta_X \text{ is small, say } < 1/3. \quad (4.6)$$

$$\lambda = \log x_{0.5} \quad (4.7)$$

Once the parameters λ and ξ are determined, we can calculate probabilities using the standard normal tables:

$$P(X < x) = P(\log X < \log x) = \Phi\left(\frac{\log x - \lambda}{\xi}\right) \quad (4.8)$$

because $\log X \sim N(\lambda, \xi^2)$.

Percentiles x_α can also be obtained from the standard normal percentiles, z_α :

$$x_\alpha = e^{\lambda + z_\alpha \cdot \xi} \quad (4.9)$$

This is because, by definition, $P(X < x_\alpha) = \alpha$, and in this case we have:

$$\begin{aligned} P(X < x_\alpha) &= P(\log X < \log x_\alpha) \\ &= \Phi\left(\frac{\log x_\alpha - \lambda}{\xi}\right) = \alpha \end{aligned}$$

the last equality implies that $\frac{\log x_\alpha - \lambda}{\xi}$ is the standard normal percentile, z_α and (4.9) follows.

Example 138. Lifetimes of a certain component are lognormally distributed with its **median 3** days and parameter $\xi = 0.5$ days.

- Find the **mean** lifetime of these components
- Find the **standard deviation** of the lifetimes

Solution: Answer: (a) 3.40 (b) 1.81

$$(a) \lambda = \log(x_{0.5}) = \log(3)$$

$$\mu_X = e^{\lambda + \frac{1}{2}\xi^2} = 3.40$$

$$(b) \delta_X = \sqrt{e^{\xi^2} - 1} = 0.533$$

$$\sigma_X = \delta_X \mu_X = 0.533 \times 3.40 = 1.81$$

□

Example 139. — Time between inspections. The time T between breakdowns of a major equipment in an oil platform follows a lognormal distribution with a median of 6 months and a coefficient of variation of 30 percent.

What should be the interval t^* between inspections and repairs in order to ensure a 95 % probability that the equipment will be operational at any time.

Solution:

We need $P(T > t^*) = 0.95$ or equivalently $P(T < t^*) = 0.05$ which means that t^* is the 5th percentile:

$$t^* = t_{0.05} = e^{\lambda + z_{0.05} \cdot \xi}$$

which gives 3.66 months.

□

Example 140. — * **An office building** is planned and designed with a lateral load-resisting structural system for earthquake resistance in a seismic zone. The seismic capacity (in term of force factor) of the proposed system has a mean of 6.5 and c.o.v. 29.8% and is assumed to have a Lognormal distribution.

(a) What is the estimated probability of damage to the office building when subjected to 5.5-magnitude earthquake?

(b) If the building survived (without any damage) a previous 4.0-magnitude earthquake, what would be its future probability of no damage under a 5.5-magnitude earthquake? (Assume that after the moderate earthquake the building remains in its original condition)

(c) What is the seismic capacity's 85th percentile ?

Solution: Answer: (a) 0.341 (b) 0.708 (c) 8.47

Random variable X: seismic capacity

Given $\mu_X = 6.5$ and $\delta_X = 0.298$

We have $\sigma_X = \mu_X \delta_X = 1.937$

We have $\delta_X < 0.3$ so,

$$\begin{aligned}\xi &= \delta_X \\ &= 0.298 \\ \lambda &= \log \mu_X - \frac{1}{2}\xi^2 \\ &= 1.827\end{aligned}$$

Or you can do

$$\begin{aligned}\xi &= \sqrt{\log(1 + \delta_X^2)} \\ &= 0.292 \\ \lambda &= \log \mu_X - \frac{1}{2}\xi^2 \\ &= 1.829\end{aligned}$$

(a) The probability of damage on a 5.5-magnitude earthquake is,

$$\begin{aligned}P(X < 5.5) &= \Phi\left(\frac{\log 5.5 - \lambda}{\xi}\right) \\ &= \Phi(-0.41) \\ &= 0.341\end{aligned}$$

(b)

$$\begin{aligned}P(X > 5.5 | X > 4) &= \frac{1 - P(X < 5.5)}{1 - P(X < 4)} \\ &= \frac{1 - 0.341}{1 - \Phi\left(\frac{\log 4 - \lambda}{\xi}\right)} \\ &= 0.708\end{aligned}$$

(c)

$$\begin{aligned}P(X < x_{85}) &= \Phi\left(\frac{\log(x_{85}) - \lambda}{\xi}\right) \\ &= 0.85\end{aligned}$$

Referring to standard normal table: $\Phi(1.04) = 0.85$

$$\begin{aligned} x_{85} &= e^{1.04\xi + \lambda} \\ &= 8.47 \end{aligned}$$

□

4.3.1 The Central Limit Theorem for products

Fact 4.5 For a set of **positive** random variables $X_i, i = 1, 2, \dots, n$, the product

$$U = \prod_{i=1}^n X_i^{a_i}$$

tends to the lognormal distribution as $n \rightarrow \infty$ with parameters:

$$\begin{aligned} \lambda_U &= \mathbf{a}^T \boldsymbol{\lambda} = \sum_{i=1}^n a_i \lambda_i, \\ \xi_U^2 &= \mathbf{a}^T (\boldsymbol{\Sigma}_{\log \mathbf{X}}) \mathbf{a} = \sum_{i=1}^n a_i^2 \xi_i^2 + \underbrace{2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \xi_{i,j}}_{0 \text{ if } X_i \text{'s are independent}} \end{aligned}$$

where

$$\begin{aligned} \mathbf{a} &= (a_1, a_2, \dots, a_n)^T \\ \lambda_i &= E(\log X_i) \\ \boldsymbol{\lambda} &= (\lambda_1, \lambda_2, \dots, \lambda_n)^T \\ \log \mathbf{X} &= (\log X_1, \log X_2, \dots, \log X_n)^T \end{aligned}$$

and:

$$\boldsymbol{\Sigma}_{\log \mathbf{X}} = E\left((\log \mathbf{X} - \boldsymbol{\lambda})(\log \mathbf{X} - \boldsymbol{\lambda})^T\right) = \begin{pmatrix} \xi_1^2 & \xi_{12} & \xi_{13} & \cdots \\ \xi_{21} & \xi_2^2 & \xi_{23} & \cdots \\ \xi_{31} & \xi_{32} & \xi_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is the Covariance matrix of the $\log X_i$, i.e. $\xi_{ij} = \text{Cov}(\log X_i, \log X_j)$ and $\xi_i^2 = V(\log X_i)$.

Note: If the X_i 's have the lognormal distribution, this result is exact, otherwise it is an approximation.

Proof. $\log U = \sum_{i=1}^n a_i \log X_i$ is a linear combination so by the CLT for linear combinations, for large n :

$$\log U \sim N(E(\log U), V(\log U))$$

with:

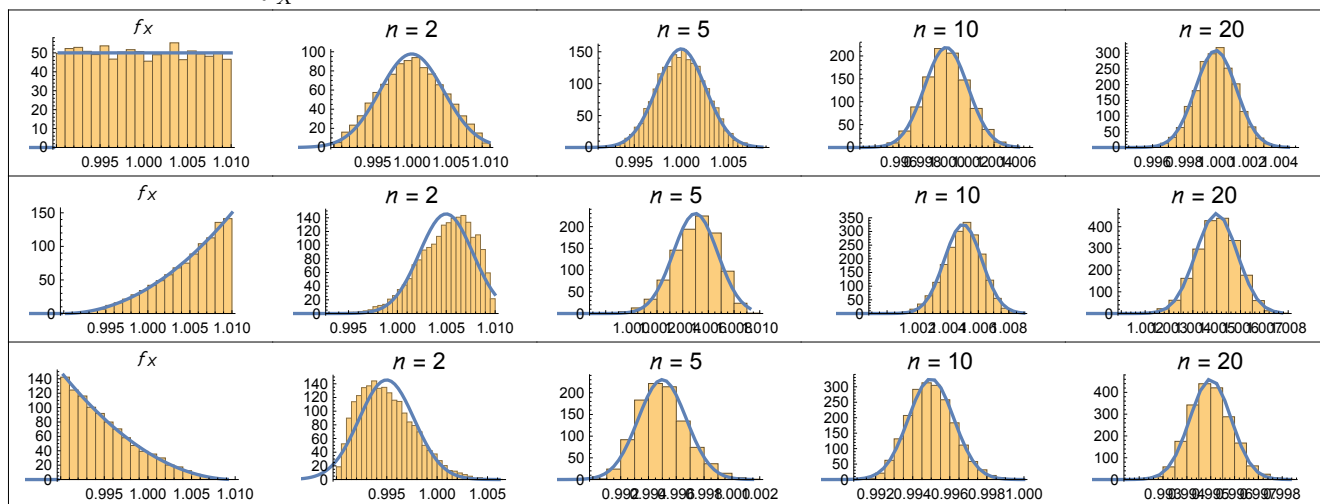
$$\begin{aligned}
 E(\log U) &= \sum_{i=1}^n a_i E(\log X_i) \\
 &= \mathbf{a}^T \boldsymbol{\lambda}, \\
 V(\log U) &= \mathbf{a}^T (\boldsymbol{\Sigma}_{\log \mathbf{X}}) \mathbf{a} \\
 &= \sum_{i=1}^n a_i^2 \xi_i^2 + \underbrace{2 \sum_{i=1}^n \sum_{j=i+1}^n a_i a_j \xi_{i,j}}_{0 \text{ if } X_i \text{'s are independent}}
 \end{aligned}$$

By definition of lognormal random variables we conclude the result. ■

The figure below shows the agreement of the CLT for the distribution of the geometric average of n random variables $X_i \sim f_X$, ie:

$$U = \prod_{i=1}^n X_i^{\frac{1}{n}}$$

for 3 distributions f_X and different values n .



It can be seen that regardless of the initial distribution f_X that CLT provides a good approximation for $n > 5$.

A corollary of Fact 4.5: If $X \sim \text{LogN}(\lambda, \xi^2)$ then

$$Y = cX \sim \text{LogN}(\log c + \lambda, \xi^2)$$

for any constant c . To see this, simply treat the constant as a lognormal rv with zero variance.

Example 141. Let:

$$U = \frac{7.58 \sqrt{X_1} X_3^2}{X_2^3 \sqrt[3]{X_4}}$$

Suppose the X_i 's are independent and that we have the following information:

i	X_i	μ_i	δ_i
1	X_1	7.	0.19
2	X_2	3.	0.05
3	X_3	3.	0.28
4	X_4	1.	0.81

- (a) Approximate the mean and variance of U .
 (b) Approximate $P(U < 94)$

Solution: (a) Using equations (4.3) we can calculate:

i	X_i	μ_i	δ_i^2	σ_i	λ_i	ξ_i^2	a_i
1	X_1	7.	0.036	1.33	1.93	0.035	1/2
2	X_2	3.	0.003	0.15	1.1	0.002	-1
3	X_3	3.	0.078	0.84	1.06	0.075	2
4	X_4	1.	0.656	0.81	-0.25	0.504	-1/3

and according to the CLT for products, we have that U tends to the lognormal distribution parameters:

$$\begin{aligned}\lambda_U &= \log c + \sum_{i=1}^n a_i \lambda_i, \\ &= \log(7.58) + \frac{1.93}{2} - 1.1 + 2 \times 1.06 + \frac{0.25}{3} = 4.0938 \\ \xi_U^2 &= \sum_{i=1}^n a_i^2 \xi_i^2 \\ &= \left(\frac{1}{2}\right)^2 0.035 + (-1)^2 0.002 + 2^2 0.075 + \left(-\frac{1}{3}\right)^2 0.504 = 0.36675\end{aligned}$$

and the mean and variance of U are approximately:

$$\begin{aligned}E(U) &= e^{\lambda_U + \xi_U^2/2} = 72.04 \\ V(U) &= e^{2\lambda_U + \xi_U^2} (e^{\xi_U^2} - 1) = 2299.26\end{aligned}$$

(b) Assuming $U \sim \text{LogN}(\lambda_U, \xi_U^2)$:

$$\begin{aligned}P(U < 94) &= \Phi\left(\frac{\log 94 - \lambda_U}{\xi_U}\right) = \Phi(0.742155) \\ &= 0.771003\end{aligned}$$

□

Example 142. Let:

$$W = \frac{1.46X_1X_4^2}{X_2^2\sqrt{X_3}}$$

Suppose the X_i 's are independent and that we have the following information:

i	X_i	μ_i	δ_i
1	X_1	4.	0.74
2	X_2	1.	0.94
3	X_3	1.	0.65
4	X_4	7.	0.06

(a) Approximate the mean and variance of W .

(b) Approximate $P(W < 4110)$

Solution: (a) Using equations (4.3) we can calculate:

i	X_i	μ_i	δ^2_i	σ_i	λ_i	ξ^2_i	a_i
1	X_1	4.	0.5476	2.96	1.17	0.4367	1
2	X_2	1.	0.8836	0.94	-0.32	0.6332	-2
3	X_3	1.	0.4225	0.65	-0.18	0.3524	$-\frac{1}{2}$
4	X_4	7.	0.0036	0.42	1.94	0.0036	2

and according to the CLT for products, we have that W tends to the lognormal distribution parameters:

$$\lambda_W = \log\left(\frac{1}{2 \cdot 32.2}\right) + \sum_{i=1}^n a_i \lambda_i,$$

$$= 1.09134$$

$$\xi_W^2 = \sum_{i=1}^n a_i^2 \xi_i^2$$

$$= 0.148124$$

and the mean and variance of W are approximately:

$$E(W) = e^{\lambda_W + \xi_W^2/2} = 2190.43$$

$$V(W) = e^{2\lambda_W + \xi_W^2} (e^{\xi_W^2} - 1) = 98,759,027$$

(b) Assuming $W \sim \text{LogN}(\lambda_W, \xi_W^2)$:

$$P(W < 4110) = \Phi\left(\frac{\log 4110 - \lambda_W}{\xi_W}\right) = \Phi(0.0188)$$

$$= 0.507$$

□

Example 143. — * **The hydraulic head loss** in a pipe may be determined by the Darcy-Weisbach equation as follows:

$$H = \frac{fLV^2}{2Dg}$$

where:

L =length of a pipe, V =flow velocity of water in a pipe, D =pipe diameter, f =coefficient of friction, g =gravitational acceleration=32.2 ft/sec². Suppose a pipe has the following properties:

i	X_i	μ_i	δ_i
1	L	100.	0.1
2	D	1.	0.1
3	f	0.02	0.2
4	V	10.	0.15

- (a) Approximate the mean and standard deviation of the hydraulic head loss of the pipe.
 (b) Approximate $P(H < 3\text{ft})$ [Hint: CLT]

Solution: (a) Using equations (4.3) we can calculate:

i	X_i	μ_i	δ_i^2	σ_i	λ_i	ξ_i^2	a_i
1	L	100.	0.01	10.	4.6	0.01	1
2	D	1.	0.01	0.1	0.	0.01	-1
3	f	0.02	0.04	0.004	-3.93	0.0392	1
4	V	10.	0.0225	1.5	2.29	0.0223	2

and according to the CLT for products, we have that H tends to the lognormal distribution parameters:

$$\begin{aligned}\lambda_H &= \log\left(\frac{1}{2 \cdot 32.2}\right) + \sum_{i=1}^n a_i \lambda_i, \\ &= 1.09134 \\ \xi_H^2 &= \sum_{i=1}^n a_i^2 \xi_i^2 \\ &= 0.148124\end{aligned}$$

and the mean and variance of H are approximately:

$$\begin{aligned}E(H) &= e^{\lambda_H + \xi_H^2/2} = 3.20722 \\ V(H) &= e^{2\lambda_H + \xi_H^2} (e^{\xi_H^2} - 1) = 1.64227\end{aligned}$$

(b) Assuming $H \sim \text{LogN}(\lambda_H, \xi_H^2)$:

$$\begin{aligned}P(H < 3) &= \Phi\left(\frac{\log 3 - \lambda_H}{\xi_H}\right) = \Phi(0.0188) \\ &= 0.507\end{aligned}$$

□

Example 144. Repeat the example above with a pipe with the following properties:

i	X_i	μ_i	δ_i
1	L	100.	0.05
2	D	1.	0.15
3	f	0.02	0.25
4	V	9.	0.2

(a) Approximate the mean and standard deviation of the hydraulic head loss of the pipe. (ans:

2.67501, 1.96151)

(b) Approximate $P(H < 3\text{ft})$ (ans: 0.684)

Example 145. — Stock Price Distribution For a given stock, let:

P_t = stock price at time period $t = 1, 2, \dots$

$R_t = (P_t - P_{t-1}) / P_{t-1}$ = rate of return for time period t

$S_t = P_t / P_{t-1} = 1 + R_t$ = return for time period t

Show that the stock price P_t tends to the lognormal distribution. Assume that P_0 is the current stock price, ie a constant.

Solution: Note:

$$\begin{aligned}
 P_t &= P_0(1 + R_1)(1 + R_2) \dots (1 + R_t) \\
 &= P_0 \prod_{j=1}^t S_j
 \end{aligned}$$

By the CLT for products $P_0 \prod_{j=1}^t S_j$ tends to the $\text{LogN}(\lambda_t, \xi_t^2)$ with parameters:

$$\lambda_t = \log[P_0] + \sum_{j=1}^t E[\log S_j], \quad \xi_t^2 = \sum_{j=1}^t V[\log S_j]$$

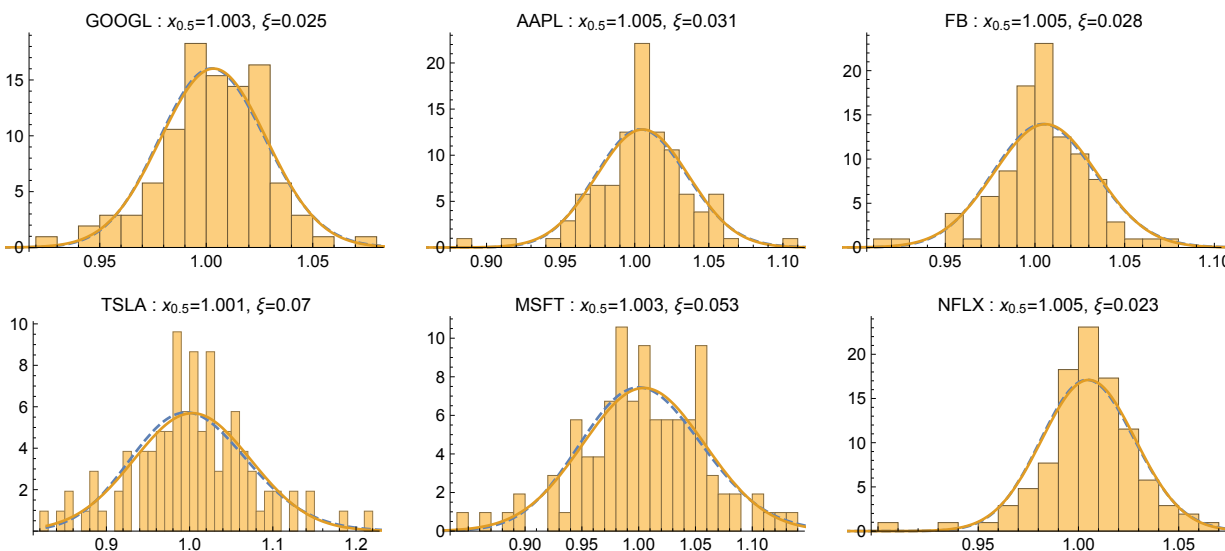
Typically, $t = 0$ is the present and we assume that future returns S_1, S_2, \dots will have a common distribution (estimated with historical data). This implies that $E[\log S_j]$ and $V[\log S_j]$ are constants, **say λ and ξ** , independent of the time period $j = 1, 2, \dots$. Therefore,

$$\lambda_t = \log[P_0] + t\lambda, \quad \xi_t^2 = t\xi^2 \tag{4.10}$$

If the returns S_t 's have the lognormal distribution, this result is exact, otherwise it is an approximation.

The figure below shows **weekly stock prices** for 6 tech companies in 2017, and it can be seen that the lognormal distribution is a good approximation.

Returns : - - - - Log-Normal — Normal



Furthermore, one may use the results in this example to produce a price forecast; in the case of Netflix we get:



In this figure we know the price of Netflix stock up to the last week of December 2017, so $P_0 = \$190$. The forecast is shown as the decile curves from the lognormal distribution with the parameters given in equation (4.10). □

Practice questions

1. On December 31st 2017, what is the probability that Netflix stock price will exceed \$350 on February 15 2018?
2. On July 1st, what is the probability that Netflix stock price will exceed \$150 on October 1st 2017?
3. what is the probability that Google stock price will exceed Apple's in two more weeks?

4.4 Bernoulli Family of Random Variables

- Bernoulli
- Binomial and multinomial
- Geometric and negative binomial

Bernoulli trial: An experiment with only two outcomes: the value 1 (success) with probability p and 0 (failure) with probability $1 - p$. For example,

- Toss a coin. Outcomes: heads or tails.
- Roll a die. Outcomes: even or odd.
- Draw a card. Outcomes: ace or not ace.

Bernoulli random variable A random variable X is said to be a Bernoulli random variable with parameter p :

$$X \sim \text{Ber}(p)$$

if its probability mass function is given by

$$p_X(x) = \begin{cases} p, & x = 1 \\ q = 1 - p, & x = 0 \end{cases}$$

and

$$\begin{aligned} E(X) &= p \\ V(X) &= pq \end{aligned}$$

Example 146. — Indicator Random Variables Let X be the random variable such that

$$X = \begin{cases} 1 & \text{if event } A \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Find the mean and variance of X in terms of $P(A)$.

Solution: Sense the probability of success here is $p = P(A)$, we have:

$$\begin{aligned} E(X) &= P(A) \\ V(X) &= P(A)(1 - P(A)) \end{aligned}$$

□

4.4.1 Binomial random variable

Consider n independent Bernoulli trials with probability of success p , and probability of failure $q = 1 - p$. If X represents **the number of successes that occur in the n Bernoulli trials**, then X is said to be a binomial random variable with parameters (n, p) .

Examples of binomial random variables are:

- Toss a coin 10 times. Let X be the number of heads.
- Roll a die 6 times. Let X be the number of even rolls.
- Draw 4 cards. Let X be the number of aces. (Is this binomial?)

X is **the number of successes in n Bernoulli trials**.

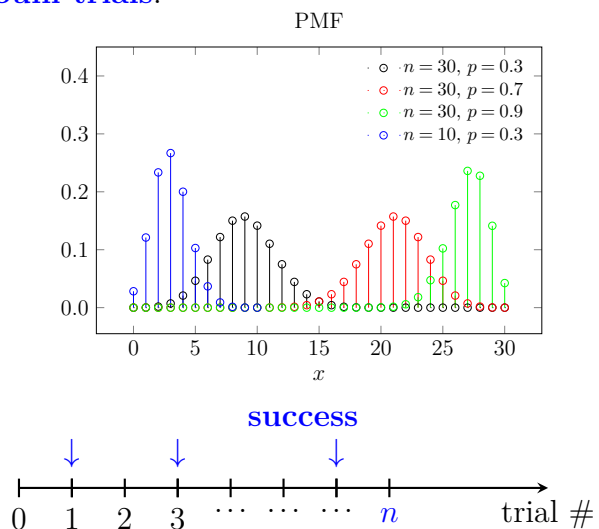
$$X \sim \text{Bin}(n, p)$$

The PMF $p_X(x) = P(X = x)$ is

$$p_X(x) = \binom{n}{x} p^x (1 - p)^{n-x} \quad (4.11)$$

and

$$\begin{aligned} E(X) &= np \\ V(X) &= npq \end{aligned}$$



Notice that the CDF $F_X(x) = P(X \leq x)$ does not simplify:

$$F_X(x) = \sum_{i=1}^x \binom{n}{i} p^i (1-p)^{n-i} \quad (4.12)$$

The logic for the PMF is as follows:

- x successes and $n - x$ failures can be arranged in exactly $\binom{n}{x}$ distinct sequences of length n trails.
- Each sequence has the same probability of occurring, $p^x(1-p)^{n-x}$.
- Therefore, the probability of one of the patterns occurring is

$$\binom{n}{x} p^x (1-p)^{n-x}.$$

The name for this random variable comes from the binomial theorem:

Fact 4.7 — The Binomial Theorem. Let n be a nonnegative integer and let a and b be any real numbers. Then

$$\begin{aligned} (a+b)^n &= a^n + \binom{n}{1} a^{n-1} b + \binom{n}{2} a^{n-2} b^2 + \cdots + \binom{n}{n-1} a b^{n-1} + b^n \\ &= \sum_{i=0}^n \binom{n}{i} a^i b^{n-i}. \end{aligned}$$

Example 147. Consider the experiment of tossing **4 fair coins**. Let X be the random variable that denotes the number of heads that result. The sample space for this experiment is illustrated in the table below, which also shows the number of heads in each possible case.

coin 1	H	H	H	H	H	H	H	H	T	T	T	T	T	T	T
coin 2	H	H	H	H	T	T	T	T	H	H	H	H	T	T	T
coin 3	H	H	T	T	H	H	T	T	H	H	T	T	H	H	T
coin 4	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H
ΣH	4	3	3	2	3	2	2	1	3	2	2	1	2	1	0

- Find the CDF and PMF of X and draw sketches of each one.
- Determine the median, upper quartile and lower quartile and show them graphically in one of the sketches of part a)
- Determine $P(0 < X \leq 3 \mid X \leq 2)$
- BONUS: suppose two players play this game, and the one with the largest number of Hs wins. Let X_1 and X_2 denote their corresponding random variables, the distribution of each one corresponding to the one you calculate it in part a). Find the joint PMF and the probability that player one wins by more than one point. Hint: X_1 and X_2 are independent.

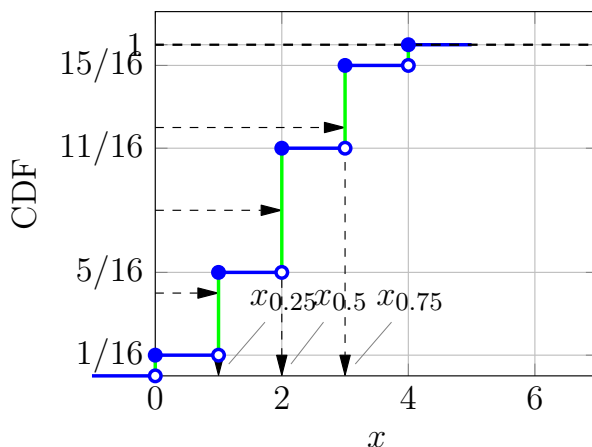
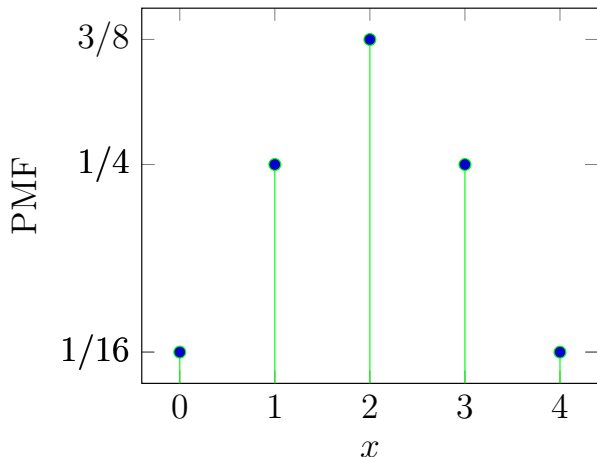
Solution:

- Find the CDF and PMF of X and draw sketches of each one.

$$X \sim \text{Bin}(n = 4, p = 1/2)$$

$$\rightarrow p_X(x) = \binom{4}{x} 0.5^x (1 - 0.5)^{4-x} = \binom{4}{x} 0.5^4 = \binom{4}{x} / 16, \text{ or:}$$

$$p_X(x) = \begin{cases} 1/16 & \text{if } x = 0 \text{ or } x = 4 \\ 4/16 & \text{if } x = 1 \text{ or } x = 3 \\ 6/16 & \text{if } x = 2 \end{cases}$$



- b) Determine the median, upper quartile and lower quartile and show them graphically in one of the sketches of part a): {2, 3, 1}
- c) $P(0 < X \leq 3 \mid X \leq 2) = 0.91$
- d) BONUS: suppose two players play this game, and the one with the largest number of Hs wins. Let X_1 and X_2 denote their corresponding random variables, the distribution of each one corresponding to the one you calculate it in part a). Find the joint PMF and the probability that player one wins by more than one point. Hint: X_1 and X_2 are independent.

□

Fact 4.8 — The sum of n Bernoulli trials has a $\text{Bin}(n, p)$ distribution. If Y_1, Y_2, \dots, Y_n are independent Bernoulli random variables,

$$Y_i \sim \text{Ber}(p)$$

for all i , and we define

$$X = \sum_{i=1}^n Y_i$$

then, by definition, $X \sim \text{Bin}(n, p)$. This fact makes it easier to compute the mean and variance of X by using the results we know for linear combinations.

Recall the experiment of tossing **4 fair coins**, if we let H=1 and T=0 we can clearly see the connection between Binomial and Bernoulli random variables:

coin 1, Y_1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0
coin 2, Y_2	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0
coin 3, Y_3	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0
coin 4, Y_4	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1
$X = \sum_{i=1}^n Y_i$	4	3	3	2	3	2	2	1	3	2	2	1	2	1	1

Fact 4.9 — Two important corollaries. :

1. The **normal approximation**

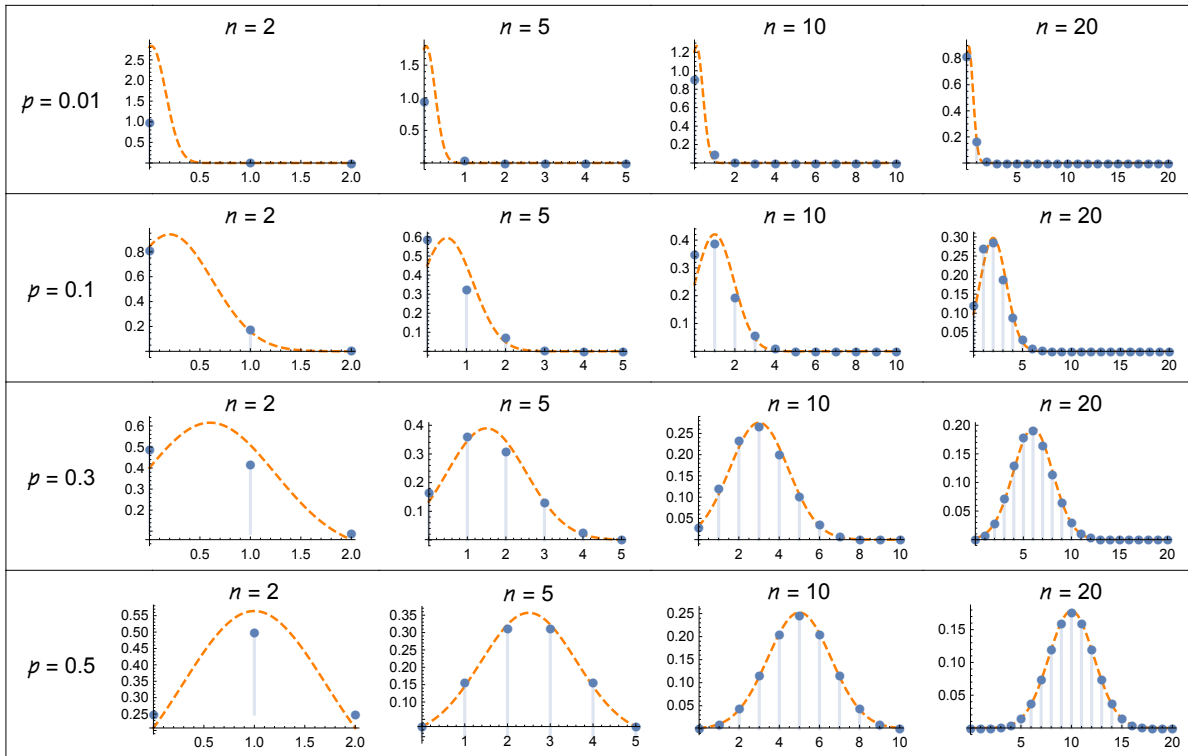
$$X \sim N(\mu = np, \sigma^2 = npq)$$

is accurate when n it is large enough (by the CLT).

