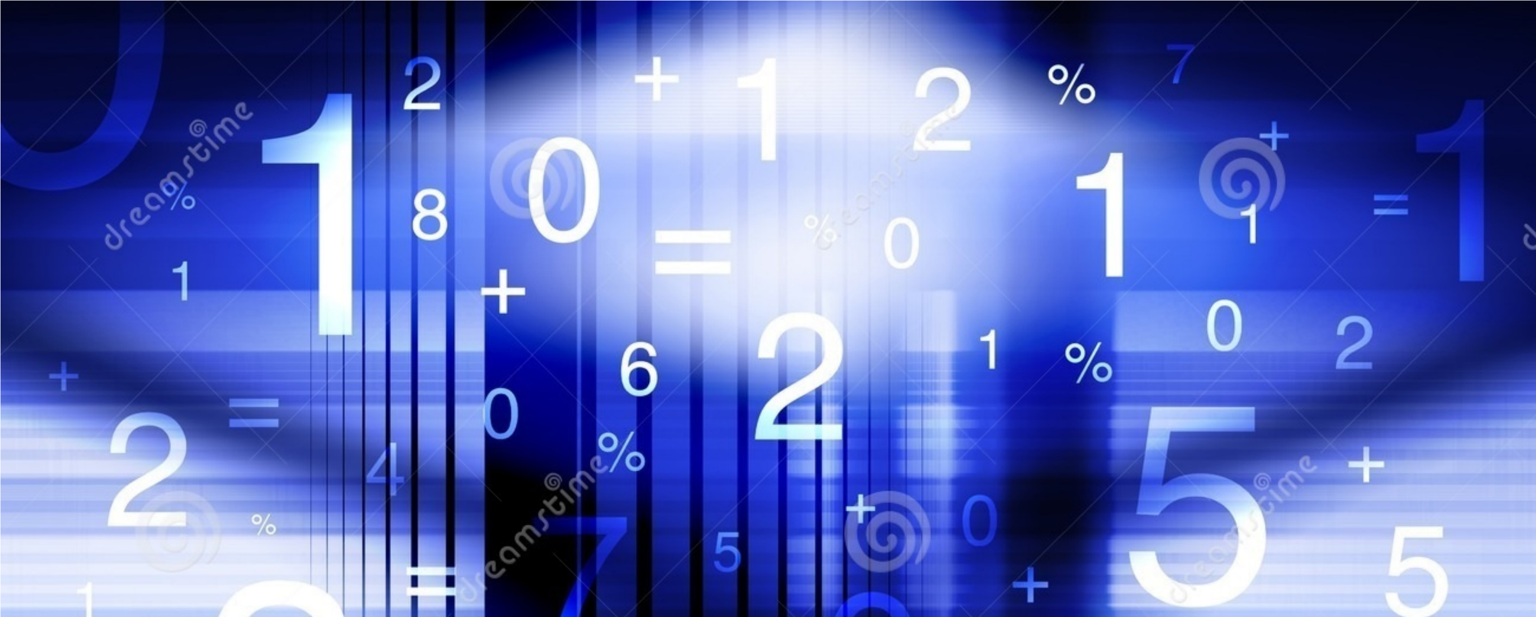
Notes on Probability and Statistics



Contentsmkboth CONTENTSCONTENTS

Probability

4. Special Distributions

4.1 Uniform Random Variable

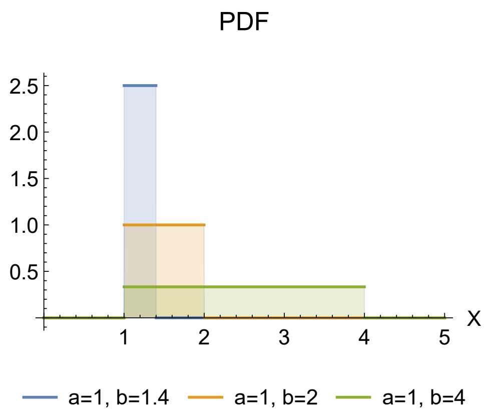
Uniform Random Variable over the interval (a,b) denoted as $X \sim \mathrm{U}(a,b)$

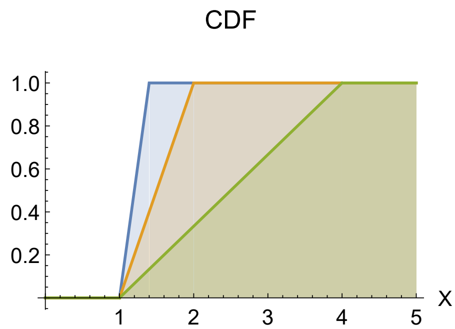
$$f_X(x) = \begin{cases} \frac{1}{b-a} & ; \quad a < x < b \\ 0 & ; \quad \text{otherwise} \end{cases}$$

$$\mathrm{F}_X(x) = \begin{cases} 0 & ; \quad x < a \\ \frac{x-a}{b-a} & ; \quad a \le x \le b \\ 1 & ; \quad b < x \end{cases}$$

$$\mathrm{E}[X] = \frac{a+b}{2}$$

$$\mathrm{V}(X) = \frac{(b-a)^2}{12}$$





Example 1. Let

$$X \sim \mathrm{U}(a,b)$$

Show that

$$\mathrm{E}[X] = \frac{a+b}{2}$$

Solution:

$$\begin{aligned}
\mathrm{E}[X] &= \int_{-\infty}^{\infty} x f(x)\,\mathrm{d}x \\
&= \int_{a}^{b} x \frac{1}{b-a}\,\mathrm{d}x = \frac{b^2-a^2}{2(b-a)} \\
&= \frac{a+b}{2}
\end{aligned}$$

$\square$

Example 2. Let

$$X \sim \mathrm{U}(a,b)$$

Show that

$$\mathrm{V}(X) = \frac{(b-a)^2}{12}$$

Solution:

$$\begin{aligned}\mathrm{E}\left[X^2\right] &= \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = \int_a^b x^2 \frac{1}{b-a}\,\mathrm{d}x = \frac{b^3-a^3}{3(b-a)} \\ &= \frac{a^2+ab+b^2}{3}\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{V}(X) &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \\
&= \frac{a^2 + ab + b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{(b-a)^2}{12}
\end{aligned}$$