2. The **sum of two binomial random variables** with the same parameter p is also binomial:

$$\text{if } X_1 \sim \text{Bin}(n_1, p) \text{ and } X_2 \sim \text{Bin}(n_2, p) \rightarrow X_1 + X_2 \sim \text{Bin}(n_1 + n_2, p)$$

The figure below shows the agreement of the normal approximation with the $\text{Bin}(n, p)$ rv for different values of parameters (n, p) .

**Continuity Correction**

We saw that the **normal approximation to a binomial random variable** $X \sim \text{Bin}(n, p)$ is:

$$X \sim N(\mu = np, \sigma^2 = npq)$$

is accurate when n it is large enough. Then, one can use

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \quad (4.13)$$

However, the error can be substantial if n is not very large. One way to improve the approximation is to use the *continuity correction*:

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b+0.5-\mu}{\sigma}\right) - \Phi\left(\frac{a-0.5-\mu}{\sigma}\right). \quad (4.14)$$

Analogous continuity corrections apply to the Poisson distribution, in which case $\mu = \theta, \sigma^2 = \theta$.

Example 148. A die is rolled 5 times. What is the probability that the result is 6, 3 times?

Solution: Let X be the number of times 6 appears.
Therefore,

$$X \sim \text{Bin}\left(5, \frac{1}{6}\right)$$

Therefore,

$$\begin{aligned} P(X = i) &= \binom{n}{i} p^i (1-p)^{n-i} \\ \therefore P(X = 3) &= \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^3 \end{aligned}$$

□

Example 149. A player bets on a number from 1 to 6, both including. Three dice are then rolled. If the number bet on by the player appears i times where $i = 1, 2, 3$, he wins i units. If the number bet on by the player does not appear on any of the dice, he loses 1 unit. A game is considered to be fair if the expected value for the player is at least 0. Is this game fair towards the player?

Solution: Let X be the player's winnings.
Let Y be the number of times the number the player bet on appeared. Therefore,

$$Y \sim \text{Bin}\left(3, \frac{1}{6}\right)$$

Therefore,

$$\begin{aligned} P(X = -1) &= P(Y = 0) \\ &= \binom{3}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^3 \\ &= \frac{125}{216} \end{aligned}$$

$$\begin{aligned} P(X = 1) &= P(Y = 1) \\ &= \binom{3}{1} \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^2 \\ &= \frac{75}{216} \end{aligned}$$

$$\begin{aligned} P(X = 2) &= P(Y = 2) \\ &= \binom{3}{2} \left(\frac{1}{6}\right)^2 \left(\frac{5}{6}\right)^1 \\ &= \frac{15}{216} \end{aligned}$$

$$\begin{aligned} P(X = 3) &= P(Y = 3) \\ &= \binom{3}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^0 \\ &= \frac{1}{216} \end{aligned}$$

Therefore,

$$\begin{aligned} E[X] &= (-1) \left(\frac{125}{216}\right) + (1) \left(\frac{75}{216}\right) + (2) \left(\frac{15}{216}\right) + (3) \left(\frac{1}{216}\right) \\ &= -\frac{17}{216} \end{aligned}$$

Therefore, as the expected value of the winnings is less than 0, the game is not fair towards the player. \square

Example 150. Tests show that about 20% of all private wells in some specific region are contaminated. What are the probabilities that in a random sample of 4 wells exactly 2, fewer than 2, or at least 2 wells are contaminated?

Solution: Here $n = 4$, $p = 0.2$ (success for being contaminated). We find

$$\begin{aligned} P(X = 2) &= \binom{4}{2} 0.2^2 0.8^{4-2} = 0.1536, \\ P(X < 2) &= P(X = 0) + P(X = 1) = \binom{4}{0} 0.2^0 0.8^4 + \binom{4}{1} 0.2^1 0.8^3 = 0.8192, \\ P(X \geq 2) &= P(X = 2) + P(X = 3) + P(X = 4) \\ &= 0.1536 + \binom{4}{3} 0.2^3 0.8^1 + \binom{4}{0} 0.2^4 0.8^0 = 0.1808. \end{aligned}$$

\square

Example 151. — Tornadoes, take 3 100 structures are located in a region where tornado wind force must be considered in its design. Suppose that from the records of tornadoes for the past 200 years, it is estimated that

1. during any given **week, at most 1 tornado can occur with probability** $p = 1/30$,
2. the number of tornadoes in different weeks are independent, and
3. if a tornado occurs, a structure will be damaged if the wind speed exceeds the structure design wind speed of 130 mph,
4. wind speeds have a median of 90 mph, a coefficient of variation of 20 percent, and follow the lognormal distribution.

Determine the following:

- a) the probability that the structure will be damaged this during a tornado?
- b) what is the probability the a structure will be damaged in the next year?
- c) calculate the mean and variance of the number of structures damaged in the next five years?
- d) If you're a contractor in charge of rehabilitating the structures in the region after a tornado damage, compute the mean and variance of your yearly income, U , if you charge c dollars per rehabilitation work.
- e) calculate the coefficient of variation of your yearly income, and comment.

Solution:

- a) the probability that the structure will be damaged this during a tornado?

Let Y be the wind speeds during a tornado,

$$Y \sim \text{LogN}(\lambda = \log 90, \xi^2 = 0.2^2)$$

and the desired probability is

$$r = P(Y > 130) = 0.033$$

- b) what is the probability the a structure will be damaged in the next year?

Let X be the rv representing the number of tornadoes on a given year, then

$$X \sim \text{Bin}(n = 52, p = 1/30)$$

Let D the event that a structure will be damaged in one year.

Since we don't know the number of tornadoes that will occur, we use the total probability rule:

$$\begin{aligned} P(D^c) &= \sum_{x=0}^n P(D^c | X = x) P(X = x) \\ &= \sum_{x=0}^n (1-r)^x P(X = x) \\ &= \sum_{x=0}^n (1-r)^x \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^n \binom{n}{x} ((1-r)p)^x (1-p)^{n-x} \\ &= (1-rp)^n = 0.944 \end{aligned}$$

and the desired probability is $1 - 0.944 = 0.0556$. The last step follows from the binomial theorem (4.7) with $a = (1 - r)p$, $b = 1 - p$.



Notice that this solution method proves that a simpler way to do this type of problems is to let Z be the rv representing **the number of tornadoes that cause damage to a structure** on a given year, then

$$Z \sim \text{Bin}(n, rp)$$

and the desired probability is also $P(Z > 0) = 1 - (1 - rp)^n = 0.0556$.

- c) solved in class
- d) solved in class
- e) solved in class

□

4.4.2 The Multinomial distribution

This is a generalization of the binomial distribution:

1. k possible outcomes,
2. each occurs with probability p_i , with $\sum_{i=1}^k p_i = 1$
3. $N_i =$ number of observations yielding the i th outcome, $i = 1, 2, \dots, k$

The (joint) distribution of the random vector $\mathbf{N} = \{N_1, \dots, N_k\}$ is

$$P(n_1, \dots, n_k) = \frac{n!}{n_1! \cdots n_k!} \prod_{i=1}^k p_i^{n_i}$$

and:

$$\begin{aligned} E(N_i) &= np_i \\ V(N_i) &= np_i(1 - p_i) \\ \text{Cov}(N_i, N_j) &= -np_i p_j \end{aligned}$$

Note that, since the occurrence of one outcome means the others cannot occur, the individual outcomes must be negatively correlated. In fact, the covariance between the i th and j th ($i \neq j$) outcome is $\text{Cov } N_i, N_j = -np_i p_j$.

Fact 4.10 — Marginal distributions. $N_i \sim \text{Bin}(n, p_i)$

Example 152. Suppose that 60% of the supply of raw material kits used in a chemical reaction can be classified as recent, 30% as moderately aged, 8% as aged, and 2% unusable. 16 kits are randomly chosen to be used for 16 chemical reactions. Let N_1, N_2, N_3, N_4 denote the number of chemical reactions performed with recent, moderately aged, aged, and unusable materials.

- a) Find the probability that exactly one of the 16 planned chemical reactions will not be performed due to unusable raw materials.
- b) Find the probability that 10 chemical reactions will be performed with recent materials, 4 with moderately aged, and 2 with aged materials.

- c) Do you expect N_1 and N_2 to be positively or negatively correlated? Explain intuitively.
- d) Find $\text{Cov}(N_1, N_2)$.

Solution: (a) According to Fact 4.10, $N_4 \sim \text{Bin}(16, 0.02)$. Thus, $P(N_4 = 1) = 16(0.02)(0.98)^{15} = 0.2363$.

(b) $P(N_1 = 10, N_2 = 4, N_3 = 2, N_4 = 0) = \frac{16!}{10!4!2!} 0.6^{10} 0.3^4 0.08^2 = 0.0377$.

c) Expect them to be negatively related: The larger N_1 is, the smaller N_2 is expected to be.

d) $\text{Cov}(N_1, N_2) = -16(0.6)(0.3) = -2.88$.

□

4.4.3 Geometric Random Variables

Let X be the number of Bernoulli trials required until the first success occurs, then

$$X \sim \text{Geo}(p)$$

The PMF $p_X(x) = P(X = x)$ is

$$p_X(x) = q^{x-1} p$$

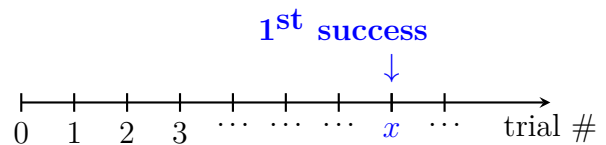
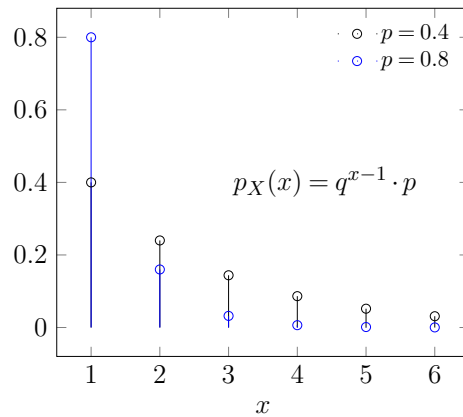
with $q = 1 - p$, and

$$E(X) = 1/p$$

$$V(X) = q/p^2$$

The CDF is given by:

$$P(X \leq x) = 1 - q^x$$



STOP! The return period T . In the case of geometric random variables where the

underlying Bernoulli trial is repeated in regular time intervals (e.g. every day, weekly, once a year...), the mean value $E(X)$ is also called the return period T , and therefore

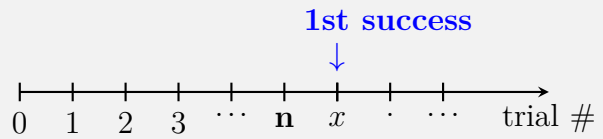
$$p = \frac{1}{T}$$

For example, if a system is designed to withstand the 100-year earthquake, the implicit assumption is that (i) the Bernoulli trial is repeated once a year, and (ii) the time between earthquakes has the geometric distribution with parameter $p = 1/100$.

Fact 4.11 — Connection between the $\text{Geo}(p)$ and $\text{Bin}(n, p)$ distributions. If

$Y \sim \text{Bin}(n, p)$, # of successes in n trials
 $X \sim \text{Geo}(p)$, # of trials until first success

then, from the picture:



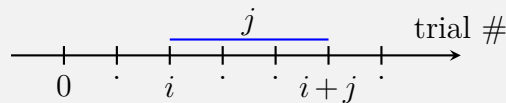
$$\begin{aligned} P(X > n) &= P(Y = 0) \\ &= \binom{n}{0} p^0 q^{n-0} \\ &= q^n \end{aligned}$$

Therefore, the CDF of the geometric distribution $F_X(x)$ is given by:

$$P(X \leq x) = 1 - q^x \quad (4.15)$$

Fact 4.12 — Memoryless Property of the Geometric Distribution. Let i, j be positive integers. $X \sim \text{Geo}(p)$. Then

$$P(X > i + j | X > i) = P(X > j)$$



This means that, if i represents the present trial number, all that matters for a geometric rv is **the number of additional trials j until the first success**, which also has the $\text{Geo}(p)$ distribution.

Proof.

$$\begin{aligned} P(X > i + j | X > i) &= \frac{P(\{X > i + j\} \cap \{X > i\})}{P(X > i)} \\ &= \frac{P(X > i + j)}{P(X > i)} \\ &= \frac{q^{(i+j)}}{q^i} = q^j \\ &= P(X > j) \end{aligned}$$

■

This result is intuitive because by definition the Bernoulli trials are independent of one another, and therefore past outcomes do not influence future outcomes.

Example 153. — **The height above sea level of a fixed offshore platform** is designed to withstand the 20-year wave height. Determine

- a) in one year, what is the probability that the platform will be flooded?

- b) the probability that the platform will be subjected to the design wave height within the return period ?
- c) the probability that the first exceedance of the design wave height will occur after the third year?
- d) If the first exceedance of the design wave height should occur after the third year, what is the probability that such a first exceedance will occur in the fifth year?

Solution:

- a) in one year, what is the probability that the platform will be flooded?

Since the return period is 20 years, the design wave height will be exceeded with $p = 1/20 = 5\%$ probability each year.

- b) the probability that the platform will be subjected to the design wave height within the return period ?

Let X be the number of years until the next flooding. Therefore,

$$X \sim \text{Geo}(1/20) \rightarrow P(X \leq 20) = 1 - (1 - 1/20)^{20} = 0.6425$$

- c) the probability that the first exceedance of the design wave height will occur after the third year?

$$P(X > 3) = 1 - P(X \leq 3) = 0.95^3$$

- d) If the first exceedance of the design wave height should occur after the third year, what is the probability that such a first exceedance will occur in the fifth year?

$$P(X = 5 | X > 3) = 0.048$$

□



STOP! Approximation for rare events. Solution b) above for the probability that the first event (flooding) happens within one return period, can be simplified to

$$P(X \leq T) \approx 1 - e^{-1} = 0.6321$$

in the case of events with long return period T , by virtue of the identity

$$\lim_{T \rightarrow \infty} \left(1 - \frac{1}{T}\right)^T = e^{-1}$$

Example 154. — * **8.5-magnitude earthquakes** in the city of San Diego, CA, have a return period of 30 years. Houses and tall buildings can suffer structural damage during such an earthquake with probabilities 50 and 20 percent, respectively.

- (a) Find the probability of damage in 100 years using the Bernoulli model where one trial = one year.
- (b) Find the probability that in 100 years there will be more than 2 damages to any particular structure.
- (c) If there are 1000 houses and 950 buildings in the region, find the probability that within 100 years there will be more structural damage to buildings than houses.

(d) BONUS: If you're a contractor in charge of rehabilitating the structures in the region, compute the probability that your yearly income, U , will exceed 1,000,000 dollars if you charge 10,000 and 200,000 thousand dollars per rehabilitation of houses and tall buildings, respectively.

Solution: a) Assume that if the building was not damaged after an earthquake it remains in its original condition.

Probability of occurrence of earthquake in a certain year:

$$p = \frac{1}{T} = 1/30$$

Probability of damage if an earthquake happens:

$$\begin{aligned} r_1 &= 0.5 && \text{(houses)} \\ r_2 &= 0.2 && \text{(buildings)} \end{aligned}$$

Probability of damage in one year:

$$\begin{aligned} p_1 &= r_1 p = 0.0166667 && \text{(houses)} \\ p_2 &= r_2 p = 0.0066667 && \text{(buildings)} \end{aligned}$$

Let X_1 and X_2 be the number of years between earthquakes that produced damage to a given house and building, respectively. Then,

$$X_i \sim Geo(p_i) \rightarrow P(X_i \leq 100) = 1 - (1 - p_i)^{100}$$

which gives that the probability of damage during the next 100 years are 0.186241 and 0.512272 for a house and building, respectively.

(b) Find the probability that in 100 years there will be more than 2 damages to any particular structure.

Let random variable Y_i be the number of damages in 100 years of a structure of type i . Then,

$$Y_i \sim Bin(100, p_i) \rightarrow P(Y_i \geq 2) = 1 - P(Y_i < 2)$$

which gives 0.233259 and 0.0297029 for a house and building, respectively.

(c) if there are $n_1 = 1,000$ houses and $n_2 = 950$ buildings, find the probability that within 100 years there will be more structural damage to buildings than houses.

Let W_1 and W_2 be the number of damages in 100 years of ALL structures of type i . Then,

$$\begin{aligned} W_i &= \sum_{j=1}^{n_i} Y_{i,j} \\ &\sim N(n_i E(Y_i), n_i V(Y_i)) \quad \text{by the CLT} \end{aligned}$$

where

$$\begin{aligned} E(Y_i) &= 100p_i \\ V(Y_i) &= 100p_i q_i \end{aligned}$$

Finally, $P(W_2 > W_1) = P(W_2 - W_1 > 0)$, which gives ≈ 0 in this case. We used

$$W_2 - W_1 \sim N \left(\underbrace{E(W_2) - E(W_1)}_{=-33.3}, \underbrace{V(W_2) + V(W_1)}_{=1284.7} \right)$$

(d) BONUS: If you're a contractor in charge of rehabilitating the structures in the region, compute the probability that your yearly income, U , will exceed 1,000,000 dollars if you charge 10,000 and 200,000 thousand dollars per rehabilitation of houses and tall buildings, respectively.

Here, we are interested in

$$U = \sum_{j=1}^{1000} \times 10,000 Y_{1,j} + \sum_{j=1}^{950} 200,000 \times Y_{2,j}$$

By CLT and linearity of expectation we get

$$U \sim N \left(\underbrace{E(U)}_{1,433}, \underbrace{V(U)}_{253,283} \right)$$

in thousands of dollars. The final answer is 0.80539. □

Example 155. Alice eats cookies one after another until she finds a chocolate cookie. For each cookie, the probability of the cookie being a chocolate cookie is $\frac{1}{10}$.

1. What is the probability that Alice eats more than 3 cookies?
2. Given that Alice has already eaten 5 cookies, and has not found a chocolate cookie, what is the probability that she will eat at least 8 more cookies?

Solution:

1.

$$\begin{aligned} P(X > 3) &= \sum_{k=4}^{\infty} P(X = k) \\ &= \sum_{k=4}^{\infty} \left(1 - \frac{1}{10}\right)^{k-1} \left(\frac{1}{10}\right) \\ &= \left(1 - \frac{1}{10}\right)^3 \sum_{j=1}^{\infty} \left(1 - \frac{1}{10}\right)^{j-1} \left(\frac{1}{10}\right) \\ &= \left(\frac{9}{10}\right)^3 \left(\frac{1}{10}\right) \left(\frac{1}{1 - \frac{9}{10}}\right) \\ &= \left(\frac{9}{10}\right)^3 \end{aligned}$$

2.

$$\begin{aligned}
 P(X \geq 13|X > 5) &= P(X > 12|X > 5) \\
 &= \frac{P(X > 12 \cap X > 5)}{P(X > 5)} \\
 &= \frac{P(X > 12)}{P(X > 5)} \\
 &= \frac{\left(\frac{9}{10}\right)^{12}}{\left(\frac{9}{10}\right)^5} \\
 &= \left(\frac{9}{10}\right)^7 \\
 &= P(X > 7)
 \end{aligned}$$

Therefore, the fact that Alice has already eaten 5 cookies does not affect the probability of her eating at least 8 more cookies. □

Example 156. A test of weld strength involves loading welded joints until a fracture occurs. For a certain type of weld, 80% of the fractures occur in the weld itself, while the other 20% occur in the beam. A number of welds are tested. Let X be the number of tests up to and including the first test that results in a beam fracture.

- (a) Find $P(X=3)$
 (b) Find the **mean** and **variance** of X

Solution: Answer: (a) 0.128 (b) 5, 20

(a)

$$\begin{aligned}
 P(X = 3) &= (0.8)^2 0.2 \\
 &= 0.128
 \end{aligned}$$

(b) X follows the geometric distribution: $X \sim Geo(0.2)$

$$E(X) = \frac{1}{0.2} = 5 \quad V(X) = \frac{1-0.2}{0.2^2} = 20$$

□

4.4.4 Negative Binomial Random Variable

Here X is **the number of Bernoulli trials required until the r -th success occurs.**

$$X \sim NB(r, p)$$

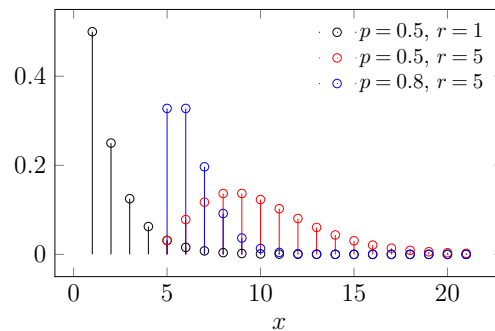
The PMF of X is

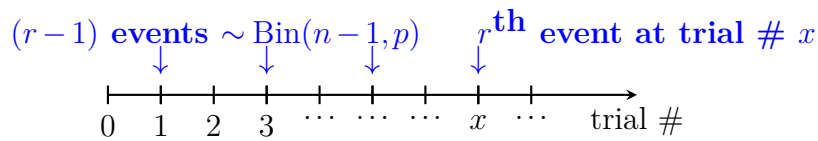
$$P(X = n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

$$E(X) = r/p$$

$$V(X) = rq/p^2$$

PMF





The logic for the PMF is as follows. The last trial must necessarily result in a success (which happens with probability p), and there must be $r-1$ success in the first $n-1$ trials (which happens according to a $\text{Bin}(n-1, p)$ random variable). Therefore, the probability distribution of X is

$$\begin{aligned}
 P(X = n) &= \binom{n-1}{r-1} p^{r-1} (1-p)^{n-r} \cdot p \\
 &= \binom{n-1}{r-1} p^r (1-p)^{n-r}
 \end{aligned}$$

Example 157. Find the expected value and variance of the number of times one must throw a die until the outcome 1 has occurred four times.

Solution: Let X be the number of times the die must be thrown for 1 to occur four times. Therefore,

$$X \sim \text{NB}\left(4, \frac{1}{6}\right)$$

Therefore,

$$\begin{aligned}
 E[X] &= \frac{r}{p} \\
 &= \frac{4}{\frac{1}{6}} \\
 &= 24 \\
 V(X) &= \frac{r(1-p)}{p^2} \\
 &= \frac{4\left(1 - \frac{1}{6}\right)}{\left(\frac{1}{6}\right)^2} \\
 &= 120
 \end{aligned}$$

□

Fact 4.13 Let

$$X_i \sim \text{Geo}(p)$$

be independent random variables, for $i \in \mathbb{N}$. Then,

$$\sum_{i=1}^n X_i \sim \text{NB}(n, p)$$

4.4.5 Hypergeometric Random Variable

A hypergeometric experiment:

1. A sample of size n is randomly selected without replacement from a population of N items.
2. In the population, k items can be classified as successes, and $N - k$ items can be classified as failures.

Let X be the number of successes in the n trials:

$$X \sim \text{HG}(n, N, k)$$

The probability distribution of X is

$$\begin{aligned} P(X = i) &= \frac{\binom{k}{i} \binom{N-k}{n-i}}{\binom{N}{n}} \\ E[X] &= \frac{nk}{N} \\ V(X) &= n \frac{k}{N} \left(1 - \frac{k}{N}\right) \left(\frac{N-n}{N-1}\right) \end{aligned}$$

Example 158. Suppose we randomly select 5 cards without replacement from an ordinary deck of playing cards. What is the probability of getting exactly 2 red cards (i.e., hearts or diamonds)?

Solution: We know the following:

$N = 52$; since there are 52 cards in a deck.

$k = 26$; since there are 26 red cards in a deck.

$n = 5$; since we randomly select 5 cards from the deck.

$i = 2$; since 2 of the cards we select are red.

$$P(X = i) = \frac{\binom{k}{i} \binom{N-k}{n-i}}{\binom{N}{n}} = 0.32513$$

□

Example 159. An extensive study undertaken by the National Highway Traffic Safety Administration reported that 17% of children under 5 use no seat belt, 29% use adult seat belt, and 54% use child seat. Set N_1, N_2, N_3 for the number of children using no seat belt, adult seat belt, and child seat, respectively. In a sample of 15 children under five. Find:

- a) the probability that exactly 10 children use child seat?
- b) the probability that exactly 10 children use child seat and 5 use adult seat?
- c) the probability that exactly 8 children use child seat, 5 use adult seat and 2 use not seat belt?
- d) $\text{Cov}(N_1, N_2)$.
- e) $\text{Cov}(N_1, N_2 + N_3)$.

Solution: office hours :)

□

4.5 Poisson Random Variables

A discrete random variable X , taking one of the values $0, 1, 2, \dots$, is said to be a **Poisson random variable** with parameter $\theta > 0$ if:

$$X \sim \text{Poi}(\theta)$$

The PMF of X is

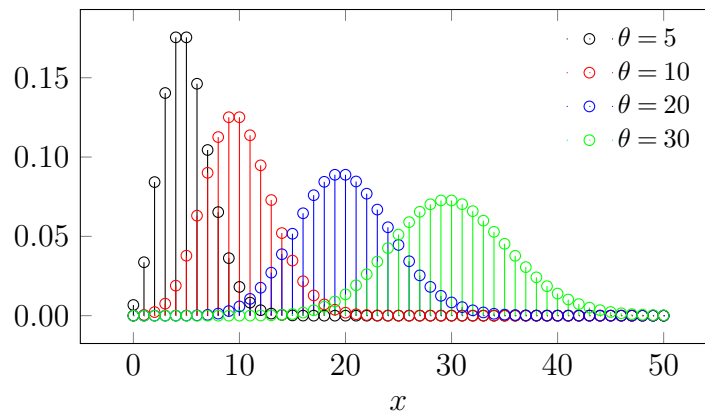
$$p_X(x) = \frac{e^{-\theta} \theta^x}{x!}$$

and

$$E(X) = \theta$$

$$V(X) = \theta$$

PMF



Practical applications

The Poisson random variable has a very wide range of applications in a very diverse number of areas such as physics, finance, biology, physics, and telecommunications. Extremely useful in modeling situations consisting of a large number of independent trials with a consistent but very small probability of occurrence.

The Poisson distribution has also been used to model a number of modern phenomena including:

1. internet traffic,
2. phone call arrivals, the number of telephone calls originating in a given locality during a certain period;
3. scoring in sporting events
4. the number of particles emitted by a lump of radioactive material undergoing radioactive decay during a certain period;
5. the occurrence of accidents at a given intersection over a certain period;
6. the breakdowns of a machine over a certain period of time;
7. the arrival of customers in a queue for service during a certain period.

In a Poisson Process events occur at a given rate

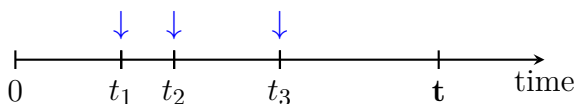
$$\lambda = \text{number of events per unit of time}$$

(or distance t , or area, or volume, or population, etc.). Then,

$$X = \# \text{ of events in } (0, t) \\ \sim \text{Poi}(\lambda t),$$

Note that the Poisson parameter $\theta = \lambda t$ has no units.

1st event



Fact 4.14 — **The Poi(θ) rv as a limit of a Bin(n, p).** Let $\theta = np$ be fixed. Then the binomial PMF tends to the Poisson PMF,

$$\lim_{n \rightarrow \infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{e^{-\theta} \theta^x}{x!}$$

This means that the Poisson distribution will be appropriate whenever the rv can be thought of as a Bin(n, p) rv **with large n and small p** .

Therefore the Poisson distribution *inherits* the 2 important properties of the binomial distribution:

1. The **normal approximation**

$$X \sim N(\theta, \theta)$$

is accurate when θ is large enough (by the CLT).

2. The **sum of two Poisson random variables** with parameters θ_1 and θ_2 is also Poisson:

$$\text{if } X_1 \sim \text{Poi}(\theta_1) \text{ and } X_2 \sim \text{Poi}(\theta_2) \rightarrow X_1 + X_2 \sim \text{Poi}(\theta_1 + \theta_2)$$

Proof. Express the binomial probability in terms of the parameter θ :

$$\begin{aligned} \lim_{n \rightarrow \infty} \binom{n}{x} p^x (1-p)^{n-x} &= \lim_{n \rightarrow \infty} \binom{n}{x} \left(\frac{\theta}{n}\right)^x \left(1 - \frac{\theta}{n}\right)^{n-x} \\ &= \lim_{n \rightarrow \infty} \frac{n!}{x!(n-x)!} \theta^x \left(\frac{1}{n}\right)^x \left(1 - \frac{\theta}{n}\right)^{-x} \left(1 - \frac{\theta}{n}\right)^n \\ &= \frac{\theta^x}{x!} \lim_{n \rightarrow \infty} \frac{n!}{(n-x)!} \frac{1}{(n-\theta)^x} \left(1 - \frac{\theta}{n}\right)^n \\ &= \frac{\theta^x}{x!} \lim_{n \rightarrow \infty} \frac{n!}{(n-x)!} \frac{1}{(n-\theta)^x} \left(1 - \frac{\theta}{n}\right)^n \end{aligned}$$

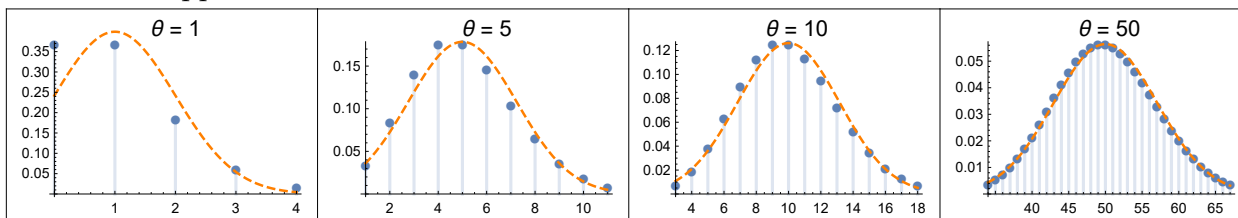
From calculus, we know that

$$\lim_{n \rightarrow \infty} \left(1 - \frac{\theta}{n}\right)^n = e^{-\theta}$$

and

$$\lim_{n \rightarrow \infty} \frac{n!}{(n-x)!} \frac{1}{(n-\theta)^x} = \lim_{n \rightarrow \infty} \frac{n(n-1)\dots(n-x+1)}{(n-\theta)(n-\theta)\dots(n-\theta)} = 1$$

The normal approximation as a function of θ :



Example 160. Consider an experiment that consists of counting the number of α -particles given off by a gram of radioactive material. If it is known that on average, 32 such α -particles are emitted **in 20 seconds**, what is the probability that no more than 2 α -particles will be emitted **in two second**?

Solution: Let X be the number of α particles emitted in **two second**. Here $\lambda = 32/20$ and $t = 2\text{s}$, so the Poisson parameter

$$\theta = \lambda t = 3.2$$

Therefore,

$$X \sim \text{Poi}(3.2)$$

and,

$$\begin{aligned} P(X \leq 2) &= P(X = 0) + P(X = 1) + P(X = 2) \\ &= \frac{e^{-3.2}\theta^0}{0!} + \frac{e^{-3.2}\theta^1}{1!} + \frac{e^{-3.2}\theta^2}{2!} \\ &\approx 0.3799 \end{aligned}$$

□

Example 161. An LCD display has 1920×1080 pixels. A display is accepted if it has 15 or fewer faulty pixels. The probability that a pixel is faulty from production is 5×10^{-5} .

(a) Find the proportion of displays that are accepted.

(b) Find the pixel failure rate required to produce 4000×2000 pixel displays and still have an acceptance rate of at least 90%.

Solution: Since there is a large number n of Bernoulli trials where the probability p of success is small, we can use the Poisson random variable with parameter $\theta = np = 1920 \times 1080 \times 5 \times 10^{-5} = 103.68$.

X : number of pixels that are faulty

(a)

$$\begin{aligned} P(\text{Accepted}) &= P(X \leq 15) \\ &= \sum_{x=0}^{15} \frac{\theta^x e^{-\theta}}{x!} \\ &= \sum_{x=0}^{15} \frac{(103.68)^x e^{-103.68}}{x!} \\ &= 1.44 \times 10^{-27} \end{aligned}$$

(b) Assume λ is the pixel failure rate required:

$$\begin{aligned} P(\text{Accepted}) &= P(X \leq 15) \\ &= \sum_{x=0}^{15} \frac{(\lambda t)^x e^{-\lambda t}}{x!} \\ &= \sum_{x=0}^{15} \frac{(\lambda \times 4000 \times 2000)^x e^{-\lambda \times 4000 \times 2000}}{x!} \\ &= 0.9 \\ \lambda &= 1.39 \times 10^{-6} \end{aligned}$$

□

Example 162. Consider that earthquakes occur with the assumptions of Poisson distributions, with $\lambda = 2$ earthquakes per week.