$\square$

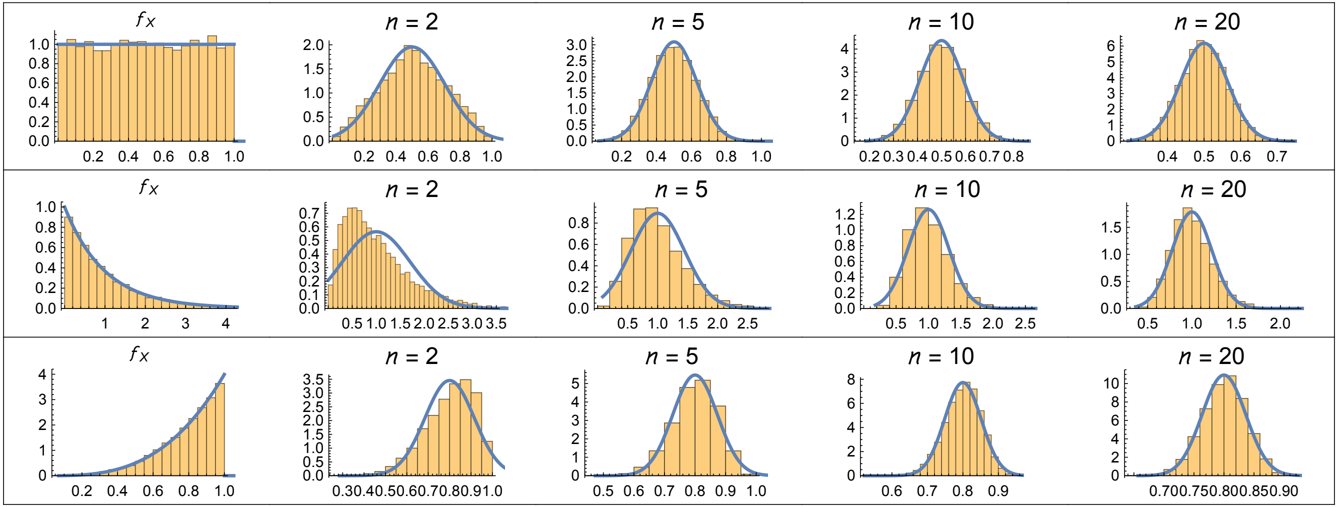
4.2 Normal Distribution

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

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{4.1}$$

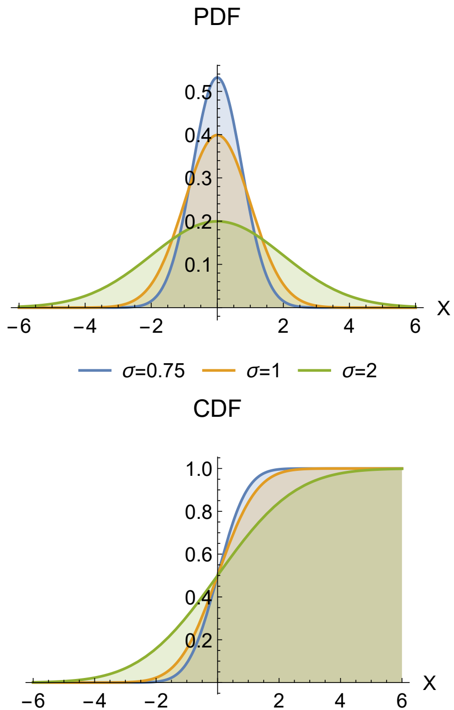
where μ and σ^2 are parameters, and

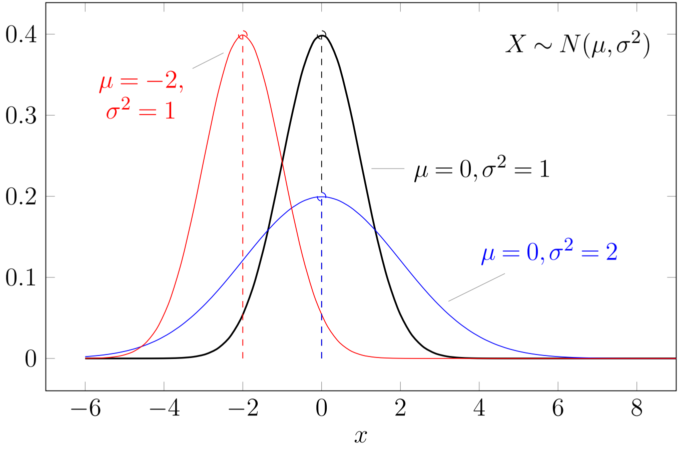
$$\mathrm{E}(X) = \mu$$
$$\mathrm{V}(X) = \sigma^2$$

The CDF of a normal rv :

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(a-\mu)^2}{2\sigma^2}} \, \mathrm{d}a$$

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:

$$\begin{aligned}\Phi(z) &= \mathrm{P}(Z \leq z) \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-a^2/2}\,\mathrm{d}a\end{aligned}$$

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.

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.

$$\mathrm{F}_Y(y) = \mathrm{P}(Y \leq y) = \mathrm{P}(aX + b \leq y) = \mathrm{P}\left(X \leq \frac{y-b}{a}\right) = \mathrm{F}_X\left(\frac{y-b}{a}\right)$$

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:

$$\boxed{x_\alpha = \mu + z_\alpha \sigma} \tag{4.2}$$

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

Fact 4.4 — The Central Limit Theorem (CLT). The linear combination

$$U = \sum_{i=1}^{n} a_i X_i = \mathbf{a}^{\mathbf{T}} \mathbf{X}$$

with $\mathbf{a} = (a_1, \ldots, a_n)^T$ **tends to the normal distribution as** $n \to \infty$ with

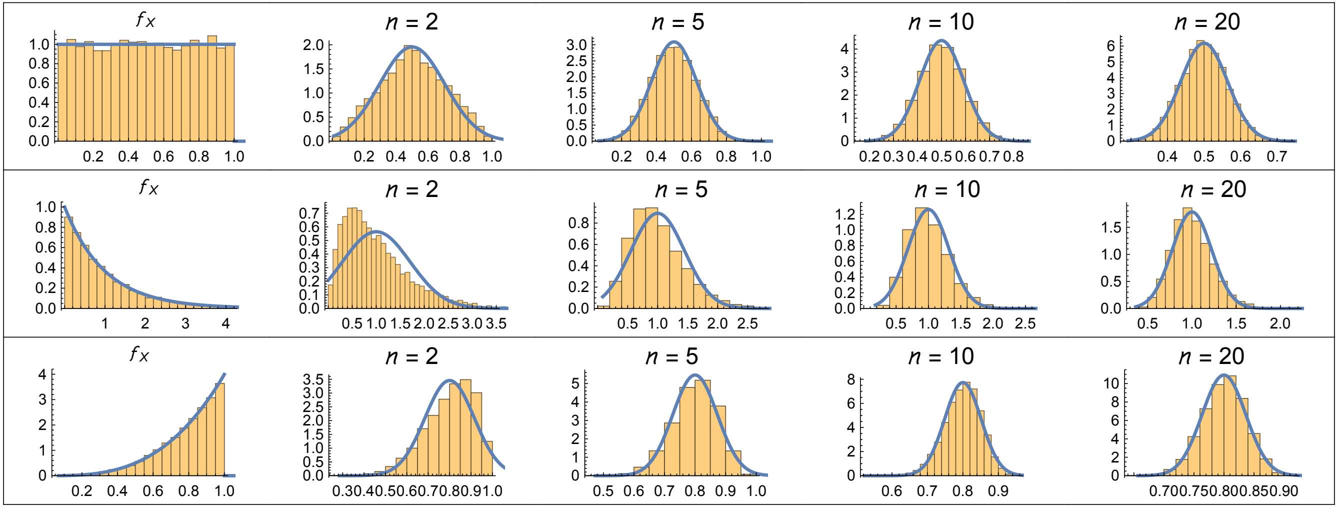
$$\begin{aligned}
\mathrm{E}(U) &= \sum_{i=1}^{n} a_i \mu_i = \mathbf{a}^T \boldsymbol{\mu}, \\
\mathrm{V}(U) &= \mathbf{a}^{\mathbf{T}} \boldsymbol{\Sigma}_{\mathbf{X}} \mathbf{a} \\
&= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \underbrace{2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_j \sigma_{i,j}}_{0 \text{ if } X_i\text{'s are independent}}
\end{aligned}$$

where

$$\Sigma_{\mathbf{X}} = \mathrm{E}\left((\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^T\right) = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \cdots \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is the Covariance matrix.

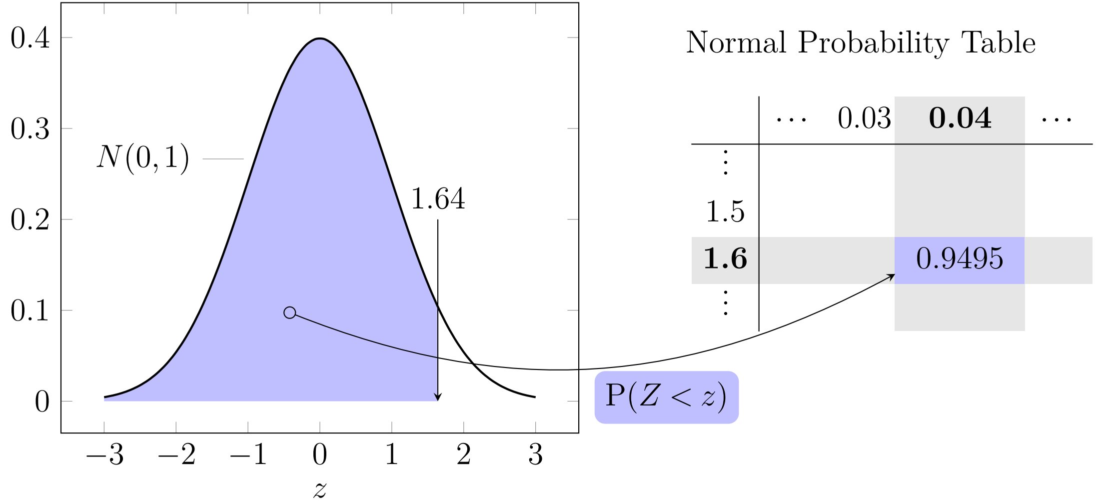
The figure below shows the agreement of the CLT for the PDF of $U = \frac{1}{n}\sum_{i=1}^{n} X_i$ where the $X_i \sim f_X$.



It can be seen that regardless of the initial distribution f_X that CLT provides a good approximation for $n > 5$.

4.2.2 How to read normal probability tables

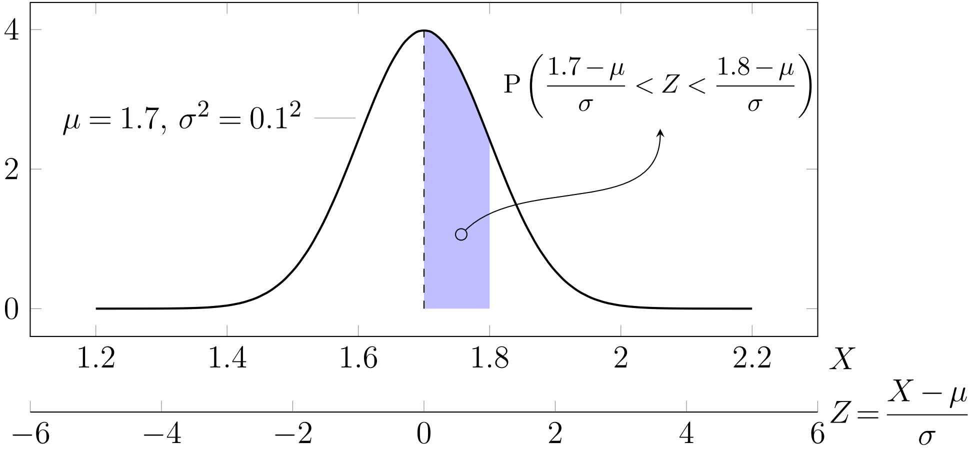
→ Download a **normal probability tables**.



→normal probability calculation with the TI 83/84.

Example 3. If $X \sim N(\mu,\sigma^2)$ with $\mu = 1.7$, $\sigma^2 = 0.1^2$, calculate $\mathrm{P}(1.7 < X < 1.8)$

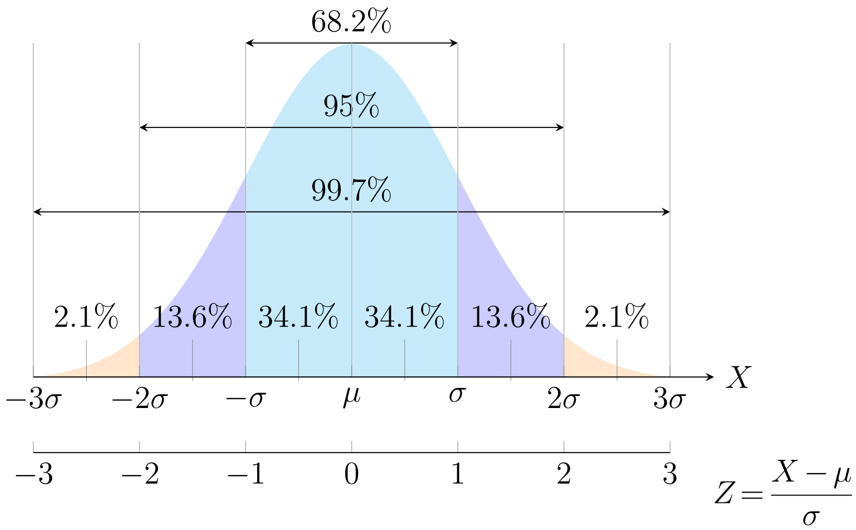
Solution:



$$\begin{aligned}
\mathrm{P}\,(1.7 < X < 1.8) &= \mathrm{P}\left(\frac{1.7-\mu}{\sigma} < Z < \frac{1.8-\mu}{\sigma}\right) \\
&= \mathrm{P}(0 < Z < 1) \\
&= \Phi(1) - \Phi(0) \\
&= 0.8413 - 0.5 \\
&= 0.3413
\end{aligned}$$

$\square$

4.2.3 The "68-95-99.7 Rule"



Example 4. 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 5. 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.
(a) When should the test be terminated if we wish to allow sufficient time for 90% of the students to complete the test?
(b) 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 $Z = \frac{X-50}{12} \sim N(0,1)$.

a) 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.

b) $P(30 < X < 60) = P(-1.67 < Z < 0.83) = 0.75$.

$\square$

Example 6. Let X be a normal random variable with parameters

$$\mu = 3$$
$$\sigma^2 = 9$$

Find

a) $\mathrm{P}(2 < X < 5)$

b) $\mathrm{P}(X > 3)$

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,

$$\begin{aligned}
\mathrm{P}(X > 3) &= \mathrm{P}\left(\frac{X-3}{3} > \frac{3-3}{3}\right) \\
&= \mathrm{P}(Z > 0) \\
&= 0.5
\end{aligned}$$

$\square$

Example 7. 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

$$1000 + 200(1.28) = 1256 \text{ points.}$$

$\square$

Example 8. — *** 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)

$$\begin{aligned} P(congestion) &= 1-\phi(\frac{350-200}{50}) \\ &= 0.00135 \end{aligned}$$

(b) Mean: $200 \times (1+10\% \times 10) = 400$

standard deviation: $\frac{50 \cdot 400}{200} = 100$

$$\begin{aligned} P(congestion) &= 1 - \phi(\frac{350-400}{100}) \\ &= 0.69146 \end{aligned}$$

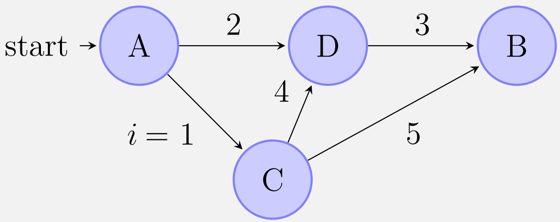
(c) New capacity:

$$\begin{aligned} Capacity &= 400 + 3 \times 100 \\ &= 700 \end{aligned}$$

$\square$

Example 9. — 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

$$\Sigma_{\mathbf{X}} = \begin{pmatrix} 4. & 0.66 & 0.66 & 0.16 & 0.4 \\ 0.66 & 10.89 & 7.62 & 0.26 & 0.66 \\ 0.66 & 7.62 & 10.89 & 1.06 & 0.66 \\ 0.16 & 0.26 & 1.06 & 0.64 & 0.16 \\ 0.4 & 0.66 & 0.66 & 0.16 & 4. \end{pmatrix}$$

a) what is the fastest route from A to B?
b) what is the probability that route A D B is faster than A C B?
c) what is the probability that route A D B is faster than A C D B?