- a) Find the probability that at least three earthquakes occur during the next two weeks.

Solution:

- a) Let X be the number of earthquakes occurring in two weeks. Therefore,

$$X \sim \text{Poi}(4)$$

Therefore,

$$\begin{aligned} P(X \geq 3) &= 1 - (P(X = 0) + P(X = 1) + P(X = 2)) \\ &= 1 - \left(\frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!} + \frac{e^{-4}4^2}{2!} \right) \\ &= 1 - 13e^{-4} \end{aligned}$$

□

Example 163. Since 1851, exactly 116 hurricanes have hit Florida. In 2005, Florida was hit by four hurricanes: Cindy, Dennis, Katrina, and Wilma. If the probability of hurricane strikes has remained the same since 1851, what is the probability of Florida being struck by four or more hurricanes in the same year?

Solution:

This is a classic Poisson distribution. We've assumed that the probability of hurricane strikes has remained the same (In reality, a bad assumption)., hence our rate is 116 hurricanes per $2016 - 1851 + 1 = 166$ years. As the question asks about a year time frame, we have to adjust the rate:

$$\theta = \frac{116}{166} \times 1.$$

Now, the probability of four or more hurricanes in the same year is

$$1 - \sum_{k=0}^3 P(X = k) = 1 - \sum_{k=0}^3 \frac{\theta^k e^{-\theta}}{k!} \approx 0.005719 = 0.57\%.$$

Notice that the return period of this event is $1/0.005719 \approx 175$: this is a “1 in 200 years” type of event. □

Example 164. In the solutions manual to a Calculus textbook, there is about one faulty solution per fifty questions. In a book with ten chapters, each with one hundred questions, what is the probability that there are at least 15 faulty solutions in the whole book? Give your answer two ways: first with a binomial distribution, then with a Poisson approximation. Use Wolfram Alpha or some other tools to find both answers numerically, and compare them.

Solution: This is exactly a binomial distribution, and approximately a Poisson distribution. A “success” is a faulty solution, hence $p = 1/50 = 0.02$. There are 1000 total problems, and so the probability that at least 15 are faulty is

$$P(X \geq 15) = \sum_{k=15}^{1000} \binom{1000}{k} \left(\frac{1}{50}\right)^k \left(\frac{49}{50}\right)^{1000-k} \approx 0.89747 = 89.7\%.$$

Using a Poisson approximation, the rate (which needs to have “per book” as its unit measurement) is

$$\theta = np = 20 \quad (\text{avg faulty solutions in 1 book}).$$

Hence,

$$P(X \geq 15) = \sum_{k=15}^{1000} \frac{\theta^k e^{-\theta}}{k!} = 0.89513 = 89.5\%.$$

□

Example 165. — **** A structure is located in a region where tornado wind force** must be considered in its design. Suppose that from the records of tornadoes for the past 20 years, the mean occurrence rate of tornadoes in the region is once every 10 years. Assume that the occurrence of tornadoes can be modeled as a Poisson process. The structure is designed to withstand a tornado force with an allowable probability of damage of 5%. (10 points)

(a) What is the distribution (and parameter(s)) of Y = the number of times the structure is damaged due to tornadoes in the next 50 years? Assume that if the structure is damaged it is immediately retrofitted to its original condition.

(b) What is the probability that the structure will be damaged in the next 50 years? (10 points)

(c) Suppose that there are 100,000 similar structures in a country. Assuming statistical independence among these structures, what is the distribution (and parameter(s)) of Z = the number of structures in the country that suffer damage due to tornadoes in the next 50 years? What is $P(Z < 22,000)$? (20 points)

Solution: Answer: (a) $Pois(0.25)$ (b) 0.2212 (c) $Bin(100000, 0.2212)$, 0.1814

(a) The mean occurrence rate of tornado is $\frac{1}{10}$, and the structure is designed to withstand a tornado force with an allowable probability of damage of 5%.

So we have:

$$\begin{aligned} \lambda &= \frac{1}{10} \times 5\% \\ &= 0.005 \end{aligned}$$

$$t = 50$$

$$\begin{aligned} \theta &= \lambda t \\ &= 0.005 \times 50 \\ &= 0.25 \end{aligned}$$

$$Y \sim Pois(0.25)$$

(b)

$$\begin{aligned} P(Y \geq 1) &= 1 - P(Y = 0) \\ &= 1 - e^{-\theta} \\ &= 1 - e^{-0.25} \\ &= 0.2212 \end{aligned}$$

(c) $Z \sim \text{Bin}(100000, 0.2212)$

Use normal approximation:

$$\begin{aligned} E(Z) &= np = 100000 \times 0.2212 = 22120 \\ V(Z) &= npq = 100000 \times 0.2212 \times (1 - 0.2212) \\ &= 17727.056 \\ \sigma_Z &= \sqrt{V(Z)} = 131.252 \\ P(Z < 22000) &= \Phi\left(\frac{22000 - E(Z)}{\sigma_Z}\right) \\ &= \Phi(-0.91) \\ &= 0.1814 \end{aligned}$$

□

4.6 Exponential Random Variable

A random variable X is said to be an exponential random variable over the interval $(0, \infty)$,

$$X \sim \text{Expo}(\lambda)$$

The PDF of X is

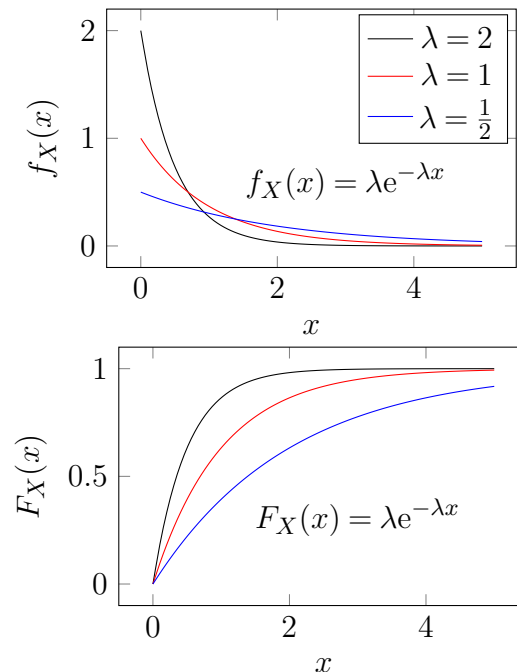
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & ; x \geq 0 \\ 0 & ; x < 0 \end{cases}$$

The CDF:

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & ; x \geq 0 \\ 0 & ; x < 0 \end{cases}$$

and

$$\begin{aligned} E(X) &= 1/\lambda \\ V(X) &= 1/\lambda^2 \end{aligned}$$



Example 166. The number of major faults on a randomly chosen 1 km stretch of highway has a Poisson distribution with mean 1.8. The random variable X is the distance (in km) between two successive major faults on the highway.

Part a) What is the probability of having at least one major fault in the next 2 km stretch on the highway? Give your answer to 3 decimal places. Part b) Which of the following describes the distribution of X , the distance between two successive major faults on the highway?

Part c) What is the mean distance (in km) and standard deviation between successive major faults?

A. mean = 3.6000; standard deviation = 3.6000 B. mean = 0.5556; standard deviation = 0.5556

C. mean = 1.8; standard deviation = 1.8 D. mean = 0.5556; standard deviation = 0.3086 E. mean = 0.2778; standard deviation = 0.2778 Part d) What is the median distance (in km) between successive major faults? Give your answer to 2 decimal places. Part e) What is the probability you must travel more than 3 km before encountering the next four major faults? Give your answer to 3 decimal places. Part f) By expressing the problem as a sum of independent Exponential random variables and applying the Central Limit Theorem, find the approximate probability that you must travel more than 25 km before encountering the next 33 major faults? Give your answer to 3 decimal places.

Solution:

□

Example 167. Assume the waiting time a customer in a restaurant is exponentially distributed with an average wait time of 5 minutes. Find the probability that the customer will have to wait no more than 10 minutes.

Solution: Let X be the the waiting time a customer spends in a restaurant, in minutes. Therefore,

$$X \sim \text{Expo}(\lambda = 1/5)$$

Therefore,

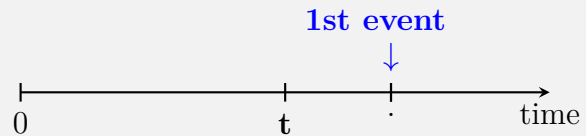
$$P(X \leq 10) = 1 - e^{-10/5} = 0.864665$$

□

Fact 4.15 — Connection between the $\text{Expo}(\lambda)$ and $\text{Poi}(\theta = \lambda t)$ distributions. If

$Y \sim \text{Poi}(\lambda t)$, # of events in $(0, t)$
 $X =$ time of first event

then, from the picture:



$$\begin{aligned} P(X > t) &= P(Y = 0) \\ &= \frac{e^{-\lambda t} (\lambda t)^0}{0!} \\ &= e^{-\lambda t} \end{aligned}$$

Therefore,

$$X \sim \text{Expo}(\lambda)$$

Fact 4.16 — The time between arrivals. of a Poisson process, $\text{Poi}(\lambda t)$ are independent, identically distributed exponential random variables having mean $1/\lambda$.

Fact 4.17 — Memoryless Property of the Exponential Distribution. Let t, s be positive real numbers and $X \sim \text{Expo}(\lambda)$. Then

$$\begin{aligned} P(X > t + s | X > t) &= P(X > s) \\ &= e^{-\lambda s} \end{aligned}$$

This means that, if t represents the present time, all that matters for an exponential rv is $Y =$ **the remaining time s until the next event**, which also has the $\text{Expo}(\lambda)$ distribution. Also,

$$E(X|X > t) = E(Y) = E(X) = 1/\lambda$$

Proof.

$$\begin{aligned} P(X > t+s|X > t) &= \frac{P(\{X > t+s\} \cap \{X > t\})}{P(X > t)} = \frac{P(X > t+s)}{P(X > t)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s) \end{aligned}$$

■



The $\text{Expo}(\lambda)$ and $\text{Geo}(p)$ are the only memoryless distributions.

For any other distribution the conditional probability depends on the present time t . **For example**, or the uniform distribution $X \sim U(0,1)$ where $P(X > x) = 1 - x$ in $0 \leq x \leq 1$ we have:

$$P(X > t+s|X > t) = \frac{P(X > t+s)}{P(X > t)} = \frac{1 - (t+s)}{1 - t} \quad (4.16)$$

which depends on the present time t . Equation (4.16) implies that $Y = \{X|X > t\}$ **the remaining time s until the next event** has the $U(0, 1-t)$ distribution (why?), so:

$$E(X|X > t) = \frac{1-t}{2} \neq E(X)$$

Example 168. A battery has a lifespan that is exponentially distributed with rate parameter $1/3000$ per hour.

- Find the probability that a random battery has a lifespan of more than 2500 hours.
- Find the probability that a random battery has a lifespan of more than 2500 hours, given that it has already worked for 2000 hours.

Solution: Let X be the battery lifespan in hours.

$$X \sim \text{Expo}(\lambda = 1/3000)$$

- Find the probability that a random battery has a lifespan of more than 2500 hours.

$$P(X \geq 2500) = 1 - F_X(2500) = e^{-2500/3000} = 0.565$$

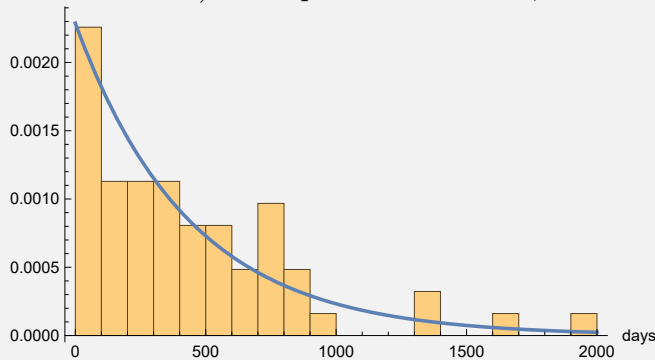
- Find the probability that a random battery has a lifespan of more than 2500 hours, given that it has already worked for 2000 hours.

According to **Fact 4.17** all that matters for an exponential rv is **the remaining time s until the first success**, which also has the $\text{Expo}(\lambda)$ distribution. So,

$$\begin{aligned} P(X \geq 2500|X > 2000) &= P(X > 500) = 1 - F_X(500) \\ &= e^{-500/3000} = 0.846 \end{aligned}$$



Example 169. Below is the histogram of time between serious (magnitude at least 7.5 or over 1000 fatalities) earthquakes worldwide, recorded from 12/16/1902 to 3/4/1977:



According to this data, the average time between serious earthquakes is 437 days. Assuming the exponential distribution for the time between earthquakes:

- if the last earthquake occurred four years ago, what is the expected time until the next earthquake?
- what is the probability of having 2 earthquakes in the next year?
- if the last earthquake occurred four years ago, what is the probability of having 2 earthquakes in the next year?

Solution: in class



4.7 Gamma (Erlang) distributions are sums of exponentials

If X_1, X_2, \dots, X_n are distributed as $\text{Expo}(\lambda)$, independently, and

$$X = X_1 + \dots + X_n$$

then $X \sim \text{Gamma}(n, \lambda)$. In general, a random variable X is said to have a $\text{Gamma}(\alpha, \beta)$ distribution when its pdf is

$$f_X(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}$$

for $x > 0$ and is 0 when $x \leq 0$. The parameter $\beta = \lambda$ is called the rate parameter and $\Gamma(a)$ is the gamma function:

$$\Gamma(a) = \int_0^\infty x^{a-1} e^{-x} dx.$$

The gamma function is a variant of the factorial function; we have $\Gamma(n) = (n-1)!$ for any positive integer n . If $X \sim \text{Gamma}(\alpha, \beta)$ then

$$\begin{aligned} E(X) &= \frac{\alpha}{\beta} \\ V(X) &= \frac{\alpha}{\beta^2} \end{aligned}$$

More on [Wikipedia](#)

Online gamma distribution [Calculator](#)

Example 170. In Example 169:

- what is the probability of having 2 earthquakes in the next year? (using Gamma)
- what is the probability that the third earthquake happens after 10 years from now?

Solution: in class □

4.8 The beta distribution: finite interval sample space

A random variable X is said to have a beta distribution with parameters α and β if its pdf is

$$f_X(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

for $0 < x < 1$ and is 0 otherwise. We then write $X \sim \text{Beta}(\alpha, \beta)$, and

$$E(X) = \alpha/(\alpha + \beta)$$

$$V(X) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

More on [Wikipedia](#)

4.9 The Bivariate Normal Distribution

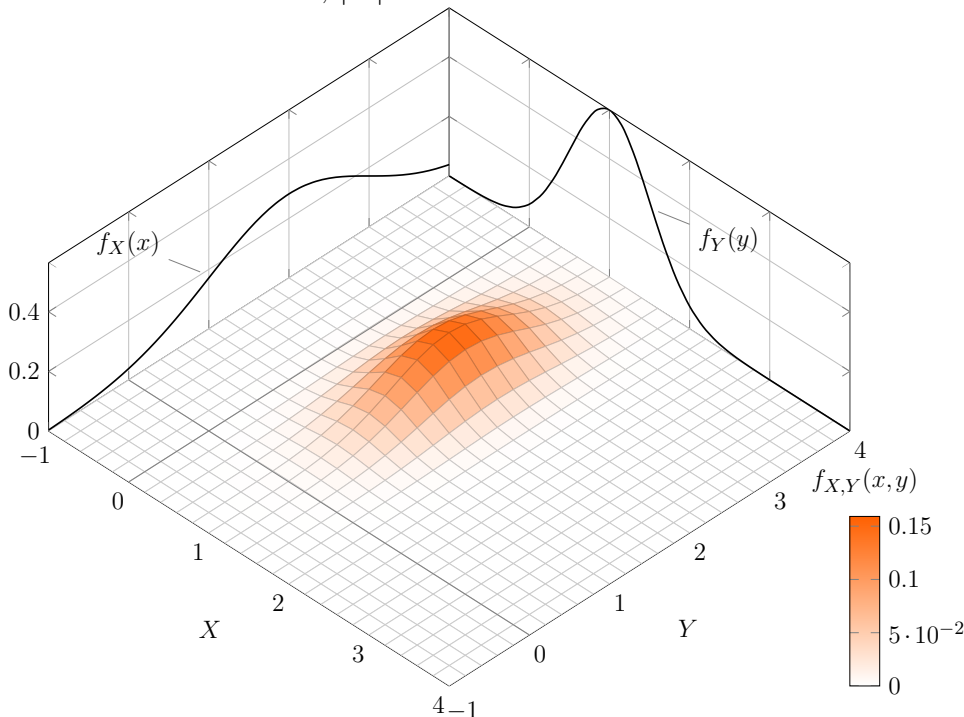
Let X_1 and X_2 have the bivariate normal joint distribution. Then, the joint pdf of (X_1, X_2) is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi|\mathbf{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right\}$$

where $\mathbf{X}^T = (X_1, X_2)$, $\boldsymbol{\mu}^T = (\mu_1, \mu_2)$ and \mathbf{V} is a full rank variance-covariance matrix, i.e.,

$$\mathbf{V}_{ij} = \text{Cov}(X_i, X_j)$$

\mathbf{V}^{-1} is the inverse of \mathbf{V} , $|\mathbf{V}|$ is the determinant of \mathbf{V} .



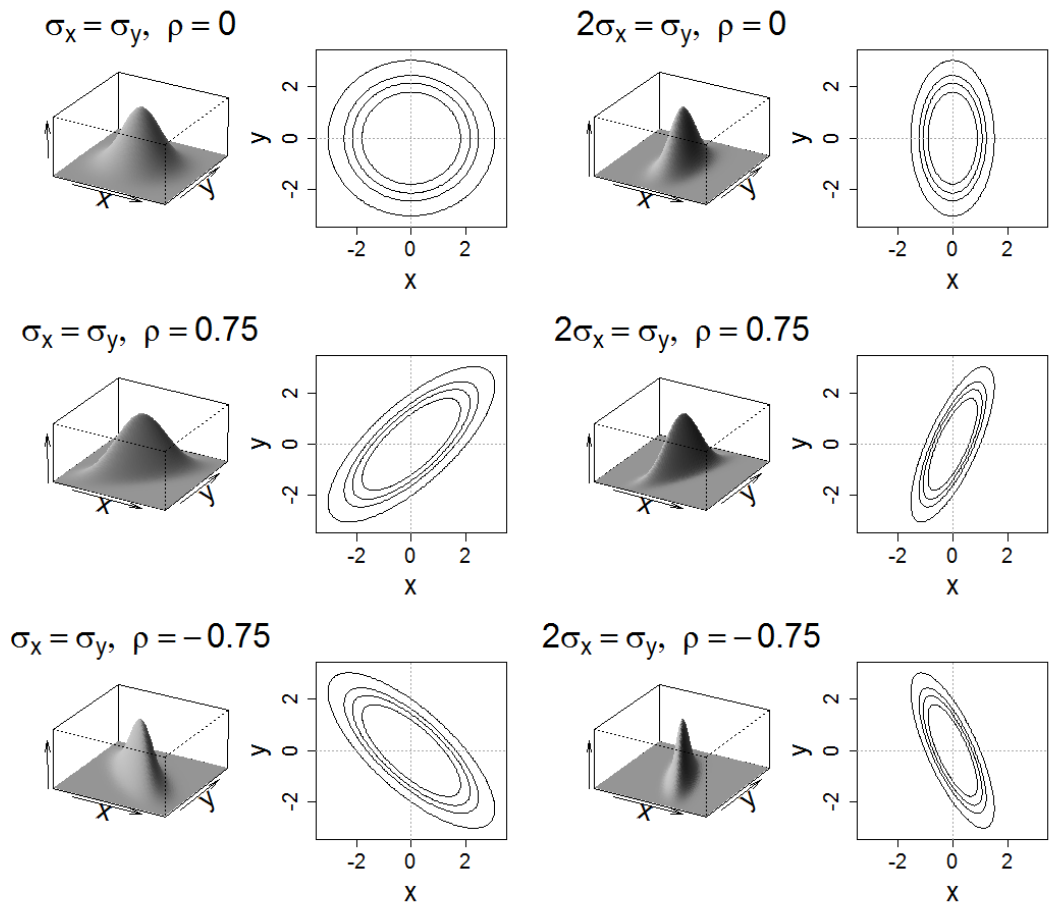
Fact 4.19 — Marginal distributions are normal. If (X_1, X_2) have a bivariate normal distribution, then the marginal distribution of X_2 is also normal with mean μ_2 and variance σ_2^2 .

Fact 4.20 — Conditional distributions are normal. If (X_1, X_2) have a bivariate normal distribution, then the conditional distribution of $X_2|X_1 = x_1$ is also normal with mean and variance given by

$$E(X_2|X_1 = x_1) = \mu_2 + \rho \frac{\sigma_2}{\sigma_1} (x_1 - \mu_1). \tag{4.17}$$

$$V(X_2|X_1 = x_1) = (1 - \rho^2)\sigma_2^2. \tag{4.18}$$

Fact 4.21 If (X_1, X_2) have a bivariate normal distribution with $\rho = 0$, X_1 and X_2 are independent.



Multi-variate normal distribution

The random vector $\mathbf{X}^T = (X_1, \dots, X_n)$ has the multivariate normal distribution if its joint density function is given by

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{n/2} |\mathbf{V}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

where $\mu^T = (\mu_1, \dots, \mu_n)$ and \mathbf{V} is a full rank variance-covariance matrix. Note that for $n = 2$ we get the density of the bivariate normal distribution.

5. Function of Random Variables

5.1 One Random Variable

We are interested in the distribution of

$$Y = g(X) \tag{5.1}$$

when the distribution of X is known.

If we only need $E(Y)$ and $V(Y)$. It is not necessary to calculate the distribution of Y :

$$E(Y) = E[g(X)] = \int g(x)f_X(x) dx$$

$$V(Y) = E[g(X)^2] - [E[g(X)]]^2$$

STOP! $E(Y) = E(g(X)) \neq g(E(X))$.

naïve approach: $g(E(X))$

correct approach: $E(g(X))$.

Example 171. X has the following distribution.

$$P(X = -1) = 0.2$$

$$P(X = 0) = 0.5$$

$$P(X = 1) = 0.3$$

Find the distribution of $Y = X^2$.

Solution: Let

$$Y = X^2$$

Using option one above, we have

$$\begin{aligned} P(Y = 0) &= P(X^2 = 0) \\ &= P(X = 0) \\ &= 0.5 \end{aligned}$$

$$\begin{aligned} P(Y = 1) &= P(X^2 = 1) \\ &= P(X = -1) + P(X = 1) \\ &= 0.5 \end{aligned}$$

□

Example 172. The probability density function of X is given by the Uniform distribution in $(0, 1)$:

$$f_X(x) = \begin{cases} 1 & ; \quad 0 \leq x \leq 1 \\ 0 & ; \quad \text{otherwise} \end{cases}$$

Find the distribution of $Y = e^X$.

Solution: Let $Y = e^X$. Therefore,

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(e^X \leq y) = P(X \leq \log y) = F_X(\log y) \\ &= \int_{-\infty}^{\log y} f_X(x) dx = \int_0^{\log y} dx = \log y \end{aligned}$$

Therefore, differentiating,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{d \log y}{dy} = \frac{1}{y}$$

□

5.1.1 Single Discrete Random Variable

If the function $g(X)$ is monotonic then, the recipe is

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) & \text{if } g^{-1}(y) \in S_X \\ 0 & \text{otherwise} \end{cases} \quad (5.2)$$

Notice from the figure that:

$$S_Y = \{y_1, y_2 \dots y_5\}$$

$$P(Y = y_i) = P(X = x_i), \quad i = 1, 2, \dots, 5$$

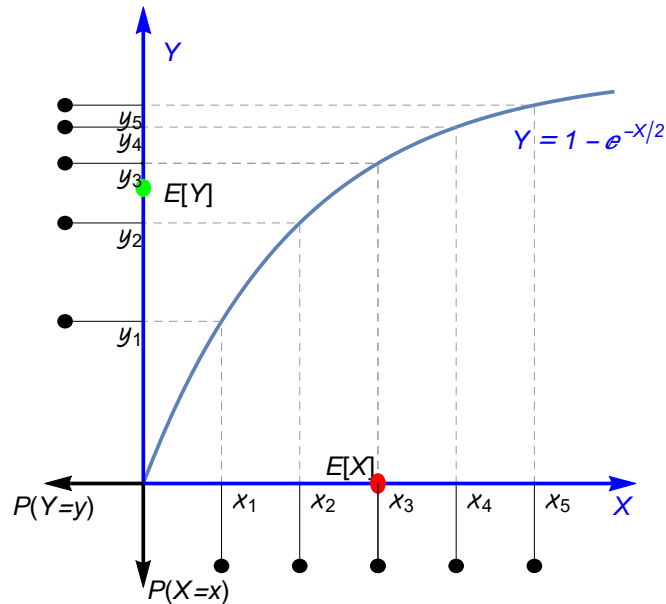
$$x_i = g^{-1}(y_i) = -2 \log(1 - y_i)$$

$$P(Y = y_i) = P(X = g^{-1}(y_i))$$

or,

$$p_Y(y_i) = p_X(g^{-1}(y_i))$$

Notice: $E(g(X)) < g(E(X))$
(always true for concave functions)



From the figure for $g(X) = 1/X$:

$$S_Y = \{y_1, y_2 \dots y_5\}$$

$$P(Y = y_i) = P(X = x_i), \quad i = 1, 2, \dots, 5$$

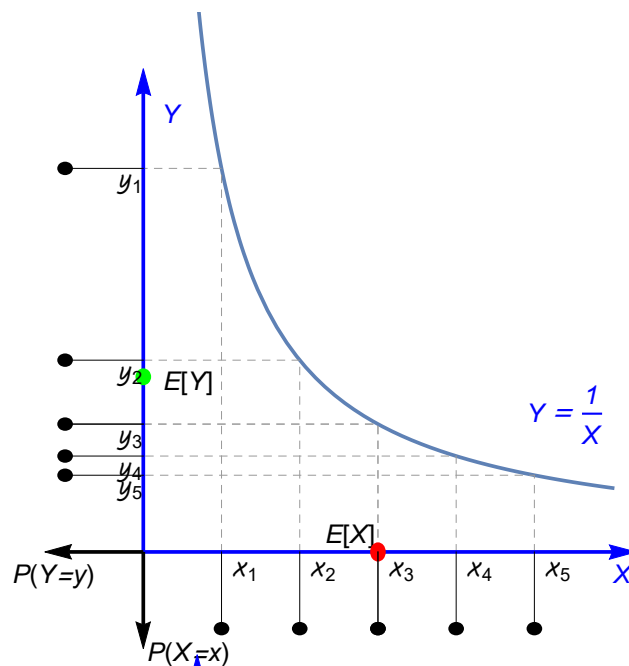
$$x_i = g^{-1}(y_i) = 1/y_i$$

$$P(Y = y_i) = P(X = g^{-1}(y_i))$$

or,

$$p_Y(y_i) = p_X(g^{-1}(y_i))$$

Notice: $E(g(X)) > g(E(X))$
(always true for convex functions)



From the figure for $g(X) = X^3$:

$$S_Y = \{y_1, y_2 \dots y_{10}\}$$

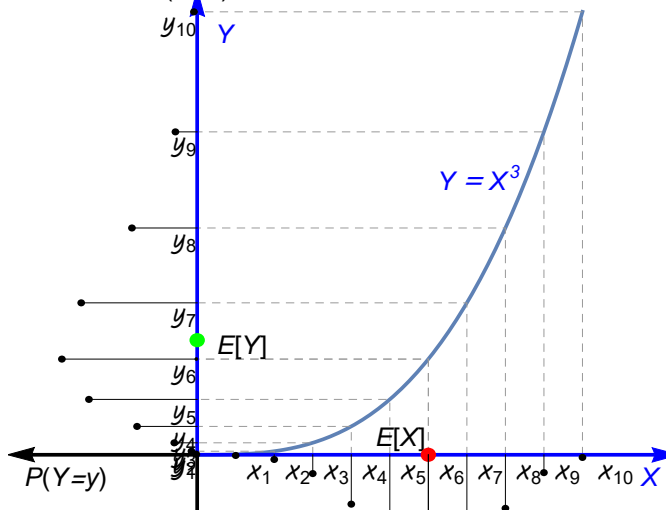
$$P(Y = y_i) = P(X = x_i), \quad i = 1, 2, \dots, 10$$

$$x_i = g^{-1}(y_i)$$

$$P(Y = y_i) = P(X = g^{-1}(y_i))$$

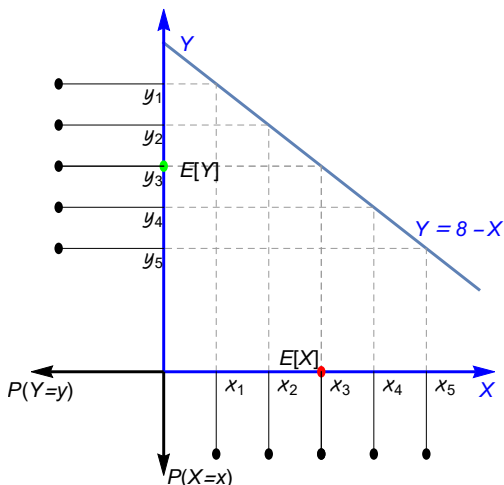
Therefore,

$$p_Y(y_i) = p_X(g^{-1}(y_i))$$



Notice: $E(g(X)) > g(E(X))$

When $g(x)$ is **linear**, the shape of the distribution remains the same:



For **non-monotonic functions**, suppose the solution of $y = g(x)$ has k roots: $x_1^*, x_2^*, \dots, x_k^*$. Therefore,

$$p_Y(y) = P(X = x_1^*) \cup P(X = x_2^*) \cup \dots \cup P(X = x_k^*)$$

$$= \sum_{i=1}^k P(X = x_i^*)$$

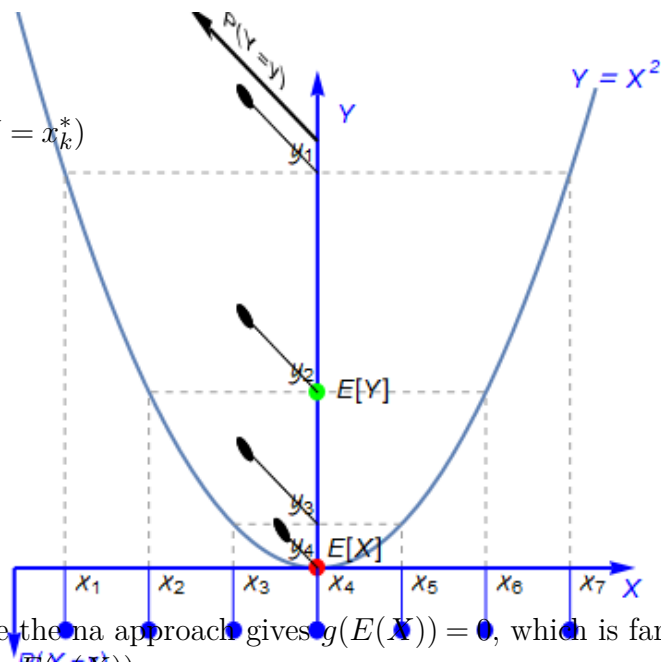
In the figure, $x_1^* = \sqrt{y}$ and $x_2^* = -\sqrt{y}$.

$$S_Y = \{y_1, y_2 \dots y_4\}$$

$$p_Y(y_i) = p_X(\sqrt{y_i}) + p_X(-\sqrt{y_i}), i = 1, 2, 3$$

$$p_Y(y_4) = p_X(0)$$

Notice: $E(g(X)) > g(E(X))$



In this case this is the worst case scenario because the naive approach gives $g(E(X)) = 0$, which is far different from the correct approach of calculating $E(g(X))$.

From the above figures we have clarified the following important inequalities,

Jensen's inequalities :

If $g(X)$ is **convex** then:

$$E(g(X)) \geq g(E(X))$$

If $g(X)$ is **concave** then:

$$E(g(X)) \leq g(E(X))$$

If $g(X)$ is **linear** then:

$$E(g(X)) = g(E(X))$$

Example 173. Given $X \sim \text{Bin}(n, p)$ and $Y = e^X$. What is the distribution of Y , $p_Y(y)$?

Solution: The recipe:

$$p_Y(y) = \begin{cases} p_X(g^{-1}(y)) & \text{if } g^{-1}(y) \in S_X \\ 0 & \text{otherwise} \end{cases}$$

Here, $x = g^{-1}(y) = \log y$ and $p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$, so:

$$p_Y(y) = \begin{cases} \binom{n}{\log y} (p)^{\log y} (1-p)^{n-\log y} & \text{when } \log y \text{ is an integer} \\ 0 & \text{otherwise} \end{cases}$$

Note: $S_Y = \{1, e, e^2, \dots, e^n\}$

□

Example 174. $X \sim \text{Bin}(n, p)$ and $Y = X^2$. What is the distribution of Y , $p_Y(y)$?

Solution: $x = g^{-1}(y) = \pm\sqrt{y}$, so $x_1^* = \sqrt{y}$ and $x_2^* = -\sqrt{y}$. In this case $x_2^* \notin S_X$ because it is negative, and therefore

$$p_Y(y) = \begin{cases} \binom{n}{\sqrt{y}} (p)^{\sqrt{y}} (1-p)^{n-\sqrt{y}} & \text{when } \sqrt{y} \text{ is an integer} \\ 0 & \text{otherwise} \end{cases}$$

□

5.1.2 Single Continuous Random Viable

Here $Y = g(X)$ is a monotone function and $f_X(x)$ is known. The PDF $f_Y(y)$ is

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \quad (5.3)$$

The derivation is analogous to the discrete case, where the key idea was $P(Y = y) = P(X = g^{-1}(y))$. In the continuous case this reads:

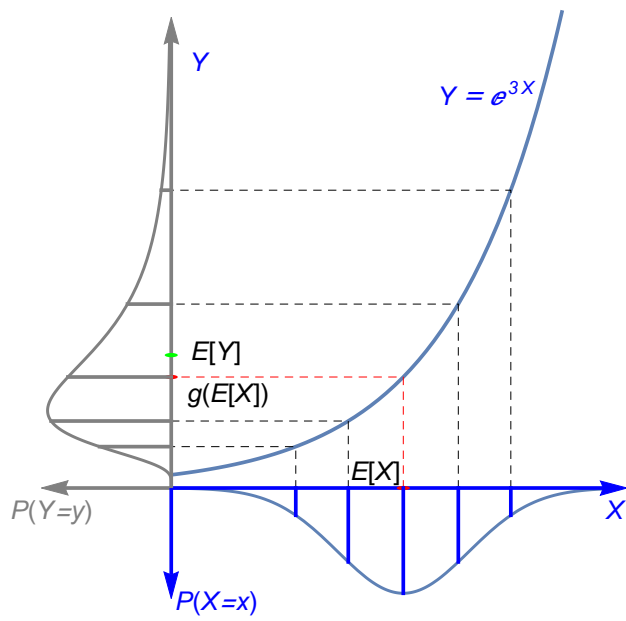
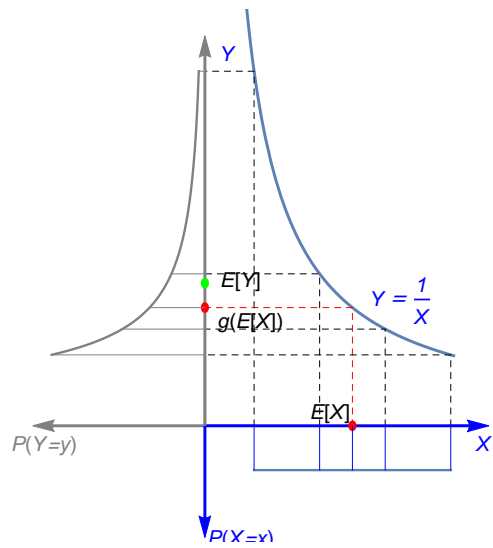
$$f_Y(y) dy = f_X(g^{-1}(y)) dx$$

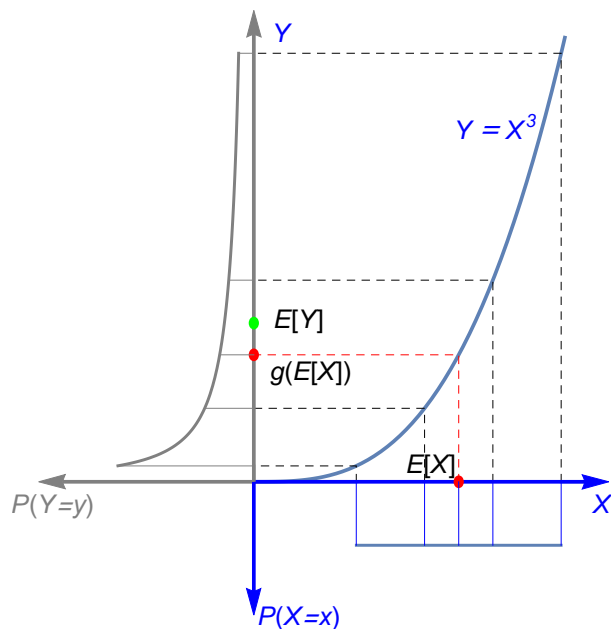
and from the figure seen in class

$$dx = \left| \frac{d}{dy} g^{-1}(y) \right| dy$$

which establishes the result.

Notice how the shape of $f_Y(y)$ changes due to $Y = g(X)$:





Example 175. $Y = e^X$, $X \sim N(\mu, \sigma^2)$. Find $f_Y(y)$.

Solution:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}}{dy} \right|$$

Since $Y = e^X = g(x)$, $g^{-1}(y) = \log(y)$, and $\left| \frac{dg^{-1}}{dy} \right| = \frac{1}{y}$. Therefore,

$$f_Y(y) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-(\log y - \mu)^2/2\sigma^2}$$

which corresponds to the lognormal distribution, $Y \sim LN(\mu, \sigma^2)$. □

Example 176. $Y = 3e^X$, $X \sim \text{Expo}(\lambda)$. Find $f_Y(y)$.

Solution:

$$f_X(x) = \lambda e^{-\lambda x}$$

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}}{dy} \right|$$

Since $Y = 3e^X = g(x)$, $g^{-1}(y) = \log(y/3)$, and $\left| \frac{dg^{-1}}{dy} \right| = \frac{3}{y}$. Therefore,

$$f_Y(y) = \frac{3}{y} \lambda e^{-\lambda \log(y/3)} = 3^{\lambda+1} \lambda y^{-\lambda-1}$$

which corresponds to the ?? □

Example 177. Let the random variable X be exponentially distributed with mean 2. You are interested in $Y = e^{-2X}$. Find $f_Y(y)$.

Example 178. The absolute velocity (X) of particles in a gas follows a Maxwell distribution, with the PDF

$$f_X(x) = \begin{cases} \frac{4x^2}{a^3\sqrt{\pi}} \exp(-\frac{x^2}{a^2}), & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

where a is a constant. Determine the PDF $f_Y(y)$ for the particle kinetic energy $Y = \frac{1}{2}mX^2$, where m is the mass of a particle.

Solution: Answer:

$$\begin{aligned} g(X) &= \frac{1}{2}mX^2 \\ g^{-1}(y) &= \pm\sqrt{\frac{2y}{m}} \\ \left| \frac{dg^{-1}}{dy} \right| &= \frac{1}{\sqrt{2my}} \\ f_Y(y) &= f_X\left(\sqrt{\frac{2y}{m}}\right) \frac{1}{\sqrt{2my}} + f_X\left(-\sqrt{\frac{2y}{m}}\right) \frac{1}{\sqrt{2my}} \\ &= f_X\left(\sqrt{\frac{2y}{m}}\right) \frac{1}{\sqrt{2my}} + 0 \\ &= \frac{8}{a^3\sqrt{\frac{2\pi m^3}{y}}} e^{-\frac{2y}{my^2}}, y > 0 \end{aligned}$$

□

5.2 Two Random Variables

Here

$$Z = g(X, Y)$$

When X, Y are discrete, assuming $p_{X,Y}$ is known:

$$p_Z(z) = \sum_{\text{all } (x,y): z=g(x,y)} p_{X,Y}(x, y)$$

If the function $g(X, Y)$ is monotone:

$$\begin{aligned} p_Z(z) &= \sum_{x \in S_X} p_{X,Y}(x, g^{-1}) \quad \text{with: } g^{-1} = g^{-1}(x, z) \quad \text{or:} \\ &= \sum_{y \in S_Y} p_{X,Y}(g^{-1}, y) \quad \text{with: } g^{-1} = g^{-1}(y, z) \end{aligned}$$

Example 179. Suppose $Z=X+Y$ where $X \sim \text{Poi}(\lambda)$, $Y \sim \text{Poi}(\mu)$ and X and Y are independent. What is the $p_Z(z)$?

Solution: Given: $p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}$ and $p_Y(y) = \frac{\mu^y}{y!} e^{-\mu}$

$$\begin{aligned} p_Z(z) &= \sum_{\text{all } (x,y): z=x+y} p_{X,Y}(x, y) \\ &= \sum_{x \in S_X} p_{X,Y}(x, g^{-1}) \quad \text{with: } g^{-1} = g^{-1}(x, z) = z - x \\ &= \sum_{x=0}^z p_X(x) \cdot p_Y(z-x) \\ &= \sum_{x=0}^z \frac{\lambda^x}{x!} \frac{\mu^{z-x}}{(z-x)!} e^{-(\lambda+\mu)} \\ &= e^{-(\lambda+\mu)} \sum_{x=0}^z \frac{\lambda^x \mu^{z-x}}{x!(z-x)!} \\ &= \frac{(\lambda+\mu)^z}{z!} e^{-(\lambda+\mu)} \\ &= \text{Poi}(\lambda+\mu) \end{aligned}$$

□

When X, Y are continuous, assuming f_X and f_Y are known:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(g^{-1}, y) \left| \frac{\partial}{\partial z} g^{-1} \right| dy \quad \text{with: } g^{-1} = g^{-1}(z, y)$$

Alternatively, we can also use

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(x, g^{-1}) \left| \frac{\partial}{\partial z} g^{-1} \right| dx \quad \text{with: } g^{-1} = g^{-1}(x, z)$$

Example 180. Suppose $Z=X+Y$ where $X \sim \text{Exp}(\lambda)$, $Y \sim \text{Exp}(\mu)$ and X and Y are independent. What is $f_Z(z)$?

Solution: Given: $f_X(x) = \lambda e^{-\lambda x}$ and $f_Y(y) = \mu e^{-\mu y}$

Since we know:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(g^{-1}, y) \left| \frac{dg^{-1}}{dz} \right| dy \quad \text{with } g^{-1} = g^{-1}(z, y)$$

Since X and Y are independent:

$$f_{X,Y}(x,y) = f_X \cdot f_Y = \lambda\mu e^{-\lambda x + \mu y}.$$

To obtain $\left| \frac{dg^{-1}}{dz} \right|$, we let $x = g^{-1} = z - y$. Therefore:

$$\left| \frac{dg^{-1}}{dz} \right| = \left| \frac{d}{dz}(z - y) \right| = |1| = 1$$

and

$$f_{X,Y}(g^{-1}, y) = \lambda\mu e^{-(\lambda(z-y) + \mu y)} dy$$

Therefore, $f_Z(z)$ can be calculated as (from the figure seen in class):

$$\begin{aligned} f_Z(z) &= \int_0^z f_{X,Y}(z-y, y) \left| \frac{dg^{-1}}{dz} \right| dy \\ &= \lambda\mu e^{-\lambda z} \int_0^z e^{-y(\mu-\lambda)} dy \\ &= \frac{\lambda\mu}{\mu-\lambda} (e^{-\lambda z} - e^{-\mu z}) \end{aligned}$$

Note: If X and Y have the same rate, it would be a Gamma (Erlang) distribution. \square

Example 181. Suppose $Z = X \cdot Y$ where $X \sim \text{Exp}(\lambda)$, $Y \sim \text{Exp}(\mu)$ and X and Y are independent. What is $f_Z(z)$?

Solution:

$$y = \frac{z}{x}, \left| \frac{dg^{-1}}{dz} \right| = \frac{1}{x}$$

$$f_Z(z) = \int_0^\infty \lambda\mu e^{-(\lambda x + \mu \frac{z}{x})} \frac{1}{x} dx$$

\square

Example 182. Suppose $Z = X + Y$ where $X \sim N(\mu_X, \sigma_X)$, $Y \sim N(\mu_Y, \sigma_Y)$ and X and Y are independent. Show that:

$$Z \sim N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$$

5.2.1 What if the distribution of X is unknown? (not covered)

If we only know $E(X) = \mu$ and $V(X) = \sigma^2$, we can still approximate $E(Y)$ and $V(Y)$ by Taylor Series around the mean of X :

$$Y = g(X) \approx g(\mu) + (X - \mu)g'(\mu) + \frac{1}{2}(X - \mu)^2 g''(\mu) + \dots$$

2nd-order approximation for $E[Y]$:

$$E(Y) \approx g(\mu) + \frac{1}{2}g''(\mu)\sigma^2 \quad (5.4)$$

1st-order approximation for $V[Y]$:

$$V(Y) \approx g'(\mu)^2\sigma^2 \quad (5.5)$$

Several random variables

Let $Y = g(X_1, X_2, \dots, X_n)$ and recall the vector notation:

$$\mathbf{X} = (X_1, X_2, \dots, X_n)^T$$

The joint PMF is not known; all we know are:

$$E(X_i) = \mu_i \quad , \quad V(X_i) = \sigma_i^2 \quad , \quad \text{Cov}(X_i, X_j) = \sigma_{ij}$$

and $\boldsymbol{\mu} = \{\mu_1, \mu_2, \dots, \mu_n\}$. A second-order Taylor series expansion of the scalar-valued function $g(\cdot)$ can be written compactly as

$$g(\mathbf{X}) = g(\boldsymbol{\mu}) + (\mathbf{X} - \boldsymbol{\mu})^T \mathbf{G} + \frac{1}{2!} (\mathbf{X} - \boldsymbol{\mu})^T \mathbf{H} (\mathbf{X} - \boldsymbol{\mu}) + \dots$$

where \mathbf{G} and \mathbf{H} are the gradient vector and the Hessian matrix of g evaluated at $\mathbf{X} = \boldsymbol{\mu}$, resp. 2nd-order approximation for $E(Y)$:

$$\begin{aligned} E(Y) &\approx g(\boldsymbol{\mu}) + \frac{1}{2} \mathbf{e}^T (\Sigma_{\mathbf{X}} \odot \mathbf{H}) \mathbf{e} \\ &= g(\boldsymbol{\mu}) + \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \sigma_{ij} \cdot \left(\frac{\partial^2 g}{\partial X_i \partial X_j} \right) \\ &= g(\boldsymbol{\mu}) + \frac{1}{2} \sum_{i=1}^n \sigma_i^2 \cdot \left(\frac{\partial^2 g}{\partial X_i^2} \right) + \underbrace{\sum_{i=1}^n \sum_{j=i+1}^n \sigma_{ij} \cdot \left(\frac{\partial^2 g}{\partial X_i \partial X_j} \right)}_{0 \text{ if } X_i \text{'s are independent}} \end{aligned} \quad (5.6)$$

where \mathbf{e} is the column vector whose entries are all 1's, and \odot is the Hadamard product, which takes two matrices of the same dimensions and produces another matrix where each element i, j is the product of elements i, j of the original two matrices. It should not be confused with the more common matrix product.

1st-order approximation for $V(Y)$:

$$\begin{aligned} V(Y) &\approx \mathbf{G}^T \Sigma_{\mathbf{X}} \mathbf{G} \\ &= \sum_{i=1}^n \sigma_i^2 \cdot \left(\frac{\partial g}{\partial X_i} \right)^2 \end{aligned} \quad (5.7)$$

Example 183. — * **The hydraulic head loss** in a pipe may be determined by the Darcy-Weisbach equation as follows:

$$H = \frac{fLV^2}{2Dg}$$

where:

L =length of a pipe, V =flow velocity of water in a pipe, D =pipe diameter, f =coefficient of friction, g =gravitational acceleration=32.2 ft/sec². Suppose a pipe has the following properties:

i	X_i	μ_i	δ_i
1	L	100.	0.1
2	D	1.	0.1
3	f	0.02	0.2
4	V	10.	0.15

(a) Approximate the mean and standard deviation of the hydraulic head loss of the pipe.

Solution: (a) We have:

	X_i	μ_i	σ_i
1	L	100.	10.
2	D	1.	0.1
3	f	0.02	0.004
4	V	10.	1.5

$$\mathbf{G} = \begin{pmatrix} \frac{fV^2}{2Dg} \\ -\frac{fLV^2}{2D^2g} \\ \frac{LV^2}{2Dg} \\ \frac{fLV}{Dg} \end{pmatrix} \text{ and } \mathbf{H} = \begin{pmatrix} 0 & -\frac{fV^2}{2D^2g} & \frac{V^2}{2Dg} & \frac{fV}{Dg} \\ -\frac{fV^2}{2D^2g} & \frac{fLV^2}{D^3g} & -\frac{LV^2}{2D^2g} & -\frac{fLV}{D^2g} \\ \frac{V^2}{2Dg} & -\frac{LV^2}{2D^2g} & 0 & \frac{LV}{Dg} \\ \frac{fV}{Dg} & -\frac{fLV}{D^2g} & \frac{LV}{Dg} & \frac{fL}{Dg} \end{pmatrix}, \text{ evaluating at } \mu \text{ gives:}$$

$$\mathbf{G} = \begin{pmatrix} 0.031 \\ -3.106 \\ 155.28 \\ 0.621 \end{pmatrix} \text{ and } \mathbf{H} = \begin{pmatrix} 0. & -0.03 & 1.55 & 0.01 \\ -0.03 & 6.21 & -155.28 & -0.62 \\ 1.55 & -155.28 & 0. & 31.06 \\ 0.01 & -0.62 & 31.06 & 0.06 \end{pmatrix}$$

and the mean and variance of H are approximately:

$$\begin{aligned} E(H) &= g(\boldsymbol{\mu}) + \frac{1}{2} \sum_{i=1}^n \sigma_i^2 \cdot \left(\frac{\partial^2 g}{\partial X_i^2} \right) \\ &= \frac{0.02 \times 100 \times 10^2}{2 \times 1 \times 32.2} + \frac{1}{2} (6.21 \times 0.01 + 0.06 \times 2.25) \\ &= 3.20652 \end{aligned}$$

$$\begin{aligned}
 V(H) &= \sum_{i=1}^n \sigma_i^2 \cdot \left(\frac{\partial g}{\partial X_i} \right)^2 \\
 &= 0.000961 \times 100. + 9.64724 \times 0.01 + 24111.9 \times 0.000016 + 0.385641 \times 2.25 \\
 &= 1.44605
 \end{aligned}$$

□

Example 184. Refer to example 183 and assume that the correlation between D and f is 0.7 and between V and f , 0.4. Show that the expected value and variance of H are now 3.237 and 1.639, respectively.

Solution: Hint:

$$\begin{aligned}
 E(H) &= \frac{0.02 \times 100 \times 10^2}{2 \times 1 \times 32.2} + \frac{1}{2} (-155.28\sigma_{2,3} - 155.28\sigma_{3,2} + 31.056\sigma_{3,4} + 31.056\sigma_{4,3} + \\
 &\quad + 6.211\sigma_2^2 + 0.062\sigma_4^2) \\
 V(H) &= -3.106 (155.28\sigma_{3,2} - 3.106\sigma_2^2) + 0.621 (155.28\sigma_{3,4} + 0.621\sigma_4^2) + \\
 &\quad + 155.28 (-3.106\sigma_{2,3} + 0.621\sigma_{4,3} + 155.28\sigma_3^2) + 0.000961\sigma_1^2
 \end{aligned}$$

□

5.3 Important distributions for statistics

5.3.1 The chi-square distribution with r degrees of freedom

The density function is given by:

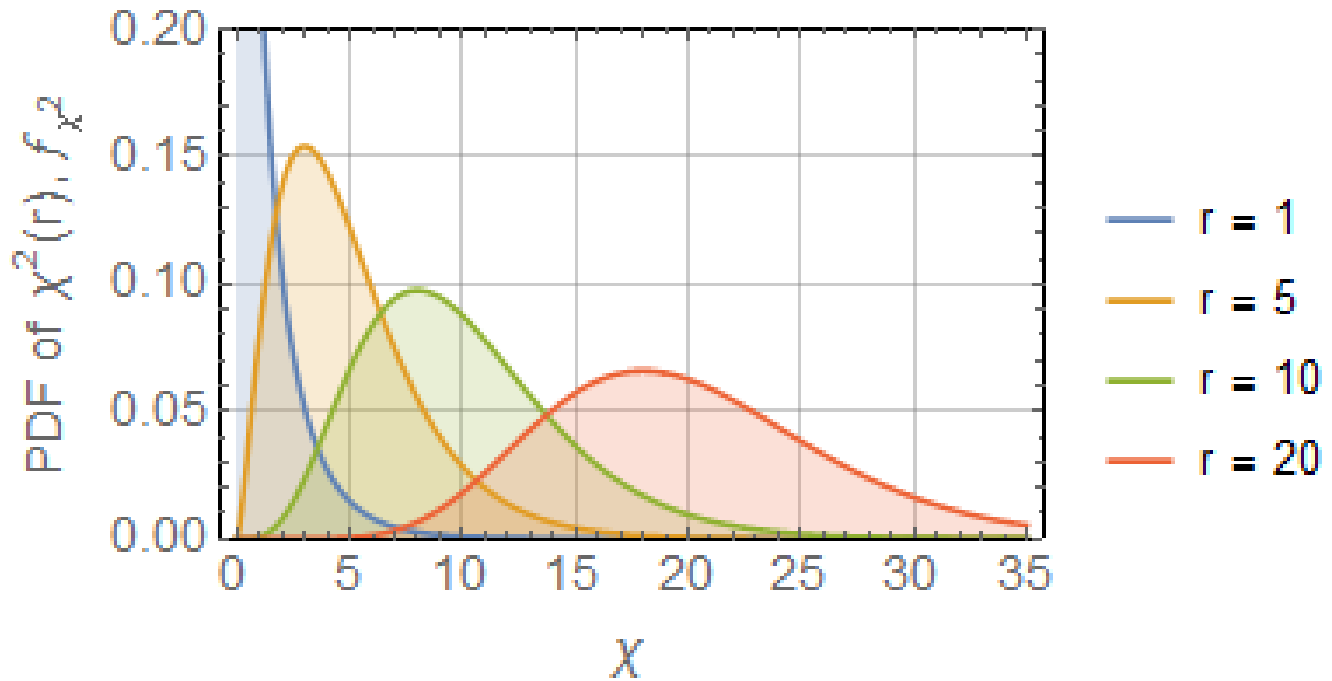
$$f_{\chi^2}(x) = \frac{2^{-r/2} e^{-x/2} x^{r/2-1}}{\Gamma\left(\frac{r}{2}\right)}, \quad x > 0.$$

and

$$E(X) = r$$

$$V(X) = 2r$$

[Chi-Sqr probability tables.](#)



It is important because:

Fact 5.5 If Y_1, Y_2, \dots, Y_r are independent standard normal random variables, $Y_i \sim N(0,1)$, then

$$\sum_{j=1}^r Y_j^2 \sim \chi^2(r).$$

Fact 5.6 If $X \sim N(0,1)$, then $Y = X^2 \sim \chi^2(1)$.

Proof Let $Y = X^2$. Then, using the techniques in this chapter

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right) = \frac{1}{\sqrt{2\pi y}} e^{-\frac{1}{2}y}, \quad y > 0.$$

■

Fact 5.7 If $Y_1 \sim \chi^2(r)$ and $Y_2 \sim \chi^2(s)$, and are independent, then

$$Y_1 + Y_2 \sim \chi^2(r+s).$$

5.3.2 Student's t -distribution

If $U \sim N(0,1)$ and $V \sim \chi^2(r)$ are independent, then

$$T = \frac{U}{\sqrt{V/r}} \sim t(r)$$

has a t -distribution with r degrees of freedom:

$$f_T(t) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{(\pi r)}\Gamma\left(\frac{r}{2}\right)} \left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}}, \quad t \in \mathbb{R}.$$

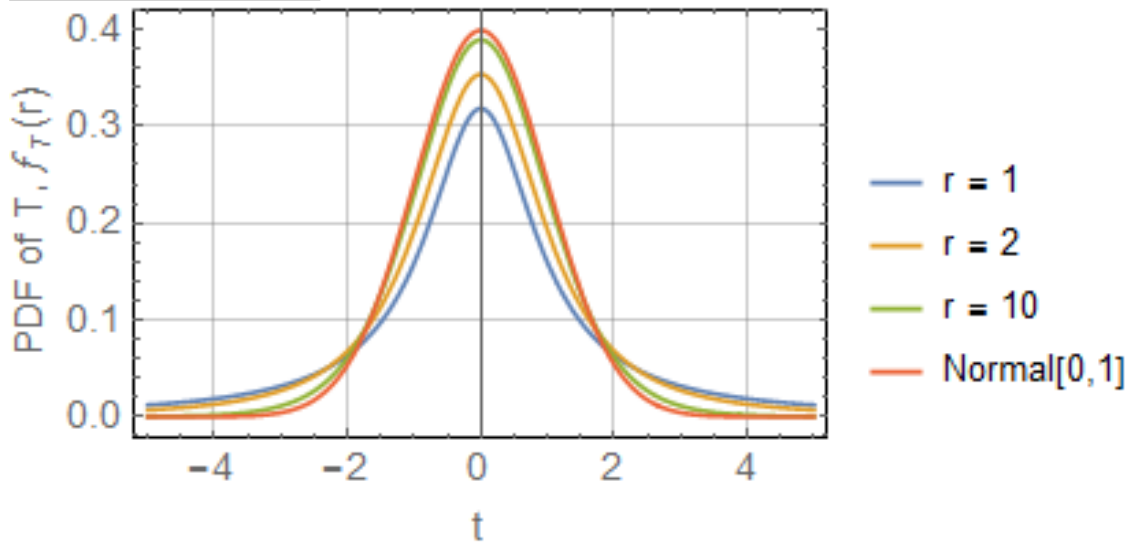
and:

$$E(T) = 0$$

$$V(T) = r/(r-2), \quad r > 2$$

Note, as $r \rightarrow \infty$ then $t(r) \rightarrow N(0, 1)$.

[t-distribution tables.](#)



n @ p a

6 Normal Random Samples 191

- 6.1 Theoretical building blocks
- 6.2 Confidence intervals
- 6.3 Hypothesis Testing

7 Linear regression 217

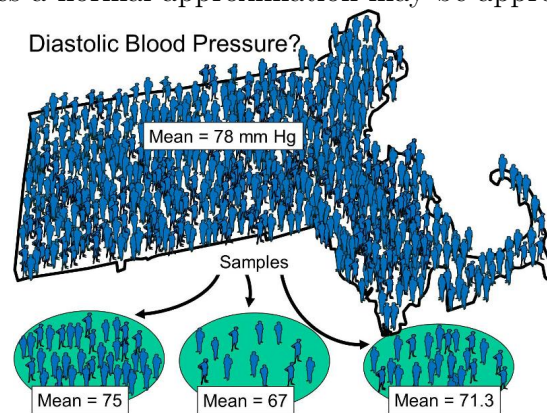
- 7.1 The regression model
- 7.2 Matrix notation
- 7.3 The method of ordinary least squares (OLS)
- 7.4 Testing the significance of coefficients
- 7.5 Goodness-of-fit: R^2
- 7.6 Assessing the model
- 7.7 Model selection
- 7.8 Making predictions
- 7.9 Simple linear regression
- 7.10 Problems

6. Normal Random Samples

Statistical theory for random samples drawn from normal distributions is very important, partly because a great deal is known about its various associated distributions and partly because the central limit theorem suggests that for large samples a normal approximation may be appropriate.

The basic assumption in statistic is that the random variable of interest, X , is distributed across a **population** according to a known distribution, typically $\text{Normal}(\mu, \sigma^2)$.

Parameters μ and σ^2 are **unknown**. **Statistics** is all about estimating them using a sample, evaluating the potential errors when the sample is small, testing hypotheses and making predictions using the available data.



Random sample. A sample of size n is a realization $\mathbf{x} = (x_1, \dots, x_n)$ of the random vector:

$$\mathbf{X} = (X_1, \dots, X_n)$$

which we call a **random sample** in this chapter. We assume X_1, \dots, X_n to be independent and identically distributed (**iid**) Normal random variables having expected value μ and variance σ^2 , denoted as:

$$X_i \stackrel{iid}{\sim} N(\mu, \sigma^2), \quad i = 1, 2, \dots, n. \quad (6.1)$$

Sample mean and sample variance

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

$$\sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

The sample mean is the “best” estimator of μ
and is normally distributed.

and

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

$$= \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

The sample variance, “best” estimator of σ^2
(shortcut formula)

Proof. (Shortcut formula for S^2)

$$\begin{aligned} (n-1)S^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i^2 + \bar{X}^2 - 2X_i\bar{X}) \\ &= \sum_{i=1}^n X_i^2 + \sum_{i=1}^n \bar{X}^2 - 2\bar{X} \sum_{i=1}^n X_i \\ &= \sum_{i=1}^n X_i^2 + n\bar{X}^2 - 2\bar{X}(n\bar{X}) \\ &= \sum_{i=1}^n X_i^2 + n\bar{X}^2 - 2n\bar{X}^2 \\ &= \sum_{i=1}^n X_i^2 - n\bar{X}^2 \end{aligned}$$

Therefore,

$$S^2 = \frac{1}{n-1} \left(\sum_{i=1}^n X_i^2 - n\bar{X}^2 \right)$$

■

STOP! **Law of large numbers:** Since $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ then

$$\bar{X} \rightarrow \mu \quad \text{as } n \rightarrow \infty$$

The standard error is the standard deviation of the sample mean:

$$SE = \sigma/\sqrt{n} \text{ or } s/\sqrt{n} \text{ if } \sigma \text{ is unknown.}$$

Example 185. The times between successive of vehicles arrivals at a toll booth were observed as follows:

$$\{1.2, 3.0, 6.3, 10.1, 5.2, 2.4, 7.1\} \text{ in sec}$$

- (a) Find the sample mean
 (b) Find the sample variance and the standard error

Solution: (a)

$$\begin{aligned} \bar{x} &= \frac{1.2 + 3.0 + 6.3 + 10.1 + 5.2 + 2.4 + 7.1}{7} \\ &= 5.04 \end{aligned}$$

(b)

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^7 (x_i - \bar{x})^2}{7 - 1} \\ &= 9.56 \\ SE &= \frac{s}{\sqrt{n}} \\ &= 1.17 \end{aligned}$$

□

6.1 Theoretical building blocks

6.1.1 The Z, T and C^2 -statistics

The standardized version of \bar{X} is the Z -statistic:

The Z -statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Proof: By the CLT

We could use the Z -statistic to calculate a range of plausible values for μ , under the assumption that σ^2 is known. **But the variance σ^2 is unknown in practice.** We need to find another statistic for the mean μ which does not contain the unknown variance parameter. It turns out that if we replace σ by S , then the distribution of the resulting statistics has the t-distribution.



The problem with the Z -statistic is that in practice the variance σ^2 is not known.

Solution: use S^2 instead.

Fact 6.1 — **The Chi-square distribution with r degrees of freedom, χ_r^2 .** If Y_1, Y_2, \dots, Y_r are independent standard normal random variables, $Y_i \sim N(0, 1)$, then

$$\sum_{j=1}^r Y_j^2 \sim \chi_r^2.$$

Additive property: If $Y_1 \sim \chi_r^2$ and $Y_2 \sim \chi_s^2$, and are independent, then

$$Y_1 + Y_2 \sim \chi_{r+s}^2$$

[Chi-Sqr probability tables.](#)

The C^2 statistic

$$C^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

Proof.

$$\begin{aligned} (n-1)S^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n ((X_i - \mu) - (\bar{X} - \mu))^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \end{aligned}$$

Dividing by σ^2 ,

$$\begin{aligned} \frac{(n-1)S^2}{\sigma^2} &= \sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 - n \left(\frac{\bar{X} - \mu}{\sigma} \right)^2 \\ &= \underbrace{\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2}_{\sim \chi_n^2} - \underbrace{\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2}_{\sim \chi_1^2} \end{aligned}$$

By the additive property of the chi-square distribution in Fact (6.1) the result follows. ■

Fact 6.2 — **Student's t -distribution.** If $U \sim N(0, 1)$ and $V \sim \chi_r^2$ are independent, then

$$T = \frac{U}{\sqrt{V/r}} \sim t_r$$

has a t -distribution with r degrees of freedom.