Solution:

a) what is the fastest route from A to B?

Let:

Route 1: $A \to D \to B$

Route 2: $A \to C \to B$

Route 3: $A \to C \to D \to B$

T_i = Travel time on route $i = 1,2,3$.

Then,

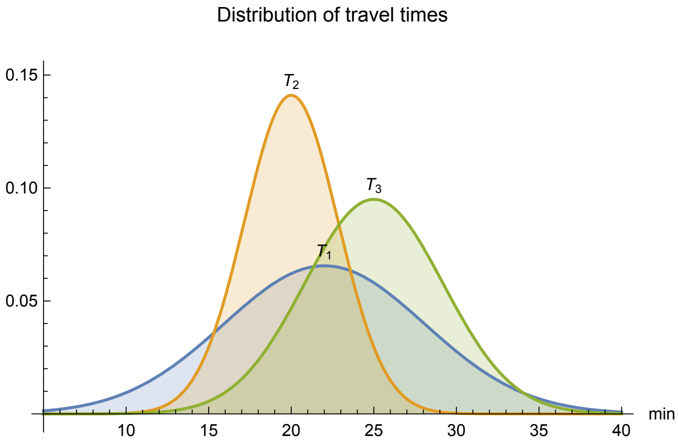
$$T_1 = X_2 + X_3$$
$$T_2 = X_1 + X_5$$
$$T_3 = X_1 + X_3 + X_4$$

Since travel times are the sum of normal random variables,
$\boxed{T_i \sim N\left(\mathrm{E}(T_i), \mathrm{V}(T_i)\right)}$ with:

$$\mathrm{E}(T_1) = \mu_2 + \mu_3 \quad = 22$$
$$\mathrm{E}(T_2) = \mu_1 + \mu_5 \quad = 20$$
$$\mathrm{E}(T_3) = \mu_1 + \mu_3 + \mu_4 \quad = 25$$

and

$$\begin{aligned}
V(T_1) &= 2\sigma_{2,3} + {\sigma^2}_2 + {\sigma^2}_3 && = 37. \\
V(T_2) &= 2\sigma_{1,5} + {\sigma^2}_1 + {\sigma^2}_5 && = 8.8 \\
V(T_3) &= 2\sigma_{1,4} + 2\sigma_{1,3} + 2\sigma_{4,3} + {\sigma^2}_1 + {\sigma^2}_4 + {\sigma^2}_3 && = 19.28
\end{aligned}$$



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?

$$
\begin{aligned}
\mathrm{P}(T_1 < T_2) &= \mathrm{P}(T_1 - T_2 < 0), \qquad \text{let } Y = T_1 - T_2 \\
&= \mathrm{P}(Y < 0) = 0.376726
\end{aligned}
$$

Since Y is a linear combination of $T_1 - T_2$, it is also normally distributed $\boxed{Y \sim N\left(\mathrm{E}(Y), \mathrm{V}(Y)\right)}$ with:

$$
\begin{aligned}
\mathrm{E}(Y) &= \mathrm{E}(T_1) - \mathrm{E}(T_2) && = 2 \\
\mathrm{V}(Y) &= \mathrm{V}(T_1) + \mathrm{V}(T_2) - 2\mathrm{Cov}(T_1, T_2) && = 40.54
\end{aligned}
$$

make sure you understand the negative sign!

$$
\mathrm{Cov}(T_1, T_2) = \sigma_{1,2} + \sigma_{1,3} + \sigma_{2,5} + \sigma_{3,5} = 2.64
$$

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:

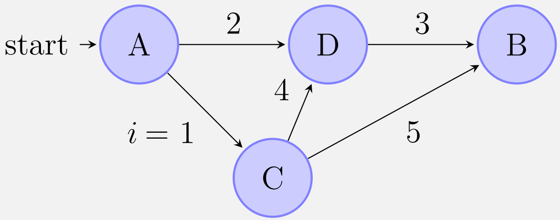
$$\mathrm{Cov}\,(T_1,T_3) = \sigma_{1,2}+\sigma_{1,3}+\sigma_{2,3}+\sigma_{2,4}+\sigma_{3,4}+{\sigma^2}_3 = 21.2$$

and $\mathrm{P}(T_1 - T_3 < 0) = 0.788644$

$\square$

Example 10. — Travel time from point A and to point B (again) In the transportation network below, let

$$X_i = \text{the travel time on link } i = 1, 2 \ldots 5$$



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

$$\Sigma_{\mathbf{x}} = \begin{pmatrix} 1. & 0. & 0. & 0. & 0. \\ 0. & 12.96 & 9.83 & 0. & 0. \\ 0. & 9.83 & 15.21 & 0.62 & 0. \\ 0. & 0. & 0.62 & 0.16 & 0. \\ 0. & 0. & 0. & 0. & 1. \end{pmatrix}$$

a) what is the fastest route from A to B?
b) what is the probability that route A D B is faster than A C B?
c) what is the probability that route A D B is faster than A C D B?

Solution:

a) what is the fastest route from A to B?

Let:
Route 1: $A \to D \to B$
Route 2: $A \to C \to B$
Route 3: $A \to C \to D \to B$
T_i = Travel time on route $i = 1, 2, 3$.
Then,

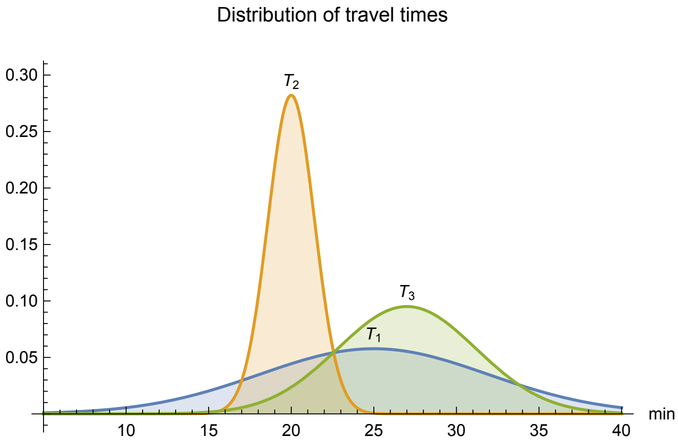
$$T_1 = X_2 + X_3$$
$$T_2 = X_1 + X_5$$
$$T_3 = X_1 + X_3 + X_4$$