[t-distribution tables.](#)

One of the key results for normal random samples is the independence of \bar{X} and S^2 , and their relationship to the mean and variance parameters of a normal distribution.

Fact 6.3 — Independence of \bar{X} and S^2 for normal samples.. If X_1, X_2, \dots, X_n are independent, identically distributed random variables with normal distribution $N(\mu, \sigma^2)$, then \bar{X} and S^2 are independent, and:

- (i) $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$
- (ii) $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$

Proof (Optional) There are various methods of proof. We will use one which delivers both independence and distribution within the same argument.

$$X_i \sim N(\mu, \sigma^2) \implies Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1)$$

Now, we know that, if \mathbf{Z} is a vector of normal random variables and \mathbf{L} is a linear transformation, then $\mathbf{Y} = \mathbf{L}\mathbf{Z}$ is also a vector of normal random variables. Suppose that \mathbf{L} is orthogonal so that $\mathbf{L}^T\mathbf{L} = \mathbf{I}$. Then

$$\mathbf{Y}^T\mathbf{Y} = \mathbf{Z}^T\mathbf{L}^T\mathbf{L}\mathbf{Z} = \mathbf{Z}^T\mathbf{Z} \quad \text{or} \quad \sum_{i=1}^n Y_i^2 = \sum_{i=1}^n Z_i^2.$$

Thus, the Y_i variables are also independent and distributed as $N(0, 1)$.

Now suppose we choose \mathbf{L} such that its first row is

$$\left(\frac{1}{\sqrt{n}}, \frac{1}{\sqrt{n}}, \dots, \frac{1}{\sqrt{n}} \right).$$

Then $Y_1 = \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i = \sqrt{n}\bar{Z}$, and

$$\sum_{i=1}^n (Z_i - \bar{Z})^2 = \sum_{i=1}^n Z_i^2 - n\bar{Z}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \sum_{i=2}^n Y_i^2,$$

which is independent of Y_1 . Thus

$$\sum_{i=1}^n (Z_i - \bar{Z})^2 \text{ is independent of } \bar{Z} \quad \implies \quad \sum_{i=1}^n (X_i - \bar{X})^2 \text{ is independent of } \bar{X}$$

since $Z_i = \frac{X_i - \mu}{\sigma}$. The independence of \bar{X} and S^2 is therefore proved.

$$(i) \quad Y_1 \sim N(0, 1) \quad \implies \quad \sqrt{n}\bar{Z} \sim N(0, 1) \quad \implies \quad \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \sim N(0, 1) \quad \implies \quad \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right);$$

(ii) From Theorem 6.1,

$$\begin{aligned} \sum_{i=2}^n Y_i^2 \sim \chi^2(n-1) &\implies \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1) \\ &\implies \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1). \end{aligned}$$

■

The T -statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Proof Recall that:

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1), \quad C^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

which are independent since \bar{X} and S^2 are independent. It turns out that T can be expressed as:

$$T = \frac{Z}{\sqrt{C^2/(n-1)}}$$

which corresponds to the definition of the t-distribution with $n - 1$ degrees of freedom!

6.1.2 Estimators

Parameters μ and σ^2 are **unknown**, so how can we estimate them?

Suppose we have a random sample $X = X_1, X_2, \dots, X_n$ drawn from a distribution with some parameter θ .

Estimators An *estimator* $\hat{\theta}$ of θ is a function of the observed data which (we hope) forms a useful approximation of the parameter:

$$\hat{\theta} = g(X_1, X_2, \dots, X_n).$$

Note that $\hat{\theta}$:

1. can depend only on the observed data, and not on any unknown parameters,
2. is itself a random variable, with a distribution, mean, variance, etc.

Examples of estimators are the sample mean and sample variance. The estimator is a function of random variables, so If the estimator is to have any use at all, it should have some nice properties. For example, we know that $\bar{X} \rightarrow \mu$ by the law of large numbers, ensuring that \bar{X} is a sensible estimator for μ .

Desirable properties of a “good” estimators

Unbiased: $\hat{\theta}$ is said to be *unbiased* if

$$E(\hat{\theta}) = \theta,$$

Consistency: $\hat{\theta}$ is said to be *consistent* if $\hat{\theta} \rightarrow \theta$ as $n \rightarrow \infty$

Efficiency (Minimum variance): $\hat{\theta}_A$ is said to be more *efficient* than $\hat{\theta}_B$ if

$$V(\hat{\theta}_A) < V(\hat{\theta}_B)$$

Even if we assume unbiasedness and consistency to be desirable, it is possible to have more than one such estimator.

Example 186. — Show that \bar{X} and S^2 are unbiased

$$\begin{aligned} E(\bar{X}) &= E\left(\frac{1}{n} \sum_{i=1}^n X_i\right) = \frac{1}{n} \sum_{i=1}^n E(X_i) = \frac{1}{n} \sum_{i=1}^n \mu \\ &= \mu \end{aligned}$$

Now for S^2 ,

$$\begin{aligned} (n-1)S^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n ((X_i - \mu) - (\bar{X} - \mu))^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \end{aligned}$$

Therefore,

$$\begin{aligned} E[(n-1)S^2] &= E\left[\sum_{i=1}^n (X_i - \mu)^2\right] - nE[(\bar{X} - \mu)^2] \\ &= \sum_{i=1}^n V(X_i) - nV(\bar{X}) \\ &= n\sigma^2 - n\frac{\sigma^2}{n} \\ &= (n-1)\sigma^2 \end{aligned}$$

Therefore,

$$E(S^2) = \sigma^2$$

which means that S^2 is unbiased.

6.2 Confidence intervals

Let $\mathbf{X} = (X_1, X_2, \dots, X_n)$ represent a random sample, and \mathbf{x} a realization. If $(a(\mathbf{X}), b(\mathbf{X}))$ is a random interval such that

$$P(a(\mathbf{X}) < \theta < b(\mathbf{X})) = 1 - \alpha,$$

then a realization of that interval, $(a(\mathbf{x}), b(\mathbf{x}))$ is said to be a $100(1 - \alpha)\%$ confidence interval for the parameter θ .

Interpretation of confidence intervals It is not easy to get to grips with what is meant by a confidence interval. One **cannot** say:

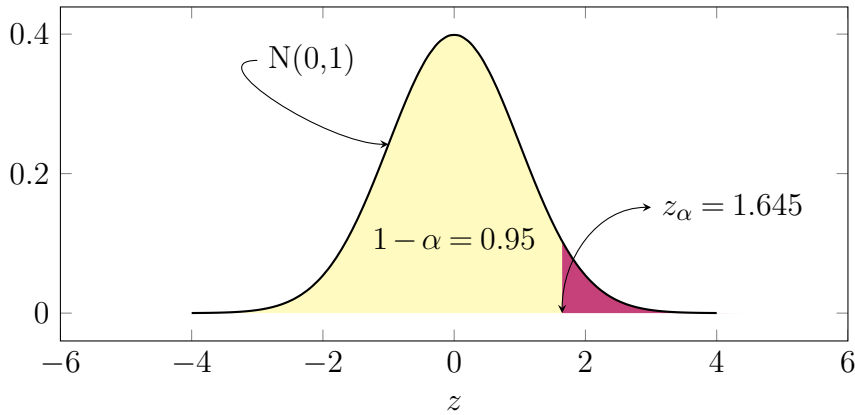
“the parameter μ has probability $(1 - \alpha)$ of lying within the calculated interval $(a(\mathbf{x}), b(\mathbf{x}))$ ”

because that statement has no random variables. Since the ends of the interval are fixed numbers,

as is θ , and without random variables being present, probability statements cannot be made: either θ lies between the two numbers or it doesn't, and we have no way of knowing which. The only **viable interpretation** is to say that we have used a procedure which, if repeated over and over again, would give intervals containing the parameter $100(1 - \alpha)\%$ of the time.

Critical value z_α . The symbol z_α is the $(1 - \alpha)$ -percentile of Z . Note that since the $N(0,1)$ is symmetric with respect to zero, so:

$$z_{1-\alpha} = -z_\alpha$$



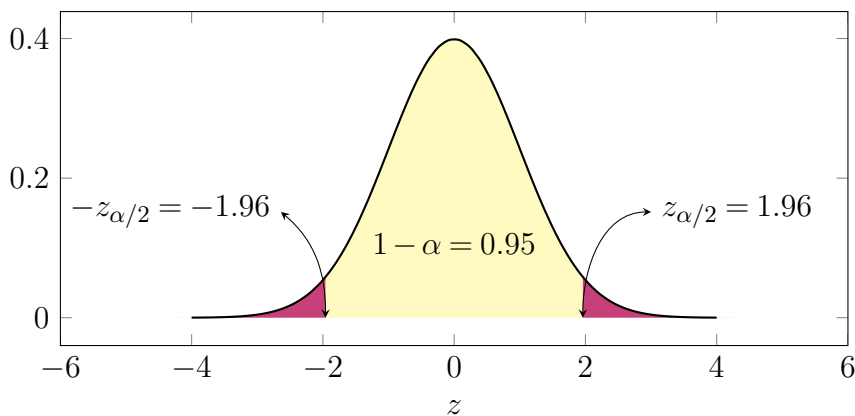
The most common value of α in use is 0.05, in which case:

$$z_\alpha = z_{0.05} = 1.654, \quad z_{\alpha/2} = z_{0.025} = \mathbf{1.96}$$

6.2.1 Confidence intervals for μ

Two-sided $100(1 - \alpha)\%$ confidence intervals. Since $Z \sim N(0,1)$, and the definition of critical values $z_{\alpha/2}$, it follows that

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha.$$



Since $Z = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}}$, then

$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

$$\implies P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Hence the appropriate **random interval** is

$$\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right), \quad \text{whose realization gives:}$$

Confidence interval for μ when σ is known The $100(1 - \alpha)\%$ confidence interval is

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right).$$

The margin of error is the half-width, h , of the confidence interval:

$$h = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

If we solve for n in the above equation we obtain:

Sample size needed for a prescribed margin of error h :

$$n = \left(\frac{\sigma z_{\alpha/2}}{h}\right)^2$$

Example 187. The times between successive arrivals of vehicles at a toll booth were observed as follows:

$$\{1.2, 3.0, 6.3, 10.1, 5.2, 2.4, 7.1\} \text{ in sec}$$

- Find the sample mean and the sample standard deviation.
- Find the 95%-confidence interval for μ when σ **is assumed to be equal to the sample standard deviation**.
- Find the margin of error.
- Find the sample size needed to reduce the margin of error by a factor of two.

Solution:

- Find the sample mean and the sample standard deviation.
 $\bar{x} = 5.04, s = \sqrt{9.56} = 3.1$
- Find the confidence interval for μ when σ **is assumed to be equal to the sample standard deviation**.

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) =$$

c) Find the margin of error.

$$h = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} =$$

d) Find the sample size needed to reduce the margin of error by a factor of two.

$$n = \left(\frac{\sigma z_{\alpha/2}}{h} \right)^2 =$$

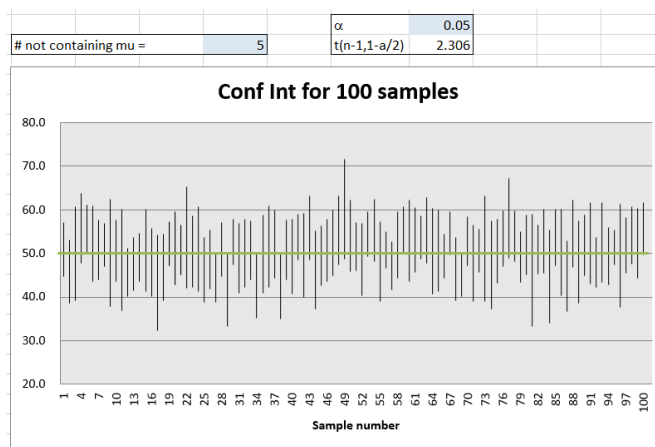
□

Recall that the only **viable interpretation** of confidence intervals is to say that we have used a procedure which, if repeated over and over again, would give intervals containing the parameter $100(1 - \alpha)\%$ of the time.

This is shown in the figure, we were repeated the following procedure 100 times:

1. take a sample of size $n = 7$: in Excel we generate seven random numbers from the $N(\mu = 50, \sigma^2 = 10)$ distribution.
2. compute the confidence interval with the above recipe
3. plot all these intervals next to each other

→ See and modify this chart in Excel



Unfortunately we do not know σ , so what should we do?

→ We **replace σ by s** , which boils down to replacing the Z -statistic with the T -statistic. Writing $t_{\alpha/2}$ for the critical values from the distribution t_{n-1} , we have

$$P\left(-t_{\alpha/2} < \frac{(\bar{X} - \mu)}{S/\sqrt{n}} < t_{\alpha/2}\right) = 1 - \alpha.$$

Re-arranging gives the random interval

$$\left(\bar{X} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + t_{\alpha/2} \frac{s}{\sqrt{n}}\right),$$

and the $100(1 - \alpha)\%$ confidence interval is a **realization** of this interval.

Confidence interval for μ when σ is unknown The $100(1 - \alpha)\%$ confidence interval is

$$\left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}\right).$$

In some circumstances, it can make more sense to express the confidence interval in only one direction – to either the lower or upper confidence limit.

One-sided CIs for μ when σ is unknown :

1. The $100(1 - \alpha)\%$ **lower** confidence interval is

$$\left(-\infty, \bar{x} + t_{\alpha} \frac{s}{\sqrt{n}}\right), \quad \text{Note: } \bar{x} + t_{\alpha} \frac{s}{\sqrt{n}} \text{ is an } \mathbf{upper} \text{ bound.}$$

2. The $100(1 - \alpha)\%$ **upper** confidence interval is

$$\left(\bar{x} - t_{\alpha} \frac{s}{\sqrt{n}}, \infty\right), \quad \text{Note: } \bar{x} - t_{\alpha} \frac{s}{\sqrt{n}} \text{ is a } \mathbf{lower} \text{ bound.}$$

Example 188. The times between successive arrivals of vehicles at a toll booth were observed as follows:

$$\{1.2, 3.0, 6.3, 10.1, 5.2, 2.4, 7.1\} \text{ in sec}$$

- Find the sample mean and the sample standard deviation.
- Find the 95%-confidence interval for μ .
- Find a margin of error.
- Find the sample size needed to reduce the margin of error by a factor of two.

Solution:

- a) Find the sample mean and the sample standard deviation.

$$\bar{x} = 5.04, s = \sqrt{9.56} = 3.1$$

- b) Find the confidence interval for μ .

$$\left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}}\right) =$$

- c) Find a margin of error.

$$h = t_{\alpha/2} \frac{s}{\sqrt{n}} =$$

- d) Find the sample size needed to reduce the margin of error by a factor of two.

$$n = \left(\frac{st_{\alpha/2}}{h}\right)^2 =$$

□

Example 189. — * **Radioactive-carbon dating** was undertaken on 8 samples from a single early

site.

Sample number	Radiocarbon age determination
C-288	2419
M-26	2485
M-195	2575
M-911	2521
M-912	2451
Y-1279	2550
Y-1280	2540

Compute the two-sided and one-sided confidence intervals of the age of the site, and calculate the sample size required to estimate the age to within ± 10 years.

Solution:

In order to estimate the age of the site, we estimate the mean of the distribution by the sample mean and write

$$\bar{x} = 2505.86$$

a) Two-sided CI:

Use a T -statistic to find a 95% confidence interval which gives a range of plausible values for the mean age. This is

$$\left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right),$$

and, putting in $n = 7$ and $\bar{x} = 2505.86$, $t_{0.025} = 2.447$, from a t -distribution with 6 degrees of freedom, $s = 56.44$, results in a 95% confidence interval of (2453.5, 2558.3), thereby giving a range of *plausible* values for μ .

b) The 100(1 - α)% **lower** confidence interval is

$$\left(-\infty, \bar{x} + \frac{t_{\alpha}}{\sqrt{n}} s \right) = (-\infty ; 2,547.31)$$

here $t_{0.05} = 1.943$.

c) The 100(1 - α)% **upper** confidence interval is

$$(2,464.4 ; \infty)$$

d) To simplify the sample size calculation, we assume a normal approximation to avoid the dependency between $t_{\alpha/2}$ and the sample size, so that we can use:

$$n \approx \left(\frac{s z_{\alpha/2}}{h} \right)^2 = 122$$

in this case, since $h = 10$ and $z_{\alpha/2} = 1.96$.

Note: $t_{\alpha/2} = 1.98$ for $n = 122$.

□

Example 190. — * **Confidence intervals for Proportion** In an opinion poll prior to the 2016 US Presidential election, out of 688 constituents chosen at random 368 said they would vote for Hillary Clinton (53.5%). The newspapers typically use these data to estimate p , the probability that a constituent selected at random would vote Clinton, but they rarely give any idea of the quality of the estimate.

Calculate a 95% confidence interval for p .

Solution:

First identify the random sample. Constituents questioned are labeled $1, \dots, 688$. Let

$$X_i = \begin{cases} 1, & \text{if } i^{\text{th}} \text{ constituent says "I will vote Clinton",} \\ 0, & \text{otherwise.} \end{cases}$$

Then $X_i \sim \text{Ber}(p)$, and the sample size n is 688. Recall that for a $\text{Ber}(p)$,

$$\mu = E(X_i) = p$$

$$\sigma^2 = V(X_i) = p(1-p)$$

We know that μ , and therefore p , can be estimated by the sample mean $\bar{x} = \frac{368}{688} = 0.535$, and we can use the usual confidence interval for the mean

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{\alpha/2}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

assuming that the variance is known, $\sigma^2 \approx \bar{x}(1-\bar{x})$. This gives **(0.498, 0.572)** as a 95% confidence interval for p , with point estimate 0.535. Margin of error = $(0.572 - 0.498)/2 = \mathbf{0.037}$.

For 99% confidence we would use $z_{0.005} = 2.576$ to replace 1.960, and get a wider interval **(0.486, 0.584)** that is less accurate! Margin of error = $(0.584 - 0.486)/2 = \mathbf{0.049}$. \square

Example 191. At a weigh station, the weights of trailer trucks were observed before crossing a highway bridge.

(a) Suppose observations on 30 trucks yielded a sample mean of 12.5 tons. Assume that the standard deviation of truck weights is known to be 2.5 tons. Determine the two-sided 99% intervals of the mean weight of trailer trucks on the particular highway.

(b) In part (a), how many additional trucks should be observed such that the mean truck weight can be estimated to within ± 1 ton with 99% confidence

Solution: (a)

$$\begin{aligned} CI &= \left(\bar{x} + z_{0.005} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{0.995} \frac{\sigma}{\sqrt{n}} \right) \\ &= \left(12.5 - 2.58 \times \frac{2.5}{\sqrt{30}}, 12.5 + 2.58 \times \frac{2.5}{\sqrt{30}} \right) \\ &= (11.3, 13.7) \end{aligned}$$

(b) Sample size calculation: $n' = \left(\frac{\sigma z_{\alpha/2}}{h} \right)^2 = \left(\frac{2.5 \times 2.58}{1} \right)^2 = 42$

So $42 - 30 = 12$ additional observation of truck weights would be required. \square

Example 192. In a traffic study, the speed of vehicles are measured by laser guns for the purpose of determining the mean vehicle speed on a particular city street. It is known that the posted speed limit is 45 kph. The following results were obtained from ten test vehicles:

45, 39, 50, 41, 47, 42, 44, 48, 48, 44 kph

(a) Determine the 95% two-sided confidence interval of the mean vehicle speed.

Solution: (a)

$$\begin{aligned}\bar{x} &= 44.8 \\ s^2 &= \frac{1}{10-1} [\sum x_i^2 - 10 \times \bar{x}^2] \\ &= 12.178 \\ t_{0.025}(r=9) &= 2.262 \\ CI &= \bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}} \\ &= 44.8 \pm (2.262) \times \frac{3.49}{\sqrt{10}} \\ &= (42.3 ; 47.3)\end{aligned}$$

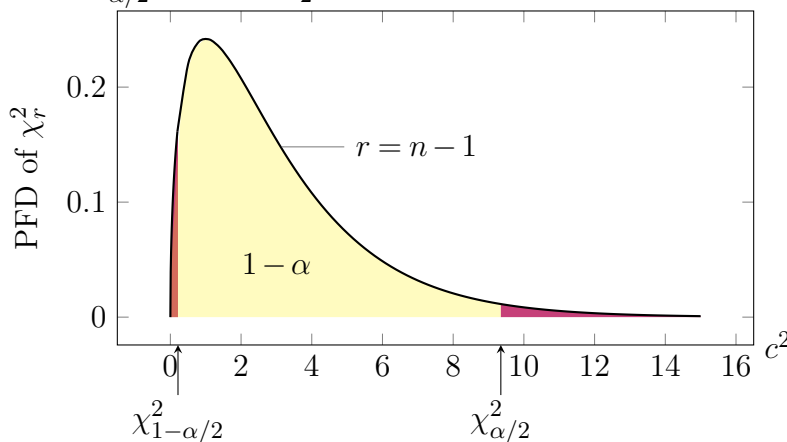
□

6.2.2 Confidence intervals for σ^2

Recall: $C^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$. Therefore,

$$P(\chi_{1-\alpha/2}^2 < C^2 < \chi_{\alpha/2}^2) = 1 - \alpha$$

where $\chi_{\alpha/2}^2$ is the $1 - \frac{\alpha}{2}$ quantile from the distribution χ_{n-1}^2



Proceedings similarly as we did with the mean μ , we find:

The $100(1 - \alpha)\%$ Confidence intervals for σ^2 :

a) The **two-sided** CI:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2} \right)$$

b) The **lower** CI:

$$\left(-\infty, \frac{(n-1)S^2}{\chi_{1-\alpha}^2} \right)$$

c) The **upper** CI:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha}^2}, \infty \right)$$

Example 193. Suppose that $n=9$, $\bar{x}=51.2$, and $s=11.7$. Find the 95% Confidence intervals for σ^2 .

Solution: Here $\alpha=0.05$, $\chi_{1-\alpha/2}^2=2.18$, $\chi_{\alpha/2}^2=17.535$, $\chi_{1-\alpha}^2=2.733$, $\chi_{\alpha}^2=15.507$, which gives

a) The **two-sided** CI:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2}, \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2} \right) = (62.6 ; 503.6)$$

b) The **lower** CI:

$$\left(-\infty, \frac{(n-1)S^2}{\chi_{1-\alpha}^2} \right) = (-\infty ; 401.7)$$

c) The **upper** CI:

$$\left(\frac{(n-1)S^2}{\chi_{\alpha}^2}, \infty \right) = (70.8 ; \infty)$$

□

6.3 Hypothesis Testing

Rather than looking at confidence intervals associated with a model parameter μ , we might formulate a question associated with the data in terms of a hypothesis. In particular, we have a so-called null hypothesis, denoted H_0 , which refers to some basic premise which to we will adhere unless evidence from the data causes us to abandon it.

Basic example Suppose the throughput of a laptop production system, in number of laptops per hour, was $N(\mu_0 = 20, \sigma^2 = 100)$ before an *improvement* performed to the system. We want to see if the improvement was effective by testing the mean after improvement, μ .

The **null hypothesis** is

$$H_0 : \mu = \mu_0 \quad \text{no difference after the improvement}$$

The **alternative hypothesis** H_1 can be either:

- a) $H_1 : \mu \neq \mu_0$ before and after are different (*2-sided*), or
- b) $H_1 : \mu > \mu_0$ new system is better (*1-sided*), or
- c) $H_1 : \mu < \mu_0$ old system is better (*1-sided*).

Example 194. A coin with probability μ of coming up tails is tossed and we hypothesize that it is fair. Therefore:

$$H_0 : \mu = \frac{1}{2} \quad \text{vs} \quad H_1 : \mu \neq \frac{1}{2}$$

Now, if there is reason to believe that the coin is biased towards tails (we suspect that $\mu > \frac{1}{2}$) then:

$$H_0 : \mu = \frac{1}{2} \quad \text{vs} \quad H_1 : \mu > \frac{1}{2}$$

6.3.1 Basic t -test about μ

Recall that:

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

Suppose: $H_0 : \mu = \mu_0$. Then, **under the null hypotheses we have that**

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t_{n-1} \quad \text{if } H_0 \text{ is true.}$$

Key idea: We take a sample to observe a realization of the random variable T :

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

If it falls “far from” its mean $E(T) = 0$ then we reject H_0 , else we **fail to reject** it.

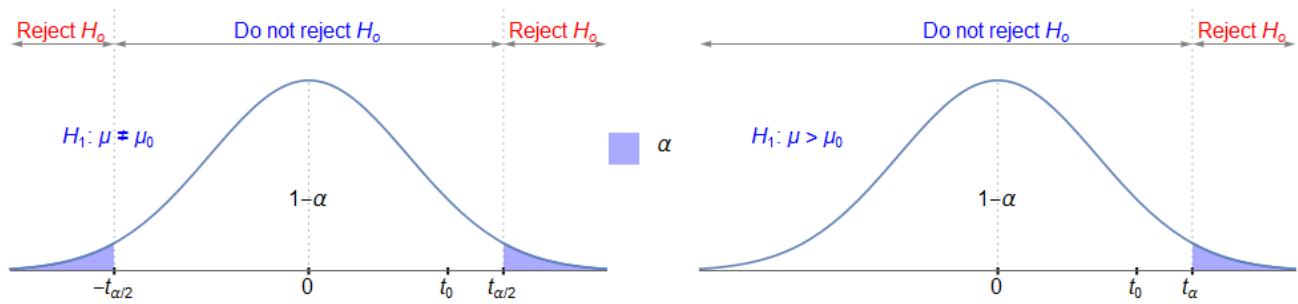
The meaning of “far from” depends on the alternative hypothesis H_1 , and on the **significance level**, α . Suppose we wanted to test

$$H_0 : \mu = \mu_0 \quad \text{against} \quad H_1 : \mu \neq \mu_0$$

In this case, we fail to reject H_0 if: $-t_{\alpha/2} < t_0 < t_{\alpha/2}$ Why? because

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1 - \alpha$$

means that if H_0 is true then $100(1 - \alpha)\%$ of the realizations of T should fall in that range.



Summary of rejection regions:

- a) if $H_1: \mu \neq \mu_0 \Rightarrow$ **reject** H_0 if: $|t_0| > t_{\alpha/2}$
 b) if $H_1: \mu > \mu_0 \Rightarrow$ **reject** H_0 if: $t_0 > t_\alpha$
 c) if $H_1: \mu < \mu_0 \Rightarrow$ **reject** H_0 if: $t_0 < -t_\alpha$



Not rejecting the hypothesis does not mean that there is strong evidence that

H_0 is true. It is recommendable to use the terminology “reject hypothesis H_0 ” or “not reject hypothesis H_0 ” but not to say “accept H_0 ”.

Example 195. — * In a traffic study, the speed of vehicles are measured by laser guns for the purpose of determining the mean vehicle speed on a particular city street. It is known that the posted speed limit is 45 kph. The following results were obtained from ten test vehicles:

45 39 55 50 47 45 44 48 51 44

- (a) Determine the 95% two-sided confidence interval of the mean vehicle speed.
 (b) Test the hypothesis that the vehicles are speeding at a 5% level of significance.
 (c) If we wish to determine the mean vehicle speed to within ± 1 kph with a 99% confidence, what should be the sample size of our observations?

Solution: (a)

$$\begin{aligned}\bar{x} &= 46.8 \\ \sum x_i^2 &= 22,082 \\ s^2 &= \frac{1}{10-1} [\sum x_i^2 - 10 \times \bar{x}^2] = 20 \\ t_{0.025}(r=9) &= 2.262 \\ CI &= \bar{x} \pm t_{0.025} \times \frac{s}{\sqrt{n}} \\ &= 46.8 \pm (2.262) \times \frac{4.47}{\sqrt{10}} \\ &= (43.6 ; 50)\end{aligned}$$

(b) $H_0 : \mu = 45, H_1 : \mu > 45$

$$t_{0.05}(r=9) = 1.833$$

$$t_0 = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{46.8 - 45}{4.47/\sqrt{10}} = 1.27$$

Since $1.27 < 1.833$ we fail to reject H_0 : “There is no evidence to suggest that vehicles are speeding”. Note, we can’t say vehicles are not speeding.

(c) Sample size calculation: Here

$$h = 1$$

$$z_{0.995} = 2.58$$

$$\text{so, } n \approx \left(\frac{sz_{\alpha/2}}{h}\right)^2 = \left(\frac{4.47 \times 2.58}{1}\right)^2 = 133$$

The sample size should be 133. □

Possible error in hypothesis testing. There are two types of possible error in hypothesis testing:

Type I error: rejecting the null hypothesis when it is, in fact, true.

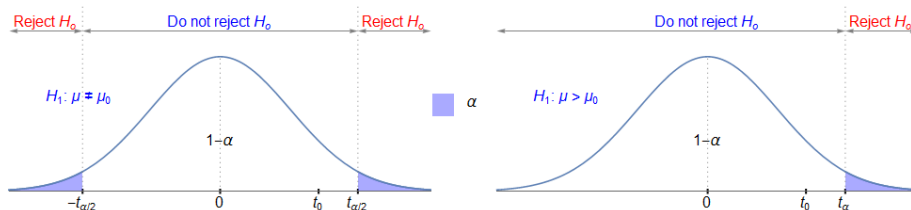
Type II error: not rejecting the null hypothesis when it is, in fact, false.

	H_0 not rejected	H_0 rejected
H_0 true	no error	Type I error
H_0 false	Type II error	no error

Thus,

$$P(\text{Type I error}) = P(\text{reject } H_0 \mid H_0) = \alpha$$

$$P(\text{Type II error}) = P(\text{accept } H_0 \mid H_1) = \beta.$$



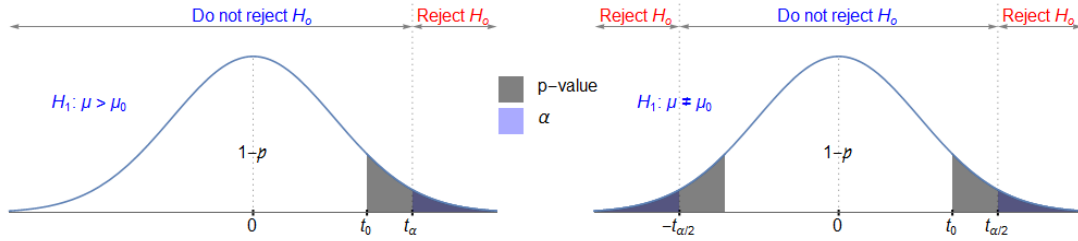
The p -value, denoted p :

1. probability of a test statistic, say T , taking a value **at least as extreme as** its observed value t_0 , assuming H_0 is true,
2. or equivalently: the smallest level α such that we would reject the null-hypothesis with the observed data.

It depends on the alternative hypothesis:

- if $H_1 : \mu > \mu_0 \Rightarrow p = P(T > t_0 \mid H_0)$

- if $H_1 : \mu < \mu_0 \Rightarrow p = P(T < t_0 | H_0)$
- if $H_1 : \mu \neq \mu_0 \Rightarrow p = 2 \min\{P(T < t_0 | H_0), P(T > t_0 | H_0)\}$



Three ways of testing hypotheses. For a null hypothesis $\mu = \mu_0$, the following are equivalent procedures to **reject** H_0 :

1. the T-test t_0 falls in the rejection region
2. the p-value $< \alpha$
3. the $(1 - \alpha)100\%$ confidence interval does not contain t_0 :

$$|t_0| > t \Leftrightarrow t_0 \notin (-t_{\alpha/2}, t_{\alpha/2}) \Leftrightarrow \mu_0 = \bar{x} + t_0 \frac{s}{\sqrt{n}} \notin \left(\bar{x} - t_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2} \frac{s}{\sqrt{n}} \right).$$

Hence we can see that there is an equivalence between the test and the interval.

6.3.2 The magic 5% significance level (or p-value of 0.05)

Question: *what is the critical level for the p-value? Is there some generally accepted level at which null hypotheses are automatically rejected?*

A significance level of $p < 0.05$ is often taken to be of interest, because it is below the “magic” level of 0.05. The p-value of 0.05 is the watershed used by the Food and Drugs Administration which licences new drugs from pharmaceutical companies. As a result it has been almost universally accepted right across the board in all walks of life.

However this level can be, to say the least, inappropriate and possibly even catastrophic. Suppose, for example, we were considering test data for safety critical software for a nuclear power station, N representing the number of faults detected in the first 10 years. Would we be happy with a p-value on trials which suggests that

$$P(N \geq 1) = 0.05?$$

We might be more comfortable if $p = 0.0001$, but even then, given the number of power stations (over 1000 in Europe alone) we would be justified in worrying. The significance level which should be used in deciding whether or not to reject a null hypothesis ought to depend entirely on the question being asked; it quite properly should depend upon the consequences of being wrong. At the very least we should qualify our rejection with something like the following.

$0.05 < p\text{-value} \leq 0.06$	“Weak evidence for rejection”
$0.03 < p\text{-value} \leq 0.05$	“Reasonable evidence for rejection”
$0.01 < p\text{-value} \leq 0.03$	“Good evidence for rejection”
$0.005 < p\text{-value} \leq 0.01$	“Strong evidence for rejection”
$0.001 < p\text{-value} \leq 0.005$	“Very strong evidence for rejection”
$0.0005 < p\text{-value} \leq 0.001$	“Extremely strong evidence for rejection”
$p\text{-value} \leq 0.0005$	“Overwhelming evidence for rejection”

Example 196. — * Concrete placed on a structure was subsequently cored after 28 days, and the following results were obtained of the compressive strengths from five test specimens:

4042, 3505, 3402, 3939, 3472 psi

- (a) Determine the 90% two-sided confidence interval of the mean concrete strength.
 (b) Suppose the confidence interval established in part (a) is too wide, and the engineer would like to have a confidence interval to be ± 300 psi of the computed sample mean concrete strength. Generally, more specimens of concrete would be needed to keep the same confidence level. However, without additional samples, what is the confidence level associated with the specified interval based on the five measurements given above?
 (c) If the required minimum compressive strength is 3500 psi, test whether the concrete satisfies these requirements by performing a one-sided hypothesis test at the 2% significance level.

Solution: (a) Sample mean:

$$\begin{aligned}\bar{x} &= \frac{4042 + 3505 + 3402 + 3939 + 3472}{5} \\ &= 3672\end{aligned}$$

Sample standard deviation:

$$\begin{aligned}s &= \sqrt{\frac{\sum_{i=1}^5 (x_i - \bar{x})^2}{5 - 1}} \\ &= 295.37\end{aligned}$$

For a 90% two-sided confidence interval:

$$\begin{aligned}t_{0.05,4} &= -2.1318 \\ t_{0.95,4} &= 2.1318 \\ CI &= \left(3672 - 2.1318 \frac{295.37}{\sqrt{5}}, 3672 + 2.1318 \frac{295.37}{\sqrt{5}}\right) \\ &= (3672 - 281.6, 3672 + 281.6) \\ &= (3390.4, 3953.6)\end{aligned}$$

(b) If the half-width of the confidence interval is 300, it means:

$$\begin{aligned}t_{\alpha/2} \frac{295.37}{\sqrt{5}} &= 300 \\ t_{\alpha/2} &= 2.2711\end{aligned}$$

Refer to T-table with 4 deg. of freedom; we have:

$$\begin{aligned} t_{0.05} &= 2.1318 \\ t_{0.025} &= 2.7764 \end{aligned}$$

We may use linear interpolation to get an approximate answer:

$$\begin{aligned} 1 - \alpha/2 &= 0.95 + (0.975 - 0.95) * \frac{2.2711 - 2.1318}{2.7764 - 2.1318} \\ &= 0.9554 \\ \Rightarrow \alpha &= 0.911 \end{aligned}$$

The required confidence level is 91.1%.

(c) If the required minimum compressive strength is 3500 psi, test whether the concrete satisfies these requirements by performing a one-sided hypothesis test at the 2% significance level.

$$\begin{aligned} H_0 &: \mu = 3500 \\ H_1 &: \mu < 3500 \end{aligned}$$

In this case, the test statistic will now be:

$$t_0 = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{3672 - 3500}{295.37/\sqrt{5}} = 1.302$$

With $f = 5 - 1 = 4$ d.o.f, we obtain the critical value of t at the 2% significance level to be $t_\alpha \approx -3$. Therefore the value of the test statistic is outside of the region of rejection, hence the null hypothesis cannot be rejected, and therefore we conclude that the concretes meet the minimum requirement. \square

Example 197. — Shoshoni bead rectangles The table below gives width-to-length ratios for 20 rectangles, analyzed as part of a study in experimental aesthetics.

Table 6.1: Shoshoni bead rectangles

Width-to-length ratios			
0.693	0.670	0.654	0.749
0.606	0.553	0.601	0.609
0.672	0.662	0.606	0.615
0.844	0.570	0.933	0.576
0.668	0.628	0.690	0.611

We want to test whether the Shoshoni instinctively made their rectangles conform to the golden ratio. That is we want to test

$$H_0 : \mu = 0.618 \quad \text{against} \quad H_1 : \mu \neq 0.618.$$

We have 20 measurements so, under the null hypothesis $\mu = 0.618$ gives

$$T = \frac{\sqrt{20}(\bar{X} - 0.618)}{S} \sim t(19),$$

where $S^2 = \frac{1}{20} \sum_{i=1}^{20} (X_i - \bar{X})^2$. For these data, $\bar{x} = 0.660$, $s = 0.093$, and the observed value of T is

$$t_0 = \frac{\sqrt{20}(\bar{x} - 0.618)}{s} = \frac{\sqrt{20}(0.660 - 0.618)}{0.093} = 2.019$$

With $r = 20 - 1 = 19$ d.o.f, we obtain the critical value of t at the 5% significance level to be $t_{\alpha/2} \approx 2.086$. Therefore the value of the test statistic is too close to call. In fact, the p -value is very close to α :

$$p\text{-value} = 2 \min\{P(T < 2.019 | H_0), P(T > 2.019 | H_0)\} = 0.058$$

This says that the case for it is very weak.

6.3.3 Paired t -test

Suppose that we have pairs of random variables (X_i, Y_i) and that $D_i = X_i - Y_i$, $i = 1, \dots, n$, is a random sample from a normal distribution, i.e. $D \sim N(\mu, \sigma^2)$ with unknown parameters. We use the test statistic

$$T = \frac{\bar{D} - \mu_0}{S_D / \sqrt{n}} \sim t_{n-1}$$

under the null hypothesis $H_0 : \mu = \mu_0$. Here S_D^2 is the sample variance of the differences D_i .

Example 198. — **Patients with glaucoma in one eye** Here is we ask “Is there a difference in corneal thickness between the eyes?”

Corneal thickness		
Glaucoma	Normal	Difference
488	484	4
478	478	0
480	492	-12
426	444	-18
440	436	4
410	398	12
458	464	-6
460	476	-16

Formally we are testing the difference μ between the corneal thicknesses.

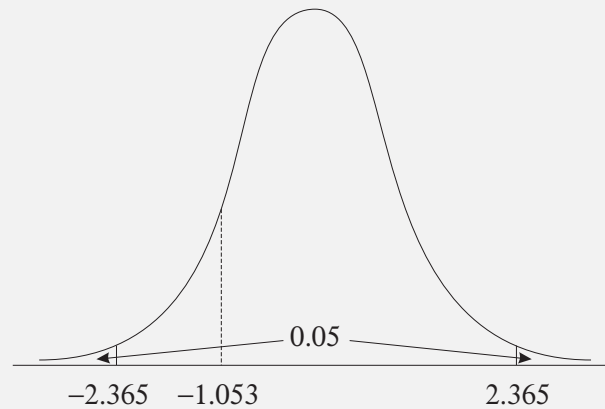
$$H_0 : \mu = 0 \quad \text{against} \quad H_1 : \mu \neq 0.$$

The mean difference is $\bar{d} = -4$ and the estimated standard deviation is $s_D = 10.744$. Under H_0

we obtain a t -statistic of

$$t_0 = \frac{\bar{d} - \mu_0}{s_D/\sqrt{n}} = \frac{-4\sqrt{8}}{10.744} = -1.053.$$

With $r = 8 - 1 = 7$ d.o.f, we obtain the critical value of t at the 5% significance level to be $t_{\alpha/2} \approx 2.365$. Therefore the value of the test statistic is outside of the region of rejection, hence the null hypothesis cannot be rejected, and therefore we cannot reject the null hypothesis of no difference in corneal thickness.



6.3.4 The two-sample t -test

One particularly important use of a t -statistic occurs when we have two samples and we wish to compare the means under the assumption of each sample having the same unknown variance. Consider two random samples X_1, \dots, X_m and Y_1, \dots, Y_n which are independent, normally distributed **with the same variance**. The null hypothesis is

$$H_0 : \mu_X = \mu_Y$$

Fact 6.8 Under H_0 the test statistic T is such that

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t_{m+n-2}, \quad (6.2)$$

where $S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$ is known as the “pooled” variance.

Proof.

Step 1: Under H_0 ,

$$\bar{X} - \bar{Y} \sim N\left(0, \sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)\right), \text{ where } \sigma^2 \text{ is the common variance.}$$

Step 2: recall that

$$\frac{(m-1)S_X^2}{\sigma^2} \sim \chi_{m-1}^2, \quad \frac{(n-1)S_Y^2}{\sigma^2} \sim \chi_{n-1}^2$$

$$\implies \frac{(m-1)S_X^2 + (n-1)S_Y^2}{\sigma^2} \sim \chi_{m+n-2}^2$$

Step 3: Thus, writing

$$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2},$$

we obtain the result:

$$T = \frac{\bar{X} - \bar{Y}}{S_p \sqrt{\left(\frac{1}{m} + \frac{1}{n}\right)}} \sim t_{m+n-2}$$

■

Example 199. — Etruscan and Italian skull widths

Table 4.3 Ancient Etruscan and modern Italian skull widths

Ancient Etruscan skulls							Modern Italian skulls					
141	147	126	140	141	150	142	133	124	129	139	144	140
148	148	140	146	149	132	137	138	132	125	132	137	130
132	144	144	142	148	142	134	130	132	136	130	140	137
138	150	142	137	135	142	144	138	125	131	132	136	134
154	149	141	148	148	143	146	134	139	132	128	135	130
142	145	140	154	152	153	147	127	127	127	139	126	148
150	149	145	137	143	149	140	128	133	129	135	139	135
146	158	135	139	144	146	142	138	136	132	133	131	138
155	143	147	143	141	149	140	136	121	116	128	133	135
158	141	146	140	143	138	137	131	131	134	130	138	138
150	144	141	131	147	142	152	126	125	125	130	133	
140	144	136	143	146	149	145	120	130	128	143	137	

The width measurements are taken with the aim of comparing modern day Italians with ancient Etruscans. The null hypothesis is therefore that the mean skull width is the same. In what follows X refers to Ancient Etruscan measurements and Y refers to Modern Italian.

$$\bar{x} - \bar{y} = 11.33, \quad m = 84, \quad n = 70.$$

Using the formulas above, the value of the test statistic turns out to be $T = 11.92 \gg t_{\alpha/2} \approx 1.98$. (As we are just asking “*is there a difference?*”, we need a 2-sided alternative hypothesis.) The test provides overwhelming evidence to suggest that the two populations are ancestrally of different origin.

6.3.5 Pearson's χ^2 test (goodness-of-fit test)

This is a method for testing how well a particular distribution F_X fits the histogram of a single random variable X from a sample of size n . **Hypothesis H_0 :** We claim that

$$P(\text{“Experiment falls in bin } i\text{”}) = P(X \in \text{bin } i), \quad i = 1, \dots, k.$$

where $P(X \in \text{bin } i)$ is calculated using F_X . Let O_i be the observed number in bin i , $i = 1, 2, \dots, k$ and E_i be the expected number in bin i :

$$E_i = n \cdot P(X \in \text{bin } i) \quad (6.3)$$

Pearson's statistic is

$$Q = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \sim \chi_{k-1-n_p}^2. \quad (6.4)$$

where n_p is the number of parameters to be estimated, if any.

How large Q should be to reject the hypothesis? Reject H_0 if $Q > \chi_{\alpha}^2(k-1-n_p)$. Further, in order to use the test, as a rule of thumb one should check that $nE_i > 5$ for all i .

Chi-Sqr probability tables.

Example 200. — Flying bomb hits on London In the south of London during World War II, 535 flying bomb hits were recorded. The total area was divided into 576 small areas of $\frac{1}{4} \text{km}^2$ each.

Flying bomb hits on London							
Number of hits in an area	0	1	2	3	4	5	≥ 6
Frequency	209	193	90	35	7	1	0

Propaganda broadcasts claimed that the weapon could be aimed accurately. If, however, this was not the case, the hits should be uniformly distributed over the entire area and a natural approximation for the number of hits in a small area would be the Poisson distribution. Is this the case?

Solution:

The first thing to do is estimate the the Poisson parameter. Since we know that $E(X) = \theta$ a good candidate is:

$$\hat{\theta} = \bar{x} = \frac{535}{576} = 0.929$$

Using the Poisson probability mass function $p_X(x) = \frac{e^{-\theta}\theta^x}{x!}$ with $\theta = 0.929$ we therefore obtain

i	0	1	2	3	≥ 4
$P(X \in \text{bin } i)$	0.3949	0.3669	0.1704	0.0528	0.015

Then we pool small cells and we obtain

Number of hits in an area	0	1	2	3	≥ 4
Frequency O_i	209	193	90	35	8
Expected frequency E_i	211.3	196.3	91.2	28.2	8

$$Q = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = 1.73$$

This is tested against $\chi^2(r = 5 - 1 - 1) = 7.814$. Clearly there is not a shred of evidence in favor of rejection. We have therefore found no evidence to reject the hypothesis that the Poisson distribution is a good model. Therefore, there is no evidence that V1 flying bomb could be aimed with any degree of precision. \square

Example 201. — **Fair dice?** In 1882, R. Wolf rolled a dice $n = 20\,000$ times and recorded the number of eyes shown

Number of eyes i	1	2	3	4	5	6
Frequency O_i	3407	3631	3176	2916	3448	3422

Was his dice fair?

Solution: Sample space $\mathcal{S} = \{1, \dots, 6\}$ and let random variable X be the number shown. Since the dice is assumed fair, all results are equally probable hence $P(X \in \text{bin } i) = P(X = i) = 1/6$. In our example $k = 6$. For Wolf's data Q is

$$\begin{aligned} E_i &= 20,000 * 1/6 = 3,333 \\ Q &= \sum_i [(O_i - E_i)^2 / E_i] \\ &= [(3,407 - 3,333)^2 / 3,333] + \dots + [(3,422 - 3,333)^2 / 3,333] \\ &= 1.6280 + 26.5816 + 7.4261 + 52.2501 + 3.9445 + 2.3585 = 94.2 \end{aligned}$$

Since $r = k - 1 = 5$ and the quantile $\chi_{0.05}^2(r) = 11.1$, we have $Q > \chi_{0.05}^2(5)$ which leads to **rejection of the hypothesis of a fair dice.** \square

Example 202. A company claims that 30% of its workforce have PhD's, 60% have an MS degree, and 10% have a BS degree. Suppose a random sample of 100 workers has 50 PhD's, 45 MS's, and 5 BS's. Is this consistent with the company's claim? Use a 0.05 level of significance.

Solution: Null hypothesis: The proportion of PhD's ($i = 1$), MS's ($i = 2$), and BS's ($i = 3$) is:

$$p_1 = 0.3, \quad p_2 = 0.6, \quad p_3 = 0.1$$

Alternative hypothesis: At least one of the proportions in the null hypothesis is false.

$$\begin{aligned} E_1 &= 100 * 0.30 = 30(\text{why?}) \\ E_2 &= 100 * 0.60 = 60 \\ E_3 &= 100 * 0.10 = 10 \\ Q &= \sum_i [(O_i - E_i)^2 / E_i] \\ &= [(50 - 30)^2 / 30] + [(45 - 60)^2 / 60] + [(5 - 10)^2 / 10] \\ &= 13.33 + 3.75 + 2.50 = 19.58 \end{aligned}$$

Here $r = k - 1 = 3 - 1 = 2$ and the quantile $\chi_{0.05}^2(r) = 5.991$. We have $Q > \chi_{0.05}^2(r)$ which leads to **rejection of the null hypothesis.** \square

7. Linear regression

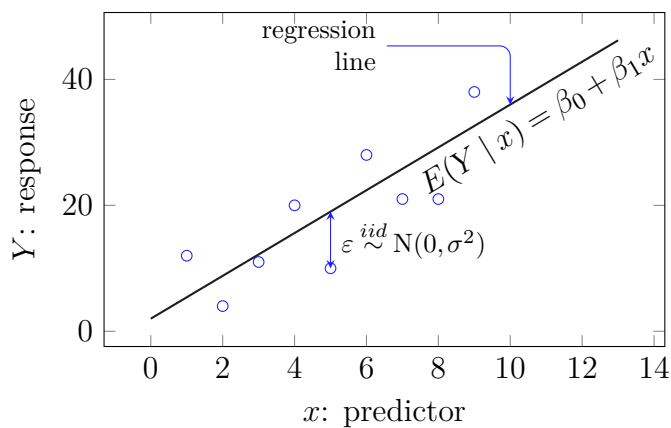
It was the pioneering work of Sir Francis Galton in the 1880s that gave rise to the technique, the original idea being the direct result of an experiment on sweet peas. He noticed that the seeds of the progeny of parents with seeds heavier than average were also heavier than average, but the difference was not as pronounced; the same effect was true for the seeds of the progeny of parents with light seeds, where again the differences from the average were not as great. He called this phenomenon *reversion* and wrote that the mean weight "reverted, or regressed, toward mediocrity".

7.1 The regression model

A **simple** linear regression model takes the form

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

(7.1)



For a sample $\{(X_1, Y_1), (X_2, Y_2) \dots (X_n, Y_n)\}$ a regression model takes the form

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (7.2)$$

where Y_1, Y_2, \dots, Y_n are observable rv's **conditional** on $X_1 = x_1, X_2 = x_2, \dots, X_n = x_n$ and

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

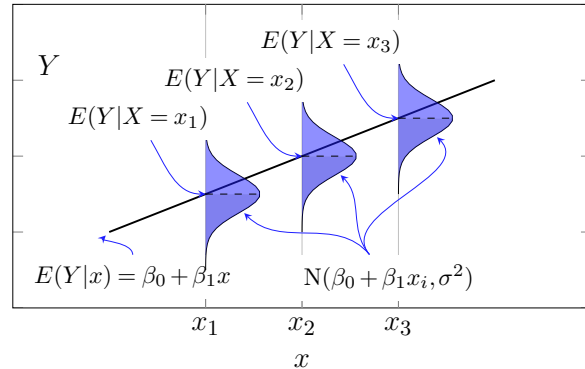
are non-observable random variables.

Terminology :

x_i is called an *explanatory variable* or *independent variable* or *predictor* or *factor*;

Y_i is the *response* or *dependent variable*;

ε_i is the *error* random variable, whose realizations are called *residual*.



Note: x_i is not considered a random variable in linear regression because it is a realization of the random variable X_i , i.e. we take the values of x_i as “given”, and the term Y_i should be interpreted as a conditional:

$$Y_i \leftrightarrow Y_i | X_i = x_i$$

The assumption of the regression model is:

$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

This means:

1. Normality: $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, \dots, n$.
2. Independence of the errors: $\varepsilon_1, \dots, \varepsilon_n$ are independent.
3. Homoscedasticity: $V(\varepsilon_i) = \sigma^2$, with σ^2 constant for all $i = 1, \dots, n$.

Therefore,

$$Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2) \quad (7.3)$$

"Linear" model means that it is *linear in the unknown parameters* $\beta = \beta_0, \beta_1, \beta_2 \dots$, and **not** in x . For example, the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n \quad (7.4)$$

is a linear regression model because it is linear in $\beta_0, \beta_1, \beta_2$.

7.2 Matrix notation

We are going to be concerned with linear model in its more general form involving several explanatory variables. The most convenient way of doing that is to write down the model in matrix notation.

The regression model (7.2) is a set of simultaneous equations which can be written more concisely as

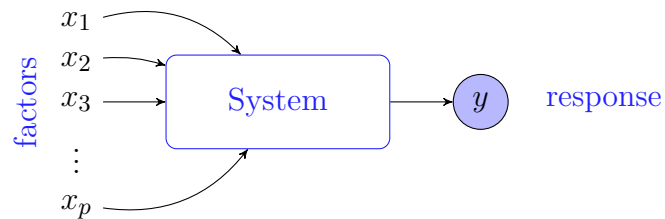
$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (7.5)$$

where

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Multiple regression model.

With p **explanatory variables**, the model takes the form:



$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \quad (7.6)$$

In matrix form it still reads as (7.5) defining: $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ and \mathbf{X} is a $n \times (1+p)$ matrix, called the **design matrix**:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}.$$

Interpretation of β_j . In the model

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3$$

we can see that

$$\frac{\partial Y}{\partial x_j} = \beta_j, \quad j = 1, 2, 3.$$

which means that the parameter β_j represents the marginal change in Y due to a change in x_j .

One should always verify that the sign of β_j accords with intuition.

A quadratic model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n$$

can be written in the form of Equations (7.5) by defining

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}.$$

Main effects and interactions give the following regression function:

$$Y = \beta_0 + \underbrace{\beta_1 x_1 + \beta_2 x_2}_{\text{main effects}} + \underbrace{\beta_3 x_1 x_2}_{\text{interaction term}} + \varepsilon$$

can be written in the form of Equations (7.5) by defining

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix}.$$

Example application: Annual income model Let:

Y is annual income (\$1000/year),

x_1 is educational level (number of years of schooling),

x_2 is number of years of work experience, and

x_3 is gender ($x_3 = 0$ is male, $x_3 = 1$ is female),

Suppose we estimated the following model

$$Y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3 + \varepsilon$$

and obtained (using statistical software),

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 = 0.8 \\ \hat{\beta}_1 = 0.8 \\ \hat{\beta}_2 = 0.5 \\ \hat{\beta}_3 = -3.0 \end{pmatrix} \quad \text{and} \quad \hat{\sigma} = 9.$$

Based on this mean function, we can determine the expected income for any person as long as we know his or her educational level, work experience, and gender.

For example, according to this mean function, a female with 12 years of schooling and 10 years of work experience would expect to earn \$12,400 annually. A male with the same credentials would expect to earn \$15,400 annually.

We can answer questions like: “what is the probability that a female with 16 years education and 28.4 yrs of work experience will earn more than \$40,000/year?”

Recall that

$$Y \sim N(\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_3, \sigma^2)$$

The mean for such a person is $0.8 * 1 + 0.8 * 16 + 0.5 * 28.4 - 3 * 1 = 24.8$, so standardizing

yields the probability:

$$\begin{aligned} P(Y > 40) &= P((Y - 24.8)/9 > (40 - 24.8)/9) \\ &= P(Z > 1.69) \\ &\approx 0.05. \end{aligned}$$

The gender variable x_3 is an **indicator variable**, since it only takes on the values 0/1 (as opposed to x_1 and x_2 which are quantitative).

The slope of an indicator variable (i.e. β_3) is the average gain for observations possessing the characteristic measured by X_3 over observations lacking that characteristic. When the slope is negative, the negative gain is a loss.

7.3 The method of ordinary least squares (OLS)

To estimate the β_j 's we minimize the **sum of squared errors, SSE**:

$$SSE = \sum_{i=1}^n \varepsilon_i^2 = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

over all possible values of the intercept and slopes. To minimize $(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, we differentiate with respect to $\boldsymbol{\beta}$ and equating to $\mathbf{0}$:

$$2\mathbf{X}^T (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{0}, \quad \rightarrow \quad \mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T \mathbf{Y}.$$

which is a set of linear simultaneous equations called the *normal equations* for the linear model.

Fact 7.8 — OLS estimators. Provided $\mathbf{X}^T \mathbf{X}$ is non-singular, the OLS estimators are:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

The matrix C. For convenience, let:

$$\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1} \quad \rightarrow \quad \hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{X}^T \mathbf{Y}.$$

Fact 7.9 — Properties of the OLS estimators. The OLS estimators have useful properties:

- $\hat{\boldsymbol{\beta}}$ is unbiased: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
- $\hat{\boldsymbol{\beta}}$ is a linear transformation of \mathbf{Y} , so it has the (multivariate) normal distribution

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma})$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of $\hat{\boldsymbol{\beta}}$, and

$$\boldsymbol{\Sigma} = \sigma^2 \mathbf{C}.$$

This result implies that

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii})$$

(7.7)

where c_{ij} is the (i, j) element of \mathbf{C} , so

$$V(\hat{\beta}_i) = \sigma^2 c_{ii}$$

$$\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$$

But the true variance σ^2 is unknown, and therefore has to be estimated.

Fact 7.10 — **An unbiased estimator for σ^2 .**

$$\hat{\sigma}^2 = \frac{SSE}{n-p-1} \quad \text{is unbiased for } \sigma^2.$$

Furthermore,

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p-1}^2.$$

Finally, $\hat{\boldsymbol{\beta}}$ is independent of $\hat{\sigma}^2$.

We conclude that the estimator of the covariance matrix is $\hat{\boldsymbol{\Sigma}} = \hat{\sigma}^2 \mathbf{C}$.

The standard errors of the coefficient estimates $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p\}$ are

$$\sqrt{V(\hat{\beta}_i)} = \hat{\sigma} \sqrt{c_{ii}}$$

7.4 Testing the significance of coefficients

Since

$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}) \quad \text{then} \quad Z_i = \frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{c_{ii}}} \sim N(0, 1).$$

Using the result that

$$\frac{(n-p-1)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p-1}^2$$

and remembering the definition of a t -distribution we conclude that

$$T = \frac{\hat{\beta}_i - \beta_i}{\sqrt{V(\hat{\beta}_i)}} \sim t_{n-p-1} \tag{7.8}$$

This enables us to carry out hypothesis tests or calculate confidence intervals for coefficients.

Significance test for β_i : $H_0 : \beta_i = 0$ against $H_1 : \beta_i \neq 0$

Let $t_0 = \hat{\beta}_i / \sqrt{V(\hat{\beta}_i)}$, then we reject H_0 if

$$|t_0| > t_{\alpha/2}$$

(7.9)

where $t_{\alpha/2}$ is the critical values from the distribution t_{n-p-1} , which is typically ≈ 2 . This test is important because **if we cannot reject H_0 it means that the variable x_i does not help explain Y** and therefore should be removed from model.

Recall that instead of this T-test we can also use the p -value, if available, and reject H_0 if $p < \alpha$.

7.5 Goodness-of-fit: R^2

The sum of squared residuals (SSE) measures the amount of variability that the linear model can not explain. The total sum of squares (SST) measures the total amount of variation in Y without considering variable x :

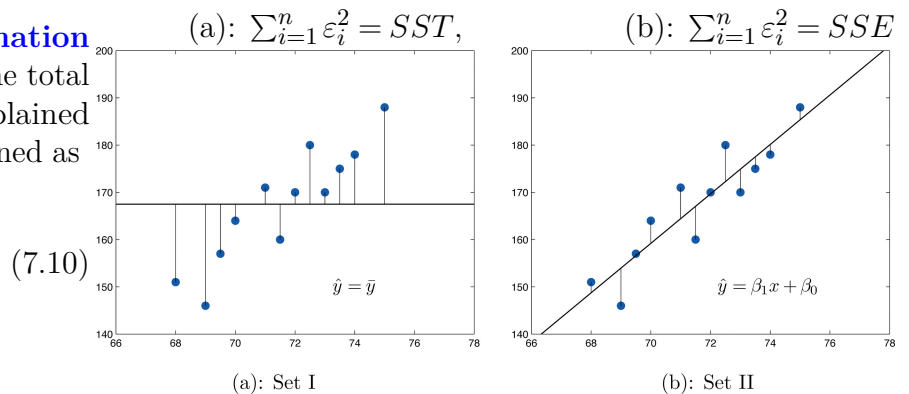
$$SST = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

Hence $\frac{SSE}{SST}$ measures the proportion of total variation that can not be explained by the linear regression. It can be shown that $0 \leq \frac{SSE}{SST} \leq 1$.

The Coefficient of determination

R^2 measures the proportion of the total variations in Y that can be explained by the linear model and it is defined as

$$R^2 = 1 - \frac{SSE}{SST}$$



It quantifies the reduction in variability of the response variable as a result of the linear relationship with X .

R is called the **sample correlation coefficient** and $\approx \rho_{X,Y}$ as $n \rightarrow \infty$, and therefore measures the strength of the linear relationship.

7.5.1 Adjusted R^2

The problem with R^2 is that it cannot decrease when additional explanatory variables are added to the model, even if they have no significant effect on Y . Since all models with the same dependent variable will have the same SST , and SSE cannot increase with additional variables, R^2 is a nondecreasing function of p . An alternative measure, computed by most econometrics packages, is

the so-called ‘Adjusted R^2 ’ :

$$\bar{R}^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)} = 1 - \frac{\hat{\sigma}^2}{s_Y^2} \quad (7.11)$$

where the numerator and denominator of R^2 are divided by their respective degrees of freedom. For a given dependent variable, the denominator does not change; but the numerator, which is $\hat{\sigma}^2$, may rise or fall as p is increased. An additional regressor uses one more degree of freedom, so $(n - (k + 1))$ declines; and SSE declines as well (or remains unchanged). If SSE declines by a larger percentage than the degrees of freedom, then \bar{R}^2 rises, and vice versa. Adding a number of regressors with little explanatory power will increase R^2 , but will decrease \bar{R}^2 – which may even become negative! \bar{R}^2 does not have the interpretation of a squared correlation coefficient, nor of a “batting average” for the model. But it may be used to compare different models of the same dependent variable.

7.5.2 One-way ANOVA

We can decompose the total variance as follows:

$$SST = SSR + SSE$$

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

where SSR = regression sum of squares, and the *predicted values* or fitted values is $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1,i} + \hat{\beta}_2 x_{2,i} + \dots$. It turns out that under the no hypotheses $H_0 : \beta_1 = \beta_2 = \dots = 0$ the statistic:

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE} \quad (7.12)$$

follows an F_{ν_1, ν_2} distribution, where $\nu_1 = p$ and $\nu_2 = n - p - 1$. As usual, we will reject the no hypothesis if the observed F statistic is greater than the critical value from the [F probability tables](#).