Since travel times are the sum of normal random variables,
$\boxed{T_i \sim N\left(\mathrm{E}(T_i), \mathrm{V}(T_i)\right)}$ with:

$$\mathrm{E}(T_1) = \mu_2 + \mu_3 \qquad = 25$$
$$\mathrm{E}(T_2) = \mu_1 + \mu_5 \qquad = 20$$
$$\mathrm{E}(T_3) = \mu_1 + \mu_3 + \mu_4 \qquad = 27$$

and

$$
\begin{aligned}
V(T_1) &= 2\sigma_{2,3} + {\sigma^2}_2 + {\sigma^2}_3 && = 47.8 \\
V(T_2) &= 2\sigma_{1,5} + {\sigma^2}_1 + {\sigma^2}_5 && = 2 \\
V(T_3) &= 2\sigma_{1,4} + 2\sigma_{1,3} + 2\sigma_{4,3} + {\sigma^2}_1 + {\sigma^2}_4 + {\sigma^2}_3 && = 17.6
\end{aligned}
$$

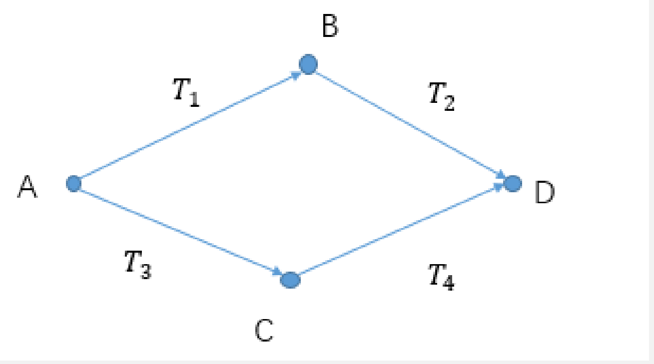


Route 2 is the fastest on average, but for risk-taking people Route 1 could be beneficial.

b) $\mathrm{P}(T_1 - T_2 < 0) = 0.239367$
c) $\mathrm{P}(T_1 - T_3 < 0) = 0.702722$

□

Example 11. — * **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:

$$\begin{aligned} T_1 &\sim N(8,4) \\ T_2 &\sim N(5,1) \\ T_3 &\sim N(5,4) \\ T_4 &\sim N(7,4) \end{aligned}$$

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:

$$\begin{aligned} T_J &= T_3+T_4 \\ \mu_{T_J} &= 5+7=12 \\ \sigma_{T_J} &= \sqrt{4+4+2\times 0.8\times 2\times 2} \\ &= 3.795 \end{aligned}$$

Hence

$$\begin{aligned} P(T_J<12) &= \Phi(\frac{12-12}{3.795}) \\ &= \Phi(0) \\ &= 0.5 \end{aligned}$$

(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. □

Example 12. — * **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:

$$\begin{aligned}P(X > 1250) &= 1 - \Phi(\frac{1250 - \mu_X}{\sigma_X}) \\ &= 1 - \Phi(-1.67) \\ &= 0.953\end{aligned}$$

$\square$

4.3 Lognormal Distribution

$$X \sim \mathrm{LogN}(\lambda, \xi^2), X > 0 \leftrightarrow \boxed{\log X \sim N(\lambda, \xi^2)}$$

$$\lambda = \mathrm{E}(\log X) \quad , \quad \xi^2 = \mathrm{V}(\log X)$$

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:

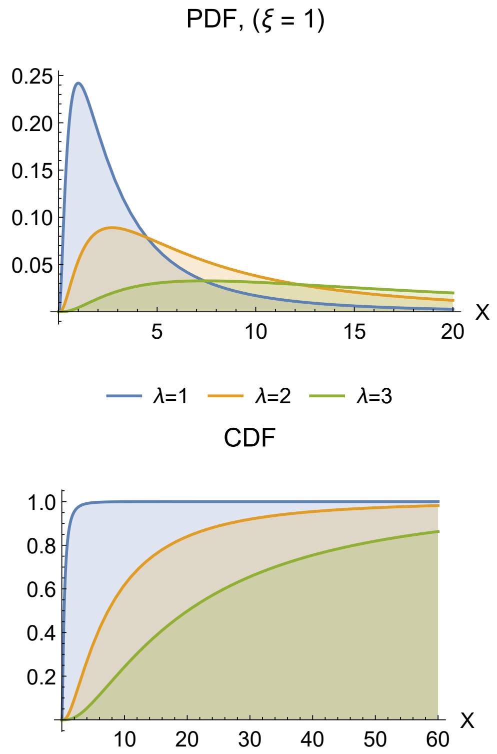
$$\mathrm{E}(X) = e^{\lambda + \xi^2/2}$$

$$\mathrm{V}(X) = e^{2\lambda + \xi^2}(e^{\xi^2} - 1)$$

and hence the coefficient of variation squared is

$$\delta_X^2 = e^{\xi^2} - 1$$
$$\approx \xi^2 \quad \text{when } \xi^2 \text{ is small, say } \xi < 1/3.$$

Note: $\lambda = \log x_{0.5}$



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:

$$\delta_X^2 = \sigma_X^2 / \mu_X^2 \tag{4.3}$$

$$\xi^2 = \log(1 + \delta_X^2) \quad \text{or } \xi^2 = \delta_X^2 \text{ if } \delta_X^2 \text{ is small, say} < 1/3. \tag{4.4}$$

$$\lambda = \log \mu_X - \xi^2/2 \tag{4.5}$$

2. if we are given $x_{0.5}, \delta_X$:

$$\xi^2 = \log(1 + \delta_X^2) \quad \text{or } \xi^2 = \delta_X^2 \text{ if } \delta_X \text{ is small, say} < 1/3. \tag{4.6}$$

$$\lambda = \log x_{0.5} \tag{4.7}$$

Once the parameters λ and ξ are determined, we can calculate probabilities using the standard normal tables:

$$\boxed{\mathrm{P}(X < x) = \mathrm{P}(\log X < \log x) = \Phi\left(\frac{\log x - \lambda}{\xi}\right)} \tag{4.8}$$

because $\log X \sim N(\lambda, \xi^2)$.
Percentiles x_α can also be obtained from the standard normal percentiles, z_α:

$$\boxed{x_\alpha = e^{\lambda + z_\alpha \cdot \xi}} \tag{4.9}$$

This is because, by definition, $\mathrm{P}(X < x_\alpha) = \alpha$, and in this case we have:

$$\begin{aligned} \mathrm{P}(X < x_\alpha) &= \mathrm{P}(\log X < \log x_\alpha) \\ &= \Phi\left(\frac{\log x_\alpha - \lambda}{\xi}\right) = \alpha \end{aligned}$$

the last equality implies that $\frac{\log x_\alpha - \lambda}{\xi}$ is the standard normal percentile, z_α and (4.9) follows.

Example 13. Lifetimes of a certain component are lognormally distributed with its **median** 3 days and parameter $\xi = 0.5$ days.
(a) Find the **mean** lifetime of these components
(b) 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) $\delta_X = \sqrt{e^{\xi^2} - 1} = 0.533$
$\sigma_X = \delta_X \mu_X = 0.533 \times 3.40 = 1.81$

$\square$

Example 14. — 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 $\mathrm{P}(T > t^*) = 0.95$ or equivalently $\mathrm{P}(T < t^*) = 0.05$ which means that t^* is the 5th percentile:

$$t^* = t_{0.05} = e^{\lambda + z_{0.05} \cdot \xi}$$

which gives 3.66 months. $\square$

Example 15. — * 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,

$$
\begin{aligned}
P(X < 5.5) &= \Phi(\frac{\log 5.5 - \lambda}{\xi}) \\
&= \Phi(-0.41) \\
&= 0.341
\end{aligned}
$$

(b)

$$\begin{aligned} P(X > 5.5 | X > 4) &= \frac{1 - P(X < 5.5)}{1 - P(X < 4)} \\ &= \frac{1 - 0.341}{1 - \Phi(\frac{\log 4 - \lambda}{\xi})} \\ &= 0.708 \end{aligned}$$

(c)

$$\begin{aligned} P(X < x_{85}) &= \Phi(\frac{\log(x_{85}) - \lambda}{\xi}) \\ &= 0.85 \end{aligned}$$

Referring to standard normal table: $\Phi(1.04) = 0.85$

$$\begin{aligned} x_{85} &= e^{1.04\xi + \lambda} \\ &= 8.47 \end{aligned}$$

$\square$

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 $n \to \infty$ with parameters:

$$\lambda_U = \mathbf{a^T} \boldsymbol{\lambda} = \sum_{i=1}^{n} a_i \lambda_i,$$

$$\xi_U^2 = \mathbf{a^T} (\boldsymbol{\Sigma}_{\log \mathbf{X}})\, \mathbf{a} = \sum_{i=1}^{n} a_i^2 \xi_i^2 + \underbrace{2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_j \xi_{i,j}}_{0 \text{ if } X_i\text{'s are independent}}$$

where

$$\begin{aligned}
\mathbf{a} &= (a_1, a_2, \dots a_n)^T \\
\lambda_i &= \mathrm{E}(\log X_i) \\
\boldsymbol{\lambda} &= (\lambda_1, \lambda_2, \dots \lambda_n)^T \\
\log \mathbf{X} &= (\log X_1, \log X_2, \dots \log X_n)^T
\end{aligned}$$

and:

$$\Sigma_{\log \mathbf{X}} = \mathrm{E}\Big((\log \mathbf{X} - \boldsymbol{\lambda})(\log \mathbf{X} - \boldsymbol{\lambda})^T\Big) = \begin{pmatrix} \xi_1^2 & \xi_{12} & \xi_{13} & \cdots \\ \xi_{21} & \xi_2^2 & \xi_{23} & \cdots \\ \xi_{31} & \xi_{32} & \xi_3^2 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

is the Covariance matrix of the $\log X_i$, i.e. $\xi_{ij} = \mathrm{Cov}\Big(\log X_i, \log X_j\Big)$ and $\xi_i^2 = \mathrm{V}(\log X_i)$.

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:

$$\log U \sim N(\mathrm{E}(\log U), \mathrm{V}(\log U))$$

with:

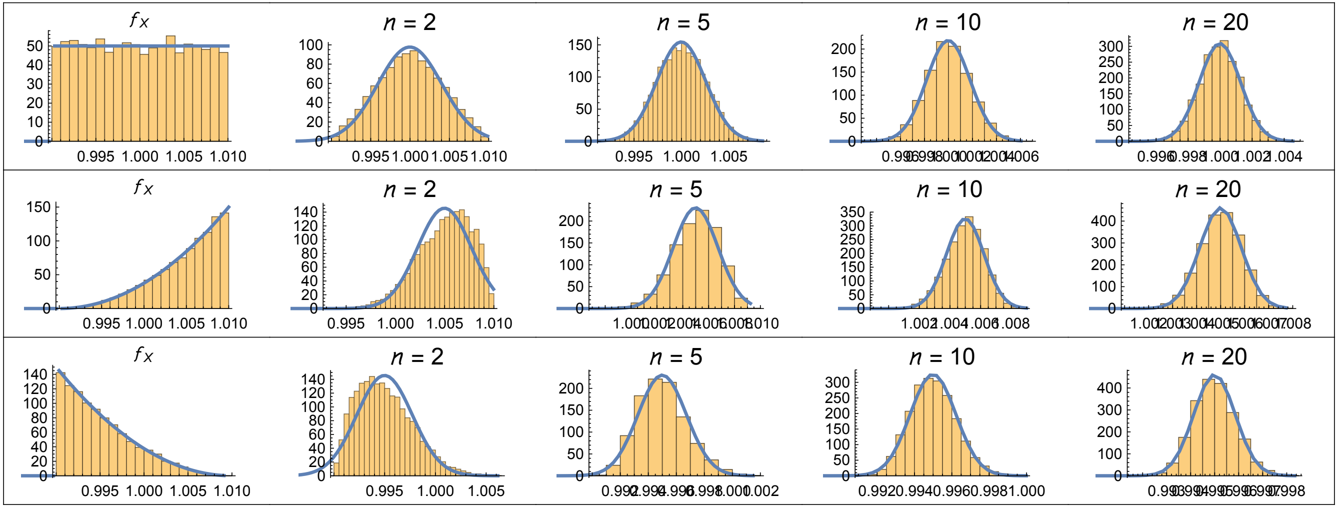
$$\begin{aligned}
\mathrm{E}(\log U) &= \sum_{i=1}^{n} a_i \mathrm{E}(\log X_i) \\
&= \mathbf{a}^T \boldsymbol{\lambda}, \\
\mathrm{V}(\log U) &= \mathbf{a}^{\mathbf{T}} (\boldsymbol{\Sigma}_{\log \mathbf{X}}) \mathbf{a} \\
&= \sum_{i=1}^{n} a_i^2 \xi_i^2 + \underbrace{2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_j \xi_{i,j}}_{0 \text{ if } X_i\text{'s are independent}}
\end{aligned}$$

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 $X \sim \mathrm{LogN}(\lambda, \xi^2)$ then
>
> $$Y = cX \sim \mathrm{LogN}(\log c + \lambda, \xi^2)$$
>
> for any constant c. To see this, simply treat the constant as a lognormal rv with zero variance.

Example 16. Let:

$$U = \frac{7.58\sqrt{X_1}X_3^2}{X_2\sqrt[3]{X_4}}$$

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	${\delta^2}_i$	σ_i	λ_i	${\xi^2}_i$	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:

$$\mathrm{E}(U) = e^{\lambda_U + \xi_U^2/2} = 72.04$$

$$\mathrm{V}(U) = e^{2\lambda_U + \xi_U^2}(e^{\xi_U^2} - 1) = 2299.26$$

(b) Assuming $U \sim \mathrm{LogN}(\lambda_U, \xi_U^2)$:

$$\begin{aligned} P(U < 94) &= \Phi(\frac{\log 94 - \lambda_U}{\xi_U}) = \Phi(0.742155) \\ &= 0.771003 \end{aligned}$$

$\square$

Example 17. Let:

$$W = \frac{1.46 X_1 X_4^2}{X_2^2 \sqrt{X_3}}$$

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 $\mathrm{P}(W < 4110)$

Solution: (a) Using equations (4.3) we can calculate:

i	X_i	μ_i	${\delta^2}_i$	σ_i	λ_i	${\xi^2}_i$	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:

$$\begin{aligned}
\lambda_W &= \log\left(\frac{1}{2 \cdot 32.2}\right) + \sum_{i=1}^{n} a_i \lambda_i, \\
&= 1.09134 \\
\xi_W^2 &= \sum_{i=1}^{n} a_i^2 \xi_i^2 \\
&= 0.148124
\end{aligned}$$

and the mean and variance of W are approximately:

$$\mathrm{E}(W) = e^{\lambda_W + \xi_W^2/2} = 2190.43$$

$$\mathrm{V}(W) = e^{2\lambda_W + \xi_W^2}(e^{\xi_W^2} - 1) = 98,759,027$$

(b) Assuming $W \sim \mathrm{LogN}(\lambda_W, \xi_W^2)$:

$$\begin{aligned} P(W < 4110) &= \Phi(\frac{\log 4110 - \lambda_W}{\xi_W}) = \Phi(0.0188) \\ &= 0.507 \end{aligned}$$