7.6 Assessing the model

The first step in looking at the adequacy of a model is to check the assumptions on which it is based:

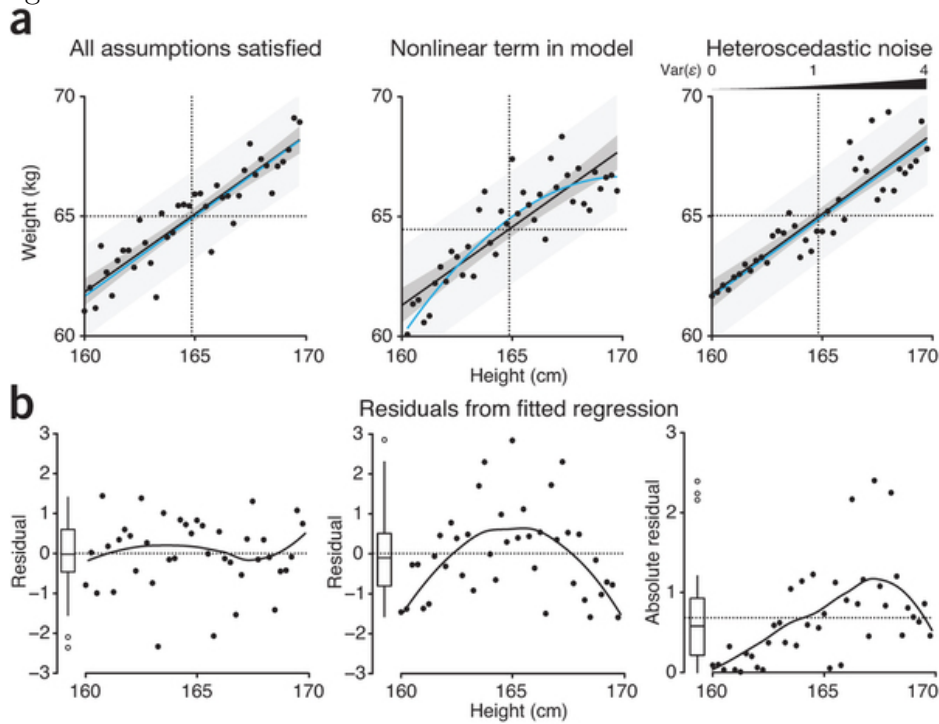
The assumption of the regression model is:

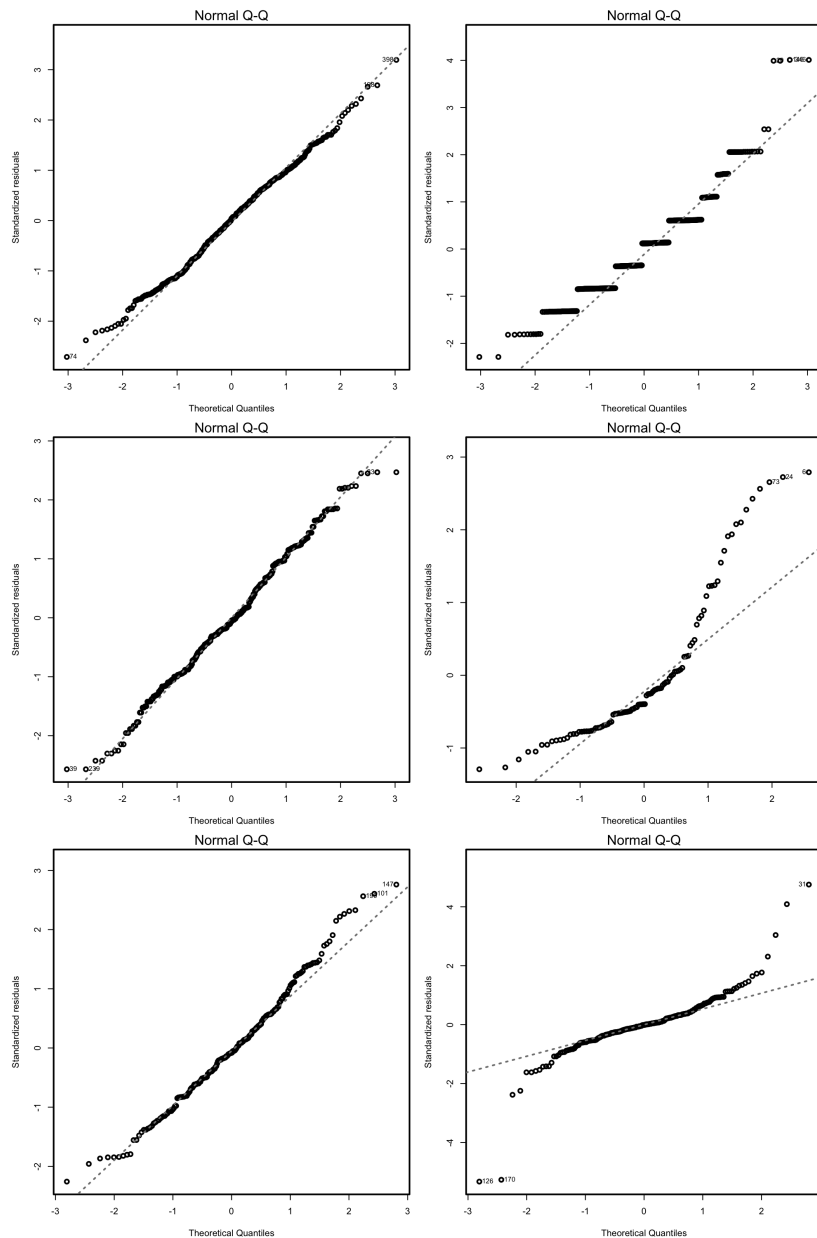
$$\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$$

This means:

1. Normality: $\varepsilon_i \sim N(0, \sigma^2)$ for $i = 1, \dots, n$.
2. Independence of the errors: $\varepsilon_1, \dots, \varepsilon_n$ are independent.
3. Homoscedasticity: $V(\varepsilon_i) = \sigma^2$, with σ^2 constant for $i = 1, \dots, n$.

Residual plots (residuals versus the fitted values) are the main tool for checking model assumptions. **No discernible pattern** in this plot is a good sign of independence, linearity and constant variance. **The Q-Q plot** (normal probability plot) is a scatterplot between the observed quantiles of residuals vs. the theoretical quantiles from the normal distribution. **A linear relationship** indicates good agreement with the normal distribution.



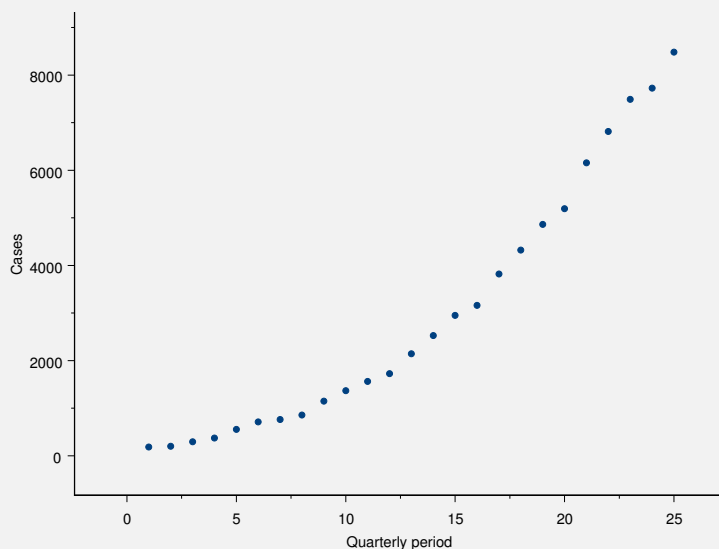


QQ-plots for datasets respecting (left column) and violating (right column) the normality assumption. [Source.](#)

Example 203. — AIDS data for the USA These data are for AIDS incidence in the USA, adjusted for reporting delays. The data are taken from Rosenberg, P.S. and Gail, M.H. (1991): Backcalculation of flexible linear models of the Human Immunodeficiency Virus infection curve. *Applied Statistics*, 40, 269-282.

Newly reported cases are recorded quarterly and the variable *Time* therefore counts 3-monthly periods, starting with the first quarter in 1982.

The scatterplot shows that the trend is not linear.



Incidence of AIDS cases in the USA against time

The plot has all the appearance of showing a functional relationship between AIDS incidence and time. One might, for example, try to fit an exponential function or some kind of power law to model the growth curve.

The plot suggests that the incidence of AIDS against time in the USA is not linearly related to time. Let's fit the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, 2, \dots, n$$

and test the coefficients β_1 , β_2 . We can use a standard computer package to carry out the regression.

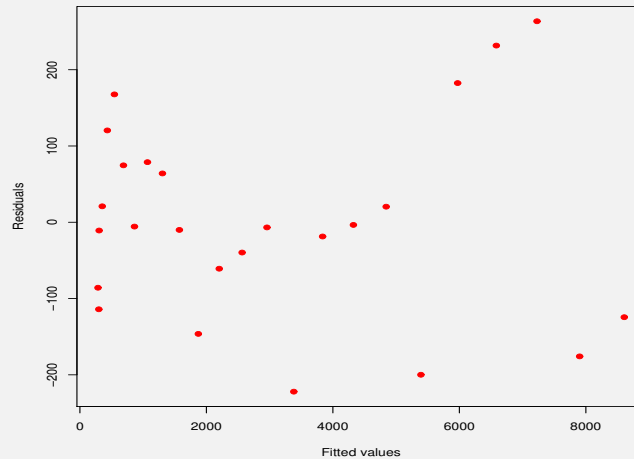
Variable	Coefficient	s.e.	t -value	p -value
Intercept	β_0 343.5913	87.7446	3.9158	0.0007
Time	β_1 -60.1380	15.5514	-3.8671	0.0008
Time ²	β_2 15.6277	0.5806	26.9158	0.0000
<hr/>				
$R^2 = 0.9976$		$d.f. = 22$	$\hat{\sigma} = 134.7$	$SSE = 399155$

This is a typical computer package output, and we need to say a few words about what some of these figures mean.

The values for the coefficients β_0 , β_1 and β_2 are simple enough to interpret. The fitted model is

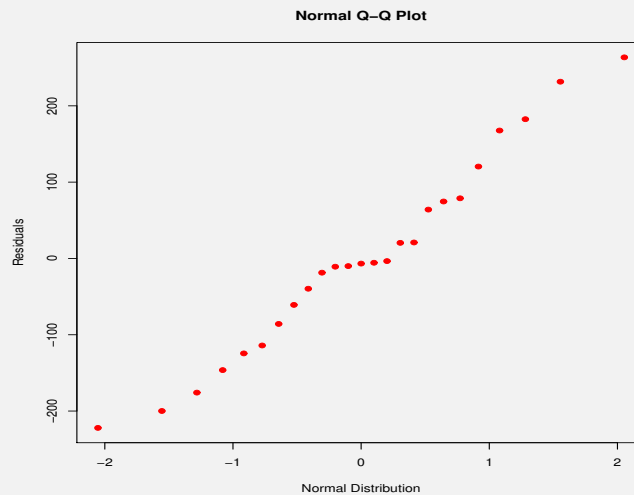
$$Cases = 343.5913 - 60.1380 \times Time + 15.6277 \times Time^2.$$

The value $R^2 = 0.9976$ above means that it is excellent. (99.76% of the variability is explained by the model, whilst the remaining 0.24% is the unexplained or *residual* variability.) But can this model be relied upon?



Plot of residuals against fitted values for AIDS data

- The points to the right seem a little more spread out but there is no indication that spread is a function of fitted value.
- There does seem to be some curvature – re-think the assumption about the model?
- We also need to check up on the third assumption.



Normal probability plot of AIDS residuals

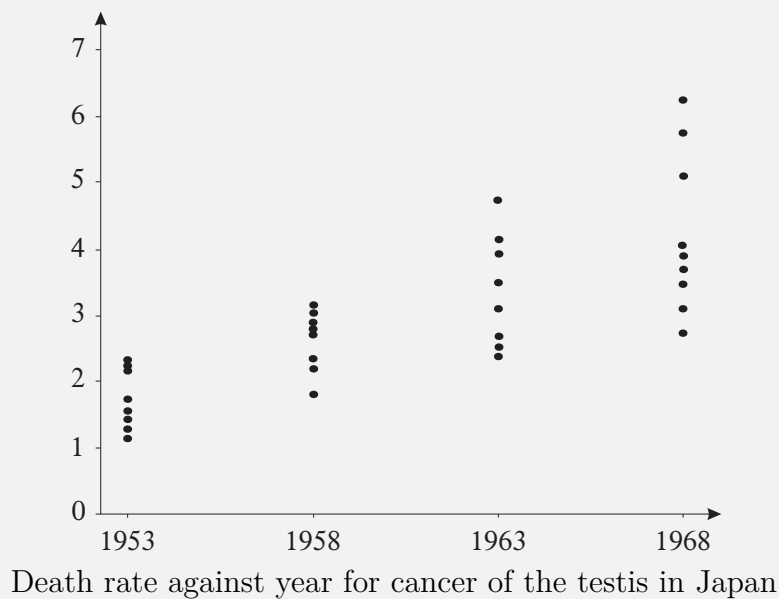
This doesn't look too bad. We can therefore conclude that we have a reasonable model which could perhaps be improved, but which fits the data pretty well for the most part.

Example 204. — **Testicular cancer** The table below comprises data from Lee, Hitosugi and Peterson (1973): Rise in mortality from tumors of the testis in Japan, 1947-70. *J. Nat. Cancer Inst.*, **51**, 1485-90. It gives the populations and numbers of deaths from testicular cancer in 5-year age groups and 5-year periods in Japan. The ages refer to the lowest age in each group and the populations are expressed in millions of persons.

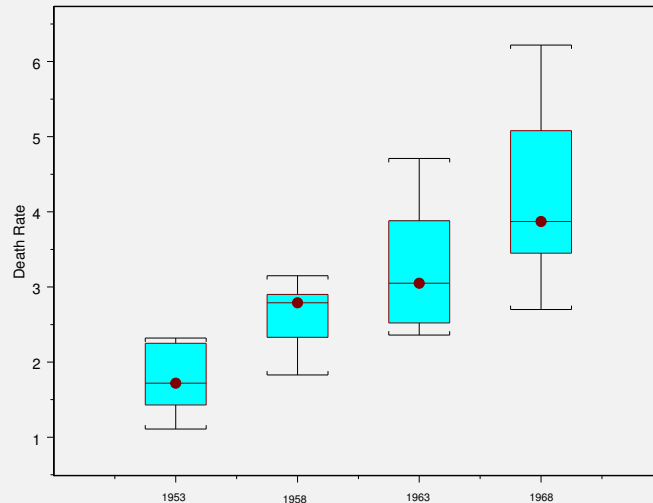
Table 5.4 Deaths in Japan from testicular cancer

Age	1951-55		1956-60		1961-65		1966-70	
	Popn.	Dths	Popn.	Dths	Popn.	Dths	Popn.	Dths
20	20.4	27	21.3	39	22.2	56	24.0	83
25	17.2	40	20.0	58	20.6	97	21.8	125
30	12.6	18	17.1	54	19.9	77	20.8	129
35	11.7	13	12.5	36	17.0	70	19.9	101
40	11.5	26	11.5	32	12.2	29	16.8	67
45	10.3	16	11.2	26	11.1	34	12.0	37
50	9.3	16	9.8	27	10.7	27	10.7	29
55	7.6	17	8.7	19	9.2	32	10.1	39
60	5.9	13	6.8	21	7.9	21	8.4	31

The scatterplot shows that the mean death rate from cancer of the testis in Japan has been rising steadily since 1951.



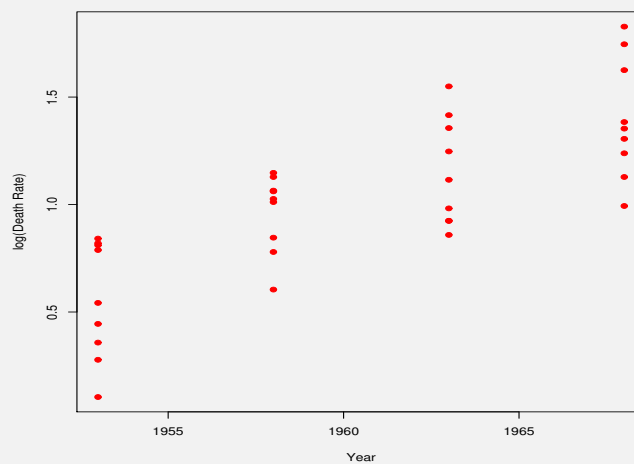
Note that there is no variability in the explanatory variable, but there is marked variability in the response. This is clearly shown in the boxplots.



Boxplots of death rates

It is clear that the variability in the data increases as the year variable increases. What should one do about this?

The answer is to look for a transformation which will stabilise the variance. Here we need a transformation which compresses large values of the response more than it compresses smaller values; something like a square root or a cube root or possibly even a log transformation. Taking the log of the death rate results in the next scatterplot.

Testicular cancer: plot of $\log(\text{Death rate})$ against year

This looks reasonable and we could now go ahead and fit a model of the form

$$\log(Y_i) = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n.$$

With the log transformed data, the fitted model turns out to be as given below.

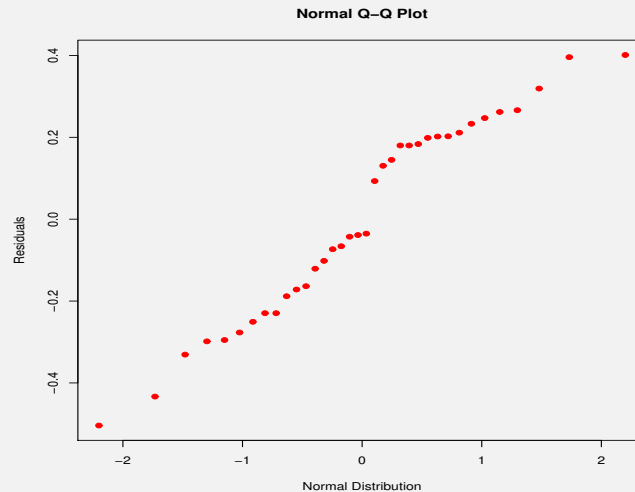
Linear Model

Response: log(Death rate)

Variable	Coefficient	s.e.	<i>t</i> -value	<i>p</i> -value
Intercept	β_0 -105.9198	14.4887	-7.3105	0.0000
Year	β_1 0.0545	0.0074	7.3808	0.0000

$r^2 = 0.6157$ $d.f. = 34$ $s = 0.2479$ $SSE = 2.0891$

Of course stable variance alone is not enough because the residuals also need to be normally distributed. This can be checked with a normal q-q plot of the residuals.



Normal probability plot of testicular cancer residuals

The plot seems to show a very rough straight line, but it is not entirely convincing.

7.7 Model selection

Question: With a large number of potential predictors, how do we choose the predictors to include in the model? Selecting the most adequate predictors for a multiple regression model is a challenging task without a unique solution. We want good prediction, but keeping the model as simple as possible:

Occam's Razor (The law of parsimony): *"Among competing hypotheses, the one with the fewest assumptions should be selected."*

→ choose the model with the fewest number of parameters that explains the data.

The inclusion of more predictors is not for free: there is a price to pay in terms more variability on the coefficients: the maximum number of predictors p that can be considered in a linear model for a sample size n : $p \leq n - 2$. Or equivalently, there is a minimum sample size n required for fitting a model with p predictors: $n \geq p + 2$, or the degrees of freedom $n - p - 1 \geq 1$.

Good fit Criteria

When we have several alternative models, we need a criteria for determining which of two models is "better". As you might suspect, there is no universally agreed upon criteria for evaluating models. A good model would give **yes** to the following questions:

- are the assumptions met?
- good $\text{Adj-}R^2$?
- good t-statistics for the β_j 's ($|t_0| > 2$) ?
- the sign of β_j 's accords with intuition?

Example 205. — * **Housing values in suburbs of Boston.** Home values for 506 Boston suburbs with potential influential factors is shown below. (Source: Belsley D. A., Kuh, E. and Welsch, R. E. (1980) Regression Diagnostics.).

X_i	Description
1	Per capita crime rate by town
2	Proportion of residential land zoned for lots over 25000 square feet
3	Proportion of non-retail business acres per town
4	Charles River dummy variable (1 if tract bounds river, 0 otherwise)
5	Nitrogen oxide concentration (parts per 10 million)
6	Average number of rooms per dwelling
7	Proportion of owner-occupied units built prior to 1940
8	Weighted mean of distances to five Boston employment centers
9	Index of accessibility to radial highways
10	Full-value property-tax rater per \$10000
11	Pupil-teacher ratio by town
13	Lower status of the population (percent)

The models below are the output of an automated model selection procedure, for the response variable $Y =$ **median value of owner-occupied homes in \$1000s**, and for models that only consider main effects:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

The parameter table gives the parameter values in the second row and the T-statistic in the third row in bold.

- Explain how the parameter values should be used to assess the adequacy of a model; give 3 example.
- decide which model is the "best" and carefully explain why.
- decide which model is the "worst" and carefully explain why.

Y = Median value of owner-occupied homes in \$1000s

Model	Adj-R ²	Parameter Table	Residuals	Q-Q plot																		
M-1	0.7	<table border="1"> <tr> <td>β_0</td> <td>β_4</td> <td>β_9</td> <td>β_{11}</td> <td>β_{13}</td> </tr> <tr> <td>4.3</td> <td>0.15</td> <td>-0.0019</td> <td>-0.039</td> <td>-0.04</td> </tr> <tr> <td>46.</td> <td>3.9</td> <td>-1.4</td> <td>-7.4</td> <td>-25.</td> </tr> </table>	β_0	β_4	β_9	β_{11}	β_{13}	4.3	0.15	-0.0019	-0.039	-0.04	46.	3.9	-1.4	-7.4	-25.					
β_0	β_4	β_9	β_{11}	β_{13}																		
4.3	0.15	-0.0019	-0.039	-0.04																		
46.	3.9	-1.4	-7.4	-25.																		
M-2	0.39	<table border="1"> <tr> <td>β_0</td> <td>β_1</td> <td>β_2</td> <td>β_5</td> <td>β_7</td> </tr> <tr> <td>3.6</td> <td>-0.018</td> <td>0.0021</td> <td>-0.8</td> <td>-0.0012</td> </tr> <tr> <td>41.</td> <td>-9.9</td> <td>2.9</td> <td>-4.2</td> <td>-1.6</td> </tr> </table>	β_0	β_1	β_2	β_5	β_7	3.6	-0.018	0.0021	-0.8	-0.0012	41.	-9.9	2.9	-4.2	-1.6					
β_0	β_1	β_2	β_5	β_7																		
3.6	-0.018	0.0021	-0.8	-0.0012																		
41.	-9.9	2.9	-4.2	-1.6																		
M-3	0.36	<table border="1"> <tr> <td>β_0</td> <td>β_2</td> <td>β_8</td> <td>β_{10}</td> </tr> <tr> <td>3.6</td> <td>0.0049</td> <td>-0.026</td> <td>-0.0013</td> </tr> <tr> <td>53.</td> <td>5.9</td> <td>-2.5</td> <td>-13.</td> </tr> </table>	β_0	β_2	β_8	β_{10}	3.6	0.0049	-0.026	-0.0013	53.	5.9	-2.5	-13.								
β_0	β_2	β_8	β_{10}																			
3.6	0.0049	-0.026	-0.0013																			
53.	5.9	-2.5	-13.																			
M-4	0.49	<table border="1"> <tr> <td>β_0</td> <td>β_1</td> <td>β_3</td> <td>β_4</td> <td>β_9</td> <td>β_{11}</td> </tr> <tr> <td>4.3</td> <td>-0.016</td> <td>-0.02</td> <td>0.2</td> <td>0.0033</td> <td>-0.055</td> </tr> <tr> <td>35.</td> <td>-8.4</td> <td>-8.5</td> <td>3.9</td> <td>1.5</td> <td>-7.9</td> </tr> </table>	β_0	β_1	β_3	β_4	β_9	β_{11}	4.3	-0.016	-0.02	0.2	0.0033	-0.055	35.	-8.4	-8.5	3.9	1.5	-7.9		
β_0	β_1	β_3	β_4	β_9	β_{11}																	
4.3	-0.016	-0.02	0.2	0.0033	-0.055																	
35.	-8.4	-8.5	3.9	1.5	-7.9																	
M-5	0.5	<table border="1"> <tr> <td>β_0</td> <td>β_2</td> <td>β_3</td> <td>β_6</td> </tr> <tr> <td>1.4</td> <td>0.00058</td> <td>-0.02</td> <td>0.29</td> </tr> <tr> <td>11.</td> <td>0.88</td> <td>-8.5</td> <td>14.</td> </tr> </table>	β_0	β_2	β_3	β_6	1.4	0.00058	-0.02	0.29	11.	0.88	-8.5	14.								
β_0	β_2	β_3	β_6																			
1.4	0.00058	-0.02	0.29																			
11.	0.88	-8.5	14.																			
M-6	0.72	<table border="1"> <tr> <td>β_0</td> <td>β_1</td> <td>β_3</td> <td>β_9</td> <td>β_{11}</td> <td>β_{13}</td> </tr> <tr> <td>4.3</td> <td>-0.01</td> <td>-0.0018</td> <td>0.0043</td> <td>-0.043</td> <td>-0.037</td> </tr> <tr> <td>49.</td> <td>-6.8</td> <td>-0.92</td> <td>2.6</td> <td>-8.4</td> <td>-21.</td> </tr> </table>	β_0	β_1	β_3	β_9	β_{11}	β_{13}	4.3	-0.01	-0.0018	0.0043	-0.043	-0.037	49.	-6.8	-0.92	2.6	-8.4	-21.		
β_0	β_1	β_3	β_9	β_{11}	β_{13}																	
4.3	-0.01	-0.0018	0.0043	-0.043	-0.037																	
49.	-6.8	-0.92	2.6	-8.4	-21.																	
M-7	0.39	<table border="1"> <tr> <td>β_0</td> <td>β_4</td> <td>β_5</td> <td>β_7</td> <td>β_9</td> <td>β_{10}</td> </tr> <tr> <td>3.9</td> <td>0.28</td> <td>-0.58</td> <td>-0.0024</td> <td>0.0067</td> <td>-0.0012</td> </tr> <tr> <td>47.</td> <td>4.8</td> <td>-2.8</td> <td>-3.2</td> <td>1.7</td> <td>-5.5</td> </tr> </table>	β_0	β_4	β_5	β_7	β_9	β_{10}	3.9	0.28	-0.58	-0.0024	0.0067	-0.0012	47.	4.8	-2.8	-3.2	1.7	-5.5		
β_0	β_4	β_5	β_7	β_9	β_{10}																	
3.9	0.28	-0.58	-0.0024	0.0067	-0.0012																	
47.	4.8	-2.8	-3.2	1.7	-5.5																	
M-8	0.47	<table border="1"> <tr> <td>β_0</td> <td>β_1</td> <td>β_4</td> <td>β_7</td> <td>β_{11}</td> </tr> <tr> <td>4.5</td> <td>-0.016</td> <td>0.2</td> <td>-0.0038</td> <td>-0.061</td> </tr> <tr> <td>38.</td> <td>-9.5</td> <td>3.7</td> <td>-7.4</td> <td>-9.3</td> </tr> </table>	β_0	β_1	β_4	β_7	β_{11}	4.5	-0.016	0.2	-0.0038	-0.061	38.	-9.5	3.7	-7.4	-9.3					
β_0	β_1	β_4	β_7	β_{11}																		
4.5	-0.016	0.2	-0.0038	-0.061																		
38.	-9.5	3.7	-7.4	-9.3																		
M-9	0.7	<table border="1"> <tr> <td>β_0</td> <td>β_3</td> <td>β_6</td> <td>β_8</td> <td>β_9</td> <td>β_{13}</td> </tr> <tr> <td>3.</td> <td>-0.0066</td> <td>0.12</td> <td>-0.031</td> <td>-0.0071</td> <td>-0.035</td> </tr> <tr> <td>20.</td> <td>-2.7</td> <td>6.7</td> <td>-4.5</td> <td>-4.8</td> <td>-16.</td> </tr> </table>	β_0	β_3	β_6	β_8	β_9	β_{13}	3.	-0.0066	0.12	-0.031	-0.0071	-0.035	20.	-2.7	6.7	-4.5	-4.8	-16.		
β_0	β_3	β_6	β_8	β_9	β_{13}																	
3.	-0.0066	0.12	-0.031	-0.0071	-0.035																	
20.	-2.7	6.7	-4.5	-4.8	-16.																	

Example 206. — * Data from 93 cars on sale in the USA in 1993. Data from 93 cars, selected at random, on sale in the US in 1993 with 27 variables. Source: Lock, R. H. (1993) 1993 New Car Data. Journal of Statistics Education 1(1).

X_i	Description
1	Minimum Price (in \$1,000): price for a basic version
2	Price (in \$1,000): average of Min.Price and Max.Price.
3	Maximum Price (in \$1,000): price for 'a premium version'
4	City MPG (miles per US gallon by EPA rating).
5	Highway MPG.
6	Number of cylinders (missing for Mazda RX-7, which has a rotary engine).
7	Engine size (litres).
8	Horsepower (maximum).
9	RPM (revs per minute at maximum horsepower)
10	Engine revolutions per mile (in highest gear).
11	Fuel tank capacity (US gallons).
12	Passenger capacity (persons).
13	Length (inches).
14	Wheelbase (inches).
15	Width (inches).
16	U-turn space (feet).
17	"Rear" seat room (inches) (missing for 2-seater vehicles).
18	Luggage capacity (cubic feet) (missing for vans).
19	Weight (pounds).

The models below are the output of an automated model selection procedure, for the response variable $Y =$ the (log of) price of the basic car, and for models that only consider the **main effects**. The parameter table gives the parameter values in the second row and the T-statistic in the third row. You can ignore the column labeled "BIC".

- Explain how the parameter values should be used to assess the adequacy of a model; give an example.
- decide which model is the "best" and carefully explain why.
- decide which model is the "worst" and carefully explain why.

Y = Log of Minimum Price (in \$1,000): price for a basic version

Model	Adj-R ²	BIC	Parameter Table						Residuals	Q-Q plot
M-1	0.76	19.	β_0	β_5	β_8	β_{13}	β_{15}	β_{16}		
			2.9	-0.025	0.0058	0.014	-0.035	-0.011		
			4.	-3.8	9.	4.6	-2.4	-0.79		
M-2	0.75	19.	β_0	β_8	β_{12}	β_{15}	β_{19}			
			3.2	0.0039	-0.068	-0.04	0.00066			
			4.5	3.9	-1.7	-2.9	5.1			
M-3	0.73	21.	β_0	β_4	β_8					
			2.6	-0.028	0.0055					
			12.	-4.5	8.1					
M-4	0.75	20.	β_0	β_7	β_8	β_{15}	β_{19}			
			3.9	0.071	0.0047	-0.051	0.0005			
			4.3	1.3	6.3	-3.2	4.9			
M-5	0.77	17.	β_0	β_4	β_8	β_{11}	β_{14}	β_{15}		
			2.6	-0.022	0.006	0.01	0.024	-0.043		
			3.4	-2.8	8.5	0.62	3.7	-3.3		
M-6	0.77	15.	β_0	β_8	β_{11}	β_{13}	β_{15}	β_{19}		
			3.	0.0052	0.022	0.01	-0.061	0.00032		
			4.3	7.5	1.3	3.2	-4.2	2.6		
M-7	0.74	21.	β_0	β_5	β_8	β_{13}				
			1.2	-0.018	0.0053	0.007				
			2.4	-2.9	8.1	3.2				
M-8	0.76	19.	β_0	β_4	β_5	β_8	β_{13}	β_{15}		
			3.1	-0.015	-0.011	0.0057	0.012	-0.04		
			4.1	-0.95	-0.75	8.6	3.9	-3.1		
M-9	0.76	21.	β_0	β_5	β_8	β_{14}	β_{15}	β_{19}		
			2.5	-0.011	0.0059	0.02	-0.042	0.0002		
			2.9	-1.3	7.4	2.2	-3.	1.1		

More examples like this here.

Solution:

- a) Explain how the parameter values should be used to assess the adequacy of a model; give an example.

What is important about the parameter value is its sign, in that it has to make intuitive sense. For example in model M-3 the parameter β_8 , which is the parameter for horsepower, is positive. This makes sense because the price of a car should increase with its horsepower.

- b) decide which model is the "best" and carefully explain why.

The residuals and future plots for all models are very similar and not particularly faulty, so

the basic assumptions are met fairly well by all models.

By the principle parsimony, the best model is the one with the fewest number of parameters, all of them significant (good t-test) and a good $\text{Adj-}R^2$. I would choose M-3, because the closest competitor, M-7, has only a slightly better $\text{Adj-}R^2$ but at the high price of one additional variable.

c) decide which model is the “worst” and carefully explain why.

I would choose model M-8 because it has six parameters, two of them not significantly different from zero. Model M-9 is very similar but has slightly better T-statistics.

□

Example 207. — * **Climate change** This data series from 1959 - 2016 includes the annual global mean surface temperature (Temp) and two possible explanatory variables, the year and the annual average fraction of CO₂ contained in the earth’s atmosphere (CO₂). Source

	Description
X_1	Year
X_2	CO ₂ atmospheric composition is defined as the number of molecules of carbon dioxide divided by the number of molecules of dry air multiplied by one million (ppm).
Y	The annual average Temperature is measured in units of 1/100 of a degree centigrade increase above the 1950-1980 mean, often referred to as the global surface temperature anomaly.

The models below are the output of an automated model selection procedure, for the response variable Y . The parameter table gives the factor that the parameter multiplies in the first row (“1.” means intercept), the parameter values in the second row and the T-statistic in the third row. You can ignore the column labeled “BIC”.

In the models below, decide which model is the best and explain why.