□

Example 18. — *** 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/sec^2. 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<3ft) [Hint: CLT]

Solution: (a) Using equations (4.3) we can calculate:

i	X_i	μ_i	${\delta^2}_i$	σ_i	λ_i	${\xi^2}_i$	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:

$$\mathrm{E}(H) = e^{\lambda_H + \xi_H^2/2} = 3.20722$$

$$\mathrm{V}(H) = e^{2\lambda_H + \xi_H^2}(e^{\xi_H^2} - 1) = 1.64227$$

(b) Assuming $H \sim \mathrm{LogN}(\lambda_H, \xi_H^2)$:

$$\begin{aligned} P(H < 3) &= \Phi(\frac{\log 3 - \lambda_H}{\xi_H}) = \Phi(0.0188) \\ &= 0.507 \end{aligned}$$

□

Example 19. 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<3ft) (ans: 0.684)

Example 20. — Stock Price Distribution For a given stock, let:

P_t = stock price at time period $t = 1, 2, \ldots$

$R_t = (P_t - P_{t-1})/P_{t-1}$ = 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:

$$\begin{aligned} P_t &= P_0(1+R_1)(1+R_2)\ldots(1+R_t) \\ &= P_0 \prod_{j=1}^{t} S_j \end{aligned}$$

By the CLT for products $P_0 \prod_{j=1}^{t} S_j$ tends to the LogN(λ_t, ξ_t^2) with parameters:

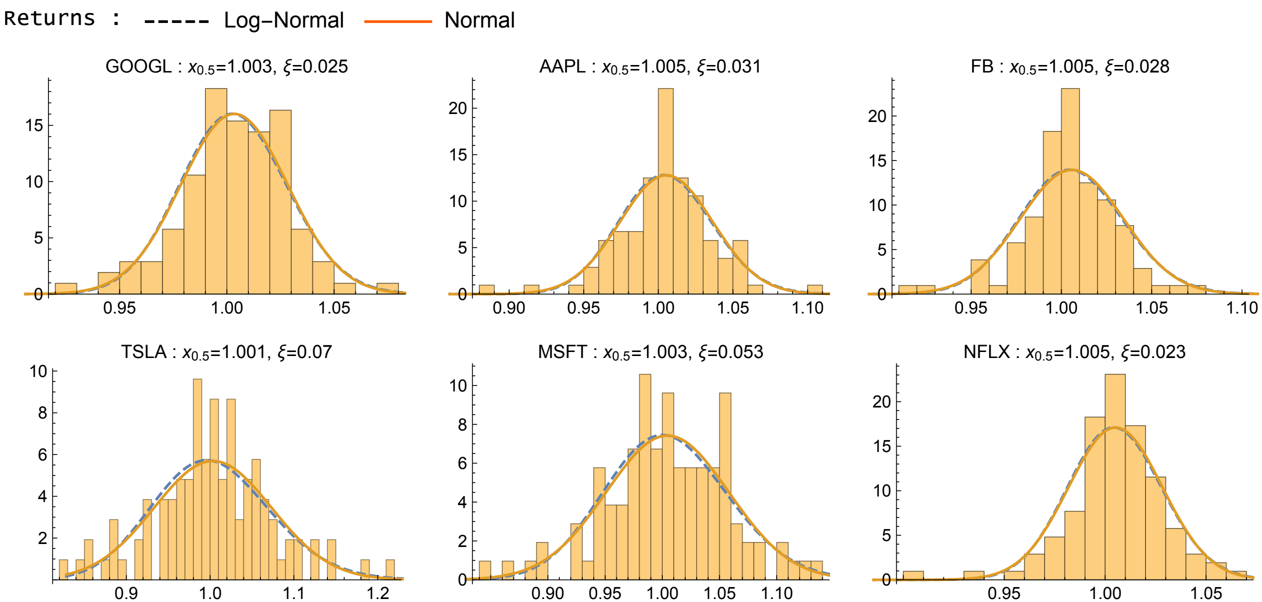
$$\lambda_t = \log[P_0] + \sum_{j=1}^{t} E[\log S_j], \quad \xi_t^2 = \sum_{j=1}^{t} V[\log S_j]$$

Typically, $t = 0$ is the present and we assume that future returns $S_1, S_2, \ldots$ 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 $j = 1, 2\ldots$ 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.



Furthermore, one may use the results in this example to produce a price forecast; in the case of Netflix we get:

Netflix stock price forecast
Price, $
240
220
200
180
160
140
120
100
Apr
Jul
Oct
Jan
2017
| 2018

□

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:

$$X \sim \mathrm{Ber}(p)$$

if its probability mass function is given by

$$p_X(x) = \begin{cases} p, & x = 1 \\ q = 1 - p, & x = 0 \end{cases}$$

and

$$\mathrm{E}(X) = p$$
$$\mathrm{V}(X) = pq$$

Example 21. — 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 $\mathrm{P}(A)$.

Solution: Sense the probability of success here is $p = \mathrm{P}(A)$, we have:

$$\mathrm{E}(X) = \mathrm{P}(A)$$
$$\mathrm{V}(X) = \mathrm{P}(A)(1 - \mathrm{P}(A))$$

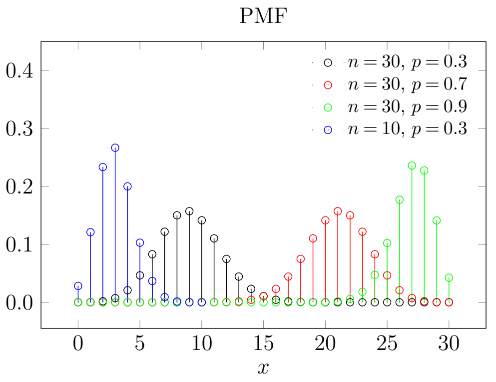
□

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**.



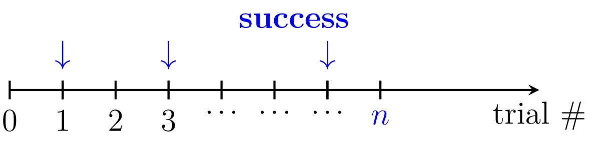
$$X \sim \mathrm{Bin}(n, p)$$

The PMF $p_X(x) = \mathrm{P}(X = x)$ is

$$\boxed{p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}} \tag{4.11}$$

and

$$\mathrm{E}(X) = np$$
$$\mathrm{V}(X) = npq$$



Notice that the CDF $F_X(x) = \mathrm{P}(X \leq x)$ does not simplify:

$$\boxed{F_X(x) = \sum_{i=1}^{x} \binom{n}{i} p^i (1-p)^{n-i}} \tag{4.12}$$

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 22. 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	T
coin 2	H	H	H	H	T	T	T	T	H	H	H	H	T	T	T	T
coin 3	H	H	T	T	H	H	T	T	H	H	T	T	H	H	T	T
coin 4	H	T	H	T	H	T	H	T	H	T	H	T	H	T	H	T
ΣH	4	3	3	2	3	2	2	1	3	2	2	1	2	1	1	0