Y = Temp

Model	Adj-R ²	BIC	Parameter Table					Residuals	Q-Q plot		
M-1	0.92	-107.	1.	x ₁	x ₂	x ₁ x ₂					
			-158.	0.068	0.92	-0.00043					
			-3.6	3.4	3.9	-3.8					
M-2	0.92	-107.	1.	x ₂	x ₁ ²	x ₁ x ₂					
			-94.	0.94	0.00002	-0.00044					
			-3.7	3.9	3.4	-3.8					
M-3	0.92	-107.	1.	x ₂	x ₂ ²	x ₁ x ₂					
			-33.	0.54	-0.0001	-0.00021					
			-4.1	3.9	-3.2	-3.8					
M-4	0.92	-103.	1.	x ₂	x ₁ ²	x ₂ ²	x ₁ x ₂				
			-118.	1.1	0.00003	0.00005	-0.00053				
			-1.2	1.7	0.89	0.27	-1.4				
M-5	0.92	-103.	1.	x ₁	x ₂	x ₂ ²	x ₁ x ₂				
			-197.	0.09	1.	0.00004	-0.0005				
			-1.1	0.89	1.8	0.22	-1.5				
M-6	0.92	-103.	1.	x ₁	x ₂	x ₁ ²	x ₁ x ₂				
			-390.	0.31	0.86	-0.00007	-0.0004				
			-0.28	0.21	2.	-0.17	-1.9				
M-7	0.92	-103.	1.	x ₁	x ₂	x ₁ ²	x ₂ ²				
			-1374.	1.4	0.16	-0.00038	-0.00013				
			-1.4	1.4	3.1	-1.4	-1.9				
M-8	0.92	-103.	1.	x ₁	x ₁ ²	x ₂ ²	x ₁ x ₂				
			-1594.	1.7	-0.00045	-0.00016	0.00009				
			-1.6	1.7	-1.8	-2.	3.1				
M-9	0.91	-99.	1.	x ₁	x ₂	x ₁ ²	x ₂ ²	x ₁ x ₂			
			3312.	-3.9	3.6	0.0011	0.0005	-0.0019			
			0.35	-0.36	0.52	0.37	0.4	-0.5			

Solution: Solved in class. □

7.8 Making predictions

Once the coefficients have been estimated and the assumptions verified, the fitted equation can be used to obtain predictions for Y for any given values $\mathbf{x}_0 = (1, x_{01}, \dots, x_{0p})$ of the explanatory variables $\mathbf{x} = (1, x_1, \dots, x_p)$. There are two type of predictions that we can do: prediction of the mean response and prediction of a particular realization.

7.8.1 Prediction of the mean response at \mathbf{x}_0 , $\mathbf{E}(Y | \mathbf{x}_0) = \mathbf{x}_0\boldsymbol{\beta}$.

The **point estimate** of $\mathbf{E}(Y | \mathbf{x}_0) = \mathbf{x}_0\boldsymbol{\beta}$ is:

$$\begin{aligned}\hat{Y}_0 &= \mathbf{x}_0\hat{\boldsymbol{\beta}}, \\ &= \hat{\beta}_0 + \hat{\beta}_1x_{01} + \hat{\beta}_2x_{02} + \cdots + \hat{\beta}_px_{0p}\end{aligned}$$

and is an unbiased estimator of $\mathbf{x}_0\boldsymbol{\beta}$. For the **confidence interval** for $\mathbf{E}(Y | \mathbf{x}_0) = \mathbf{x}_0\boldsymbol{\beta}$, we note that \hat{Y}_0 is a linear combination of the random vector $\hat{\boldsymbol{\beta}}$ and therefore must be normally distributed with

$$\mathbf{E}(\hat{Y}_0) = \mathbf{x}_0\boldsymbol{\beta}, \quad \mathbf{V}(\hat{Y}_0) = \mathbf{x}_0\boldsymbol{\Sigma}\mathbf{x}_0^T = \sigma^2\mathbf{x}_0\mathbf{C}\mathbf{x}_0^T.$$

Thus,

$$T = \frac{\mathbf{x}_0\hat{\boldsymbol{\beta}} - \mathbf{x}_0\boldsymbol{\beta}}{\hat{\sigma}\sqrt{\mathbf{x}_0\mathbf{C}\mathbf{x}_0^T}} \sim t_{n-p-1}$$

and an inequality can be constructed and re-arranged for $\mathbf{x}_0\boldsymbol{\beta}$ in the usual way:

$(1 - \alpha)\%$ **Confidence interval for $\mathbf{E}(Y | \mathbf{x}_0)$:**

$$\mathbf{x}_0\hat{\boldsymbol{\beta}} \pm t_{\alpha/2} \hat{\sigma}\sqrt{\mathbf{x}_0\mathbf{C}\mathbf{x}_0^T}$$

7.8.2 Prediction of a particular realization of $Y_0 = \mathbf{x}_0\boldsymbol{\beta} + \varepsilon_0$

The **point estimate** of $Y_0 = \mathbf{x}_0\boldsymbol{\beta} + \varepsilon_0$, with $\varepsilon_0 \sim N(0, \sigma^2)$ is also:

$$\hat{Y}_0 = \mathbf{x}_0\hat{\boldsymbol{\beta}}$$

and is an unbiased estimator of Y_0 . For the **confidence interval**, we know that $Y_0 \sim N(\mathbf{x}_0\boldsymbol{\beta}, \sigma^2)$, and therefore $Y_0 - \hat{Y}_0$ has a normal distribution with mean

$$\mathbf{E}(Y_0 - \hat{Y}_0) = \mathbf{x}_0\boldsymbol{\beta} - \mathbf{x}_0\boldsymbol{\beta} = 0$$

$$\begin{aligned}\mathbf{V}(Y_0 - \hat{Y}_0) &= \mathbf{V}(y_0) + \mathbf{V}(\mathbf{x}_0\hat{\boldsymbol{\beta}}) \\ &= \sigma^2 + \sigma^2\mathbf{x}_0\mathbf{C}\mathbf{x}_0^T \\ &= \sigma^2(1 + \mathbf{x}_0\mathbf{C}\mathbf{x}_0^T).\end{aligned}$$

since Y_0 and $\hat{\boldsymbol{\beta}}$ are independent. Then,

$$T = \frac{Y_0 - \mathbf{x}_0\hat{\boldsymbol{\beta}}}{\hat{\sigma}\sqrt{1 + \mathbf{x}_0\mathbf{C}\mathbf{x}_0^T}} \sim t_{n-p-1}. \quad (7.13)$$

$(1 - \alpha)\%$ **Confidence interval for a particular realization of $Y | \mathbf{x}_0$:**

$$\mathbf{x}_0\hat{\boldsymbol{\beta}} \pm t_{\alpha/2}\hat{\sigma}\sqrt{1 + \mathbf{x}_0\mathbf{C}\mathbf{x}_0^T}$$

7.9 Simple linear regression

In this case:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \mathbf{X}^T \mathbf{Y} = \begin{pmatrix} \sum Y_i \\ \sum x_i Y_i \end{pmatrix}, \quad \mathbf{X}^T \mathbf{X} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}$$

$$\mathbf{C} = \frac{\begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}}{n \sum (x_i - \bar{x})^2}$$

It will be convenient to define :

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(Y_i - \bar{Y}) & \parallel & \quad S_{xx} = \sum (x_i - \bar{x})^2 & \parallel & \quad S_{yy} = \sum (Y_i - \bar{Y})^2 \\ &= \sum x_i Y_i - n \bar{x} \bar{Y} & & \quad = \sum x_i^2 - n \bar{x}^2 & & \quad = \sum Y_i^2 - n \bar{Y}^2 \\ &\approx n \text{Cov}(X, Y) & & \quad = (n-1) S_X^2 & & \quad = (n-1) S_Y^2 \end{aligned}$$

where $S_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ and, $S_Y^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$ are the unbiased variance estimators of previous chapters.

so that

$$\mathbf{C} = \frac{1}{n S_{xx}} \begin{pmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{pmatrix}$$

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \mathbf{C} \mathbf{X}^T \mathbf{Y} = \frac{1}{n S_{xx}} \begin{pmatrix} \sum x_i^2 \sum Y_i - \sum x_i \sum x_i Y_i \\ -\sum x_i \sum Y_i + n \sum x_i Y_i \end{pmatrix} = \frac{1}{n S_{xx}} \begin{pmatrix} n \bar{Y} \sum x_i^2 - n \bar{x} \sum x_i Y_i \\ n S_{yy} \end{pmatrix}.$$

Fact 7.17 The OLS expressions for $\hat{\beta}_0$ and $\hat{\beta}_1$ are:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \quad \text{and} \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Fact 7.18 — The variance of $\hat{\beta}_0$ and $\hat{\beta}_1$.

$$V(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{S_{xx}}, \quad V(\hat{\beta}_0) = \hat{\sigma}^2 \frac{\sum x_i^2}{n S_{xx}}, \quad \text{and} \quad \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\hat{\sigma}^2 \frac{\sum x_i}{n S_{xx}}$$

where $\hat{\sigma}^2 = \frac{SSE}{n-2}$ and $\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-2}^2$

→ Since S_{xx} is proportional to $V(X)$, a more precise estimate the slope is obtained for x -values that are **more spread out**.

Significance test for β_1 : $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$

→ Reject H_0 if

$$\left| \frac{\hat{\beta}_1}{\hat{\sigma} / \sqrt{S_{xx}}} \right| > t_{\alpha/2}.$$

This test is important because if we can't reject H_0 it means that the variable X_i does not help explain Y and therefore should be removed from model.

The regression line passes by (\bar{x}, \bar{Y})

$$\begin{aligned} E(Y_i) &= \hat{\beta}_0 + \hat{\beta}_1 x_i \\ &= \bar{Y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i \\ &= \bar{Y} + \hat{\beta}_1 (x_i - \bar{x}) \end{aligned}$$

We can also estimate β_0 and β_1 without matrices by differentiating

$$SSE(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

and solving:

$$\begin{aligned} \frac{\partial SSE}{\partial \beta_0} &= -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial SSE}{\partial \beta_1} &= -2 x_i \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \end{aligned}$$

for β_0 and β_1 . We would get the same result.

SSE, R^2

$$\begin{aligned} SSE &= S_{yy} - \hat{\beta}_1 S_{xy} \\ &= S_{yy} - S_{xy}^2 / S_{xx}. \quad (\text{proof}) \end{aligned}$$

Since $SST = S_{yy}$ and $SSE = S_{yy} - S_{xy}^2 / S_{xx}$, we can see that

$$R^2 = \frac{S_{xy}^2}{S_{xx} S_{yy}}$$

$R \approx \rho_{X,Y}$, the correlation between X and Y . Recall

$$\begin{aligned}\rho(X,Y) &= \frac{\text{Cov}(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{E[(X - E[X])^2]E[(Y - E[Y])^2]}} \\ &\approx \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2 \frac{1}{n} \sum (y_i - \bar{y})^2}} \\ &= \frac{\frac{1}{n} S_{xy}}{\sqrt{\frac{1}{n} S_{xx} \frac{1}{n} S_{yy}}} \\ &= \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}} = R\end{aligned}$$

Connection between the slope and the correlation coefficient:

$$\hat{\beta}_1 = \frac{S_Y}{S_X} R.$$

→ Since $\hat{\rho} = R$, the test $H_0 : \beta_1 = 0$ is similar to $H_0 : \rho = 0$.

To see this,

$$\begin{aligned}R^2 &= \frac{S_{xy}^2}{S_{xx} S_{yy}} = \frac{S_{xy}^2}{S_{xx}^2} \frac{S_{xx}}{S_{yy}} = \hat{\beta}_1^2 \frac{S_{xx}}{S_{yy}} \\ &= \hat{\beta}_1^2 \frac{S_X^2}{S_Y^2}.\end{aligned}$$

and the result follows.

7.9.1 Predictions

Suppose we wish to predict $E(Y)$ at the point x_0 . Then $\mathbf{x}_0 = (1, x_0)$ so

$$V(\hat{Y}_0) = \hat{\sigma}^2 \mathbf{x}_0 \mathbf{C} \mathbf{x}_0^T = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right)$$

And we obtain:

$(1 - \alpha)\%$ **Confidence interval for $E(Y | \mathbf{x}_0)$:**

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

$(1 - \alpha)\%$ **Confidence interval for a particular realization of Y :**

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

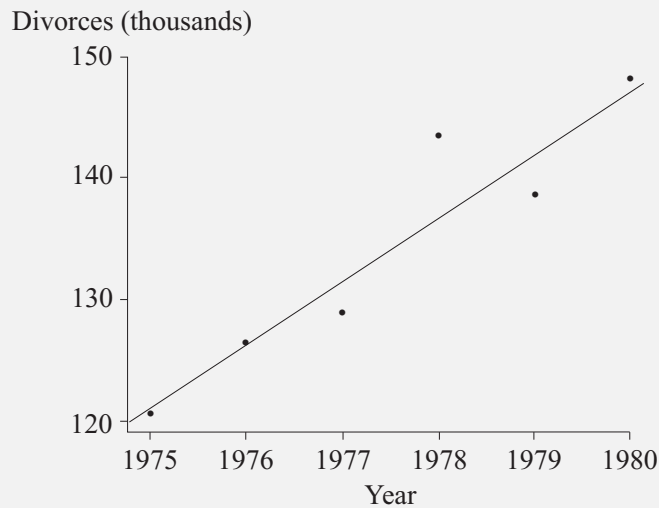
Example 208. — Divorces in England and Wales The table below gives the data and summary for:

Y = the annual number of divorces recorded in England and Wales between 1975 and 1980.
 x = years since 1974 ($x = 1$ means year 1975).

							Total
x_i	1	2	3	4	5	6	21
y_i	120.5	126.7	129.1	143.7	138.7	148.3	807.
$x_i y_i$	120.5	253.4	387.3	574.8	693.5	889.8	2919.3
x_i^2	1	4	9	16	25	36	91
y_i^2	14520.3	16052.9	16666.8	20649.7	19237.7	21992.9	109120.

We therefore obtain

n	\bar{x}	\bar{y}	S_{xy}	S_{xx}	S_{yy}	$\hat{\beta}_1$	$\hat{\beta}_0$	$\hat{\sigma}^2$	R^2
6.	3.5	134.5	94.8	17.5	578.72	5.417	115.54	16.294	0.887



Divorces in England and Wales with fitted line

Our estimate of the rate of increase of divorces is $\hat{\beta}_1 = 5.417$ and we would like to answer the question “Is the divorce rate changing?” In other words, we would like to test the null hypothesis

$$H_0 : \beta_1 = 0$$

Under H_0

$$T = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \sim t(4)$$

giving

$$t = \frac{5.417}{\sqrt{16.294/17.5}} = 5.614.$$

We therefore have strong evidence to reject the null hypothesis that the divorce rate is not changing; that is, there is strong evidence of an increasing divorce rate.

Example 209. — US Cars :

Data from 93 cars on sale in the USA in 1993.
 Data from 93 cars, selected at random, on sale in the US in 1993 with 27 variables Source: Lock, R. H. (1993) 1993 New Car Data. Journal of Statistics Education 1(1).

Y = City MPG (miles per US gallon by EPA rating).
 x = weight (pounds).

----- 95% Single Prediction CI - - - - 95% Mean Response CI

Data summary:

n	\bar{x}	\bar{y}	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$
93.	3072.9	22.366	6.13449×10^6	9.10188×10^8	49426.

we obtain the following:

S_{xy}	S_{xx}	S_{yy}	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	R^2
-257149.	3.2014×10^7	2905.57	47.048	-0.008	9.231	0.711

Software output:

	Estimate	Standard Error	t-Statistic	P-Value
α	-47.0484	1.67991	-28.0064	1.62701×10^{-46}
β	-0.00803239	0.000536985	-14.9583	2.96705×10^{-26}

More examples here.

Example 210. You recorded the speed of 7 individual vehicles on a highway segment with posted speed limit of 55 mph, Y, in mph, and the rainfall, X, at the time of each particular measurement, in millimeters per hour, mm/h. The following descriptive statistics were obtained:

$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$	n
125	377.44	5846.55	3012.5	21,580.36	7

- (a) Find the linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and interpret the meaning of the estimated parameters in this case.
- (b) Conduct a hypothesis test (5% significance level) to determine whether rainfall is a useful

linear predictor of vehicle speeds.

(c) Use hypotheses testing to assess whether or not the average speed in this highway exceeds the speed limit by 10 mph on non-rainy days.

(d) Consider the confidence interval for the speed of a particular vehicle. At what level of rainfall is this interval the narrowest? Calculate this interval and interpret its meaning.

(e) Test the hypotheses that the true variance of the regression model is at least 100.

(f) Estimate the probability that an individual driver will be traveling at 10 mph over the speed limit.

Solution:

$$\begin{aligned}
 S_{xx} &= \sum (x_i - \bar{x})^2 \\
 &= \sum x_i^2 - n\bar{x}^2 \\
 &= 780.36 \\
 S_{xy} &= \sum (x_i - \bar{x})(Y_i - \bar{Y}) \\
 &= \sum x_i Y_i - n\bar{x}\bar{Y} \\
 &= -893.45 \\
 S_{yy} &= \sum (Y_i - \bar{Y})^2 \\
 &= \sum Y_i^2 - n\bar{Y}^2 \\
 &= 1228.80 \\
 \hat{\beta}_1 &= \frac{S_{xy}}{S_{xx}} \\
 &= -1.1449 \\
 \hat{\beta}_0 &= \bar{y} - \beta_1 \bar{x} \\
 &= 74.365 \\
 SSE &= S_{yy} - \hat{\beta}_1 S_{xy} \\
 &= 205.89 \\
 \hat{\sigma}^2 &= \frac{SSE}{n-2} \\
 &= 41.178
 \end{aligned}$$

(a)

$$\begin{aligned}
 \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\
 &= 74.365 - 1.1449\hat{x}
 \end{aligned}$$

In this straight line, which shows the relationship between vehicle speed (Y) and rainfall (X), the intercept $\hat{\beta}_0$ is average speed when a non-rainy day, and the slope $\hat{\beta}_1$ represents the change in vehicle speed due to a unit change in the rainfall.

(b) We test $H_0 : \beta_1 = 0$ against $H_1 \neq 0$

$$\begin{aligned}
 \left| \frac{\hat{\beta}_1}{\hat{\sigma}/\sqrt{S_{xx}}} \right| &= \left| \frac{-1.1449}{\sqrt{41.178}/\sqrt{780.36}} \right| \\
 &= 4.98 > t_{0.025,5} = 2.5706
 \end{aligned}$$

so we reject H_0 which means rainfall is a useful linear predictor of vehicle speeds.

(c) We test $H_0 : \beta_0 + \beta_1 x = 65$ against $H_1 : \beta_0 + \beta_1 x > 65$, but on a non-rainy day, $x = 0$, so we are testing $H_0 : \beta_0 = 65$ against $H_1 : \beta_0 > 65$

$$\begin{aligned} \text{Var}(\hat{\beta}_0) &= \hat{\sigma}^2 \frac{\sum x_i^2}{nS_{xx}} \\ &= 41.178 \times \frac{3012.5}{7 \times 780.36} \\ &= 22.709 \\ \frac{\hat{\beta}_0 - 65}{\sqrt{\text{Var}(\hat{\beta}_0)}} &= \frac{74.365 - 65}{\sqrt{22.709}} \\ &= 1.9652 < t_{0.05,5} = 2.0150 \end{aligned}$$

so we accept H_0 , which means vehicle speeds do not exceed the speed limit by 10 mph on non-rainy days.

(d) The confidence interval is narrowest when rainfall level equals its average value: $x_0 = \bar{x} = 17.86$: the 95% confidence interval is

$$\begin{aligned} \hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{\alpha/2, n-2} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} &= 74.365 - 1.1449 \times 17.86 \pm 2.0150 \times 6.85 \\ &= 53.92 \pm 13.82 \\ &= (40.10, 67.74) \end{aligned}$$

so the confidence interval for vehicle speed when rainfall level is 17.86 mm/h is (40.10, 67.74) mph. As with any confidence interval, the interpretation is that the method we use to compute this interval will contain a realization of a particular vehicle speed 95% of the time; that is, if the experiment of taking a sample and computing this interval is repeated *many* times, then on average 95% of those intervals will contain a realization of a particular vehicle speed when $x_0 = \bar{x}$.

We *cannot* say: "This means, when rainfall level is 17.86 mm/h, the probability that vehicle speed is between 40.10 and 67.74 mph is 0.95."

(e)

$$\begin{aligned} H_0 &: \sigma^2 = 100 \\ H_1 &: \sigma^2 > 100 \\ C^2 &= \frac{(7-2)\hat{\sigma}^2}{100} \sim \chi_{n-2}^2 \\ &= 2.0589 \end{aligned}$$

Because $2.0589 < \chi_{df=5, \alpha=0.05}^2 = 11.07$, we cannot reject H_0 .

(f) Since rainfall is not specified, we assume $x_0 = \bar{x}$:

$$\begin{aligned} \hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x_0 \\ &= 74.365 - 1.1449 \times 17.86 = 53.92 \end{aligned}$$

Since Y is normally distributed with this mean and variance: $\hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 6.85$, we get

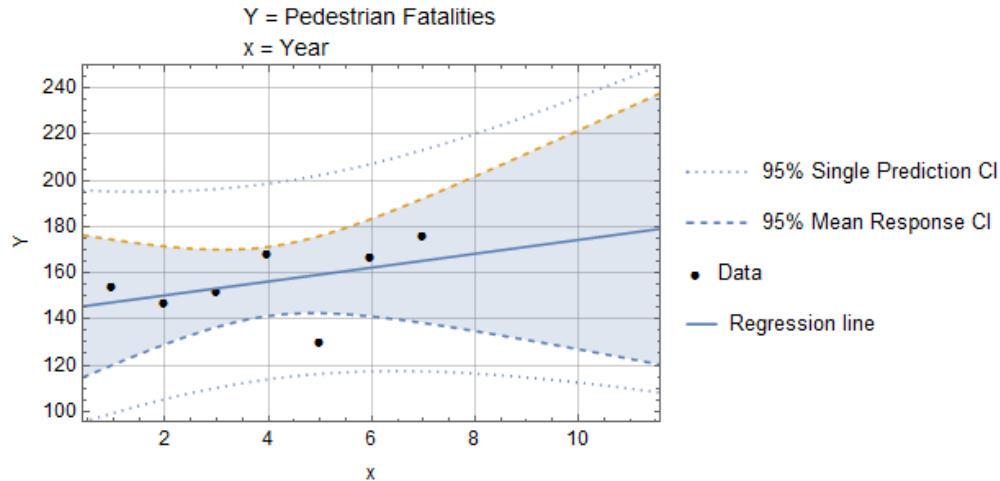
$$\begin{aligned} P(Y > 75) &= 1 - \Phi\left(\frac{75 - 53.92}{6.85}\right) \\ &= 0.053 \end{aligned}$$

and the probability that an individual driver will be traveling at 10 mph over the speed limit is 5.3%.

□

7.10 Problems

Problem 7.1 — Pedestrian Fatalities in Georgia. The number of yearly pedestrian fatalities in the state of Georgia (Y) is presented below for the years 2007-2013 ($x = 1$ means year 2007). Source. Given the data summary, answer the following questions.

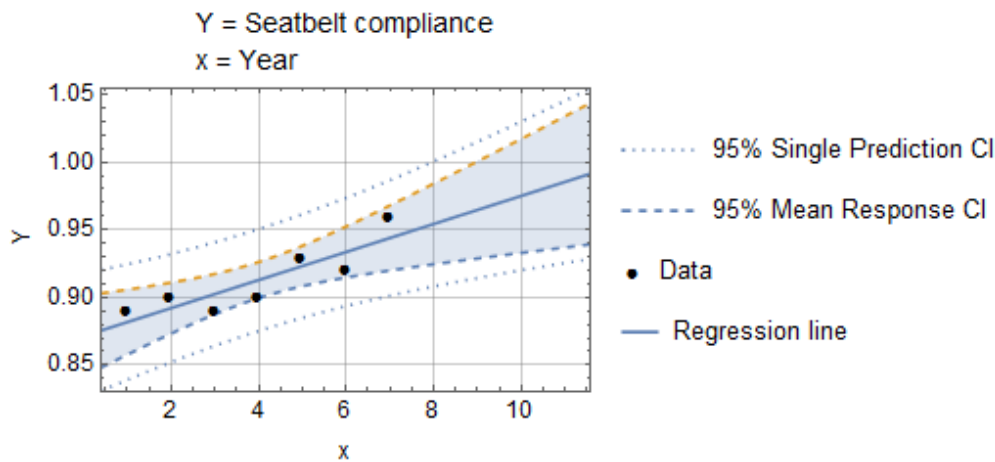


Data summary:

n	\bar{x}	\bar{y}	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$
7.	4.	156.286	4460.	140.	172418.

- Find the linear regression model $y = \hat{\beta}_0 + \hat{\beta}_1 x$ and interpret the meaning of the estimated parameters in this case.
- Calculate $\hat{\sigma}^2$ and R^2 and interpret their meaning.
- Would you say there is statistical evidence suggesting an increasing trend in pedestrian fatalities?
- Use the model to test the hypotheses that the expected number of pedestrian fatalities in 2017 ($x=11$) will be less than 120.
- Use the model to test the hypotheses that the true variance of the regression model is less than 200.

Problem 7.2 — Seatbelt compliance in Georgia. Data for seatbelt compliance is presented below for the years 2007-2013 ($x = 1$ means year 2007). Source. Given the data summary, answer the following questions.

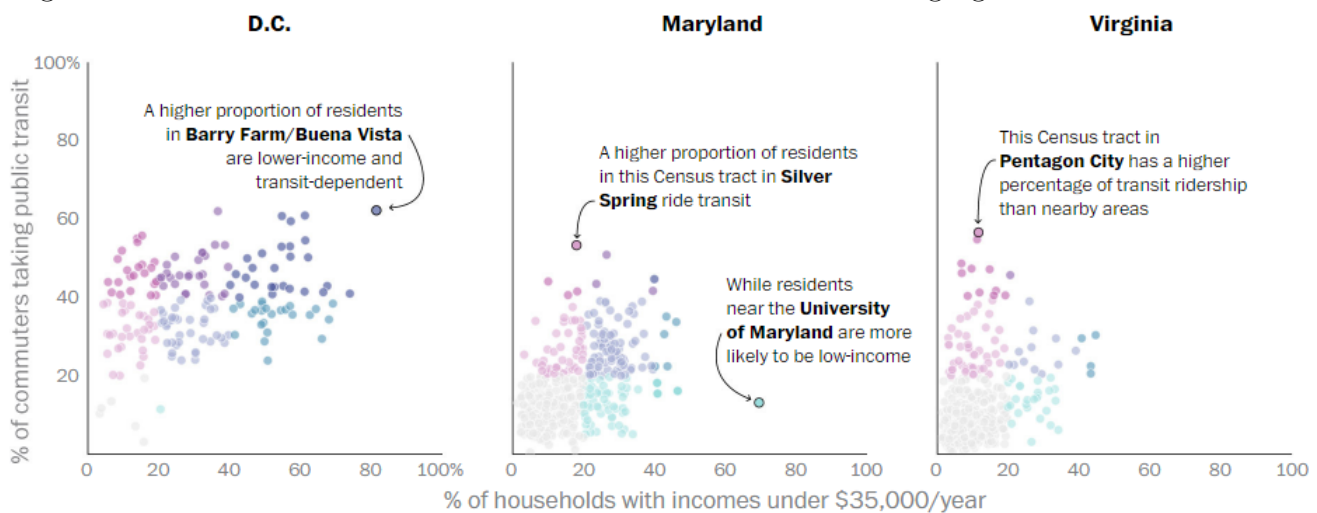


Data summary:

n	\bar{x}	\bar{y}	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$
7.	4.	0.913	25.85	140.	5.837

- Find the linear regression model for seatbelt compliance $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ and interpret the meaning of the estimated parameters in this case.
- Calculate $\hat{\sigma}^2$ and R^2 and interpret their meaning.
- Would you say there is statistical evidence suggesting an increasing trend in seatbelt compliance?
- Use the model to test the hypotheses that the expected seatbelt compliance in 2017 ($x=11$) will be less than 0.95.
- Use the model to test the hypotheses that the true variance of the regression model is less than 200.

Problem 7.3 — Washington low-income, transit-reliant residents. This article (<https://goo.gl/omGFCd>) appeared recently on the Washington Post, analyzes the relationship between income and transit usage in the D.C. area. The data can be summarized in the following figures.



The author of the article, however, did not give any statistical foundations to his observations and conclusions, the most prominent of which is that “D.C. has a higher concentration of low-income, transit-reliant residents than nearby counties in Virginia and Maryland.”

You are asked to use the DC data and Virginia data to fill this gap by using the statistical techniques

that you deem appropriate to verify/disprove the claims in this article. It is expected that you use at least two techniques and that you compare and comment the results.

Problem 7.4 The figure shows the average global temperature (relative to the year 1921) from 1880 to 2005. Although it may seem obvious from the figure that the temperature is increasing at a higher rate since the 70's, many people believe that such an increase can be explained by random fluctuations. The factor X denotes the number of years since 1879; i.e., 1880 corresponds to $x = 1$.

Sample	$\sum x_i$	$\sum y_i$	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$	n
1970-2005	3815	9.87	1180	419405	6.78	35
1880-1969	4186	-21.7	-693	255346	9.35	91
1880-2005	8001	-11.83	486	674751	16.13	126

- Using the whole sample, test the hypotheses that global warming can be explained by statistical fluctuations around a constant mean that does not grow in time.
- Test the hypotheses that the global warming rate has increased since the 70's (1970-2005) compared to the rate in (1880-1969).
- Provide a 95% confidence interval prediction for the global temperature for the year where this interval would be the narrowest, and interpret the meaning of this interval.
- It is believed that vehicle emissions, Z , are proportional to the square of the global temperature due to the increased use of air conditioning. Estimate the probability that emissions in the year 2017 will double the levels observed in year 2000.

Example 211. From the same survey on the previous question, here we fit a simpler model with only one factor:

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Based on the model estimation results summarized below,

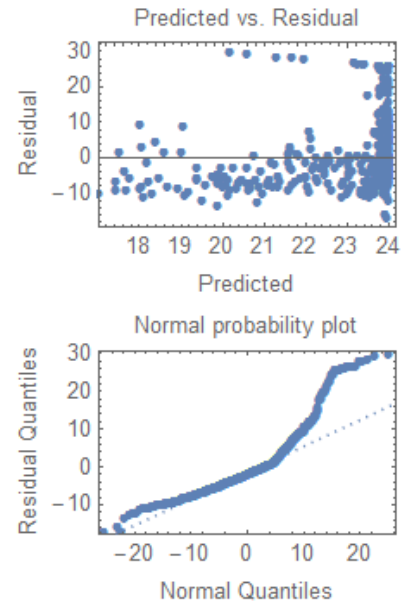
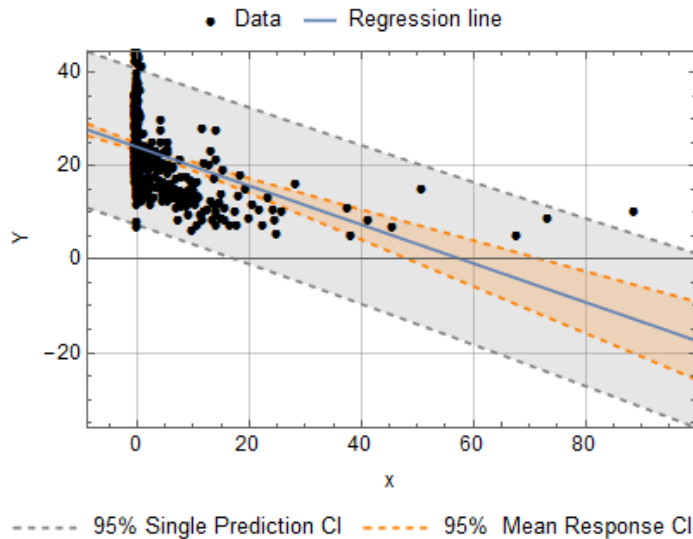
- What would you say are the problems of the fitted model? Clearly justify your answer in each case.
- As you can see from the scatter plot, the 95 percent confidence interval for Y when $x = 40$ is between 5 and 10. How can we interpret this interval?
- Do you agree with the slope of the regression line? Why?

Housing values in suburbs of Boston.

Home values for 506 Boston suburbs with potential influential factors. Source: Belsley D. A., Kuh, E. and Welsch, R. E. (1980) Regression Diagnostics. Identifying Influential Data and Sources of Collinearity. New York: Wiley.

Y = Median value of owner-occupied homes in \$1000s

x = Per capita crime rate by town



Data summary:

n	\bar{x}	\bar{y}	$\sum x_i y_i$	$\sum x_i^2$	$\sum y_i^2$
506.	3.614	22.533	25687.1	43970.3	299626.

we obtain the following:

S_{xy}	S_{xx}	S_{yy}	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\sigma}^2$	R^2
-15512.8	37363.2	42716.3	24.033	-0.415	71.975	0.151

Software output:

	Estimate	Standard Error	t-Statistic	P-Value
α	24.0331	0.409142	58.7403	1.34172×10^{-227}
β	-0.41519	0.0438904	-9.45971	1.17399×10^{-19}