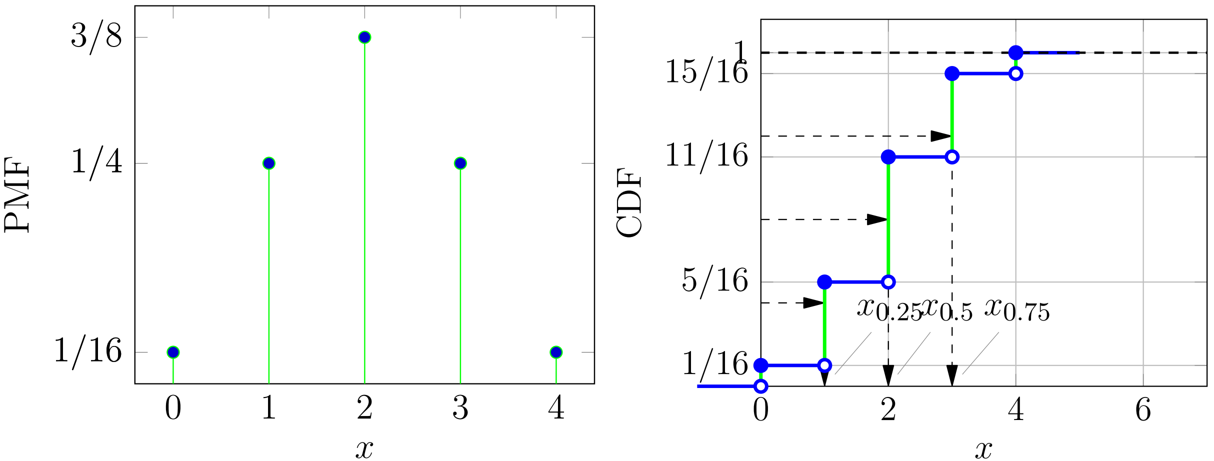
a) Find the CDF and PMF of X and draw sketches of each one.
b) Determine the median, upper quartile and lower quartile and show them graphically in one of the sketches of part a)
c) Determine $\mathrm{P}(0 < X \leq 3 \mid X \leq 2)$
d) BONUS: suppose two players play this game, and the one with the largest 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.

Solution:

a) Find the CDF and PMF of X and draw sketches of each one.

$X \sim \text{Bin}(n=4, p=1/2)$
$\rightarrow p_X(x) = \binom{4}{x} 0.5^x (1-0.5)^{4-x} = \binom{4}{x} 0.5^4 = \binom{4}{x}/16$, 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 $Y_1, Y_2, \ldots, Y_n$ 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	0
coin 2, Y_2	1	1	1	1	0	0	0	0	1	1	1	1	0	0	0	0
coin 3, Y_3	1	1	0	0	1	1	0	0	1	1	0	0	1	1	0	0
coin 4, Y_4	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0
$X = \sum_{i=1}^{n} Y_i$	4	3	3	2	3	2	2	1	3	2	2	1	2	1	1	0

Fact 4.9 — Two important corollaries. :

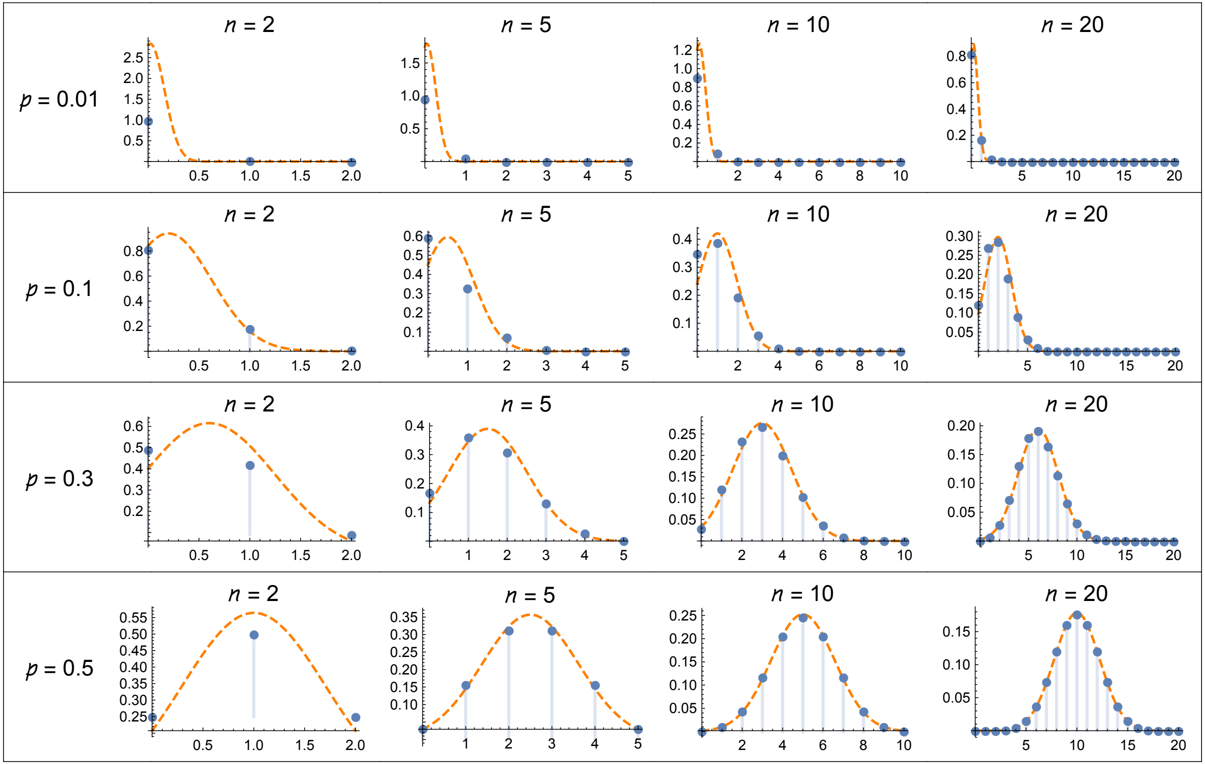
1. The **normal approximation**

$$X \sim N(\mu = np, \sigma^2 = npq)$$

   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:

   if $X_1 \sim \mathrm{Bin}(n_1, p)$ and $X_2 \sim \mathrm{Bin}(n_2, p) \to X_1 + X_2 \sim \mathrm{Bin}(n_1 + n_2, p)$

The figure below shows the agreement of the normal approximation with the $\mathrm{Bin}(n,p)$ rv for different values of parameters (n,p).



n = 2
n = 5
n = 10
n = 20
p = 0.01
p = 0.1
p = 0.3
p = 0.5

Continuity Correction

We saw that the **normal approximation to a binomial random variable** $X \sim \mathrm{Bin}(n,p)$ is:

$$X \sim N(\mu = np, \sigma^2 = npq)$$

is accurate when n it is large enough. Then, one can use

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right). \tag{4.13}$$

However, the error can be substantial if n is not very large. One way to improve the approximation is to use the *continuity correction*:

$$P(a \leq X \leq b) \approx \Phi\left(\frac{b\mathbf{+0.5}-\mu}{\sigma}\right) - \Phi\left(\frac{a\mathbf{-0.5}-\mu}{\sigma}\right). \tag{4.14}$$

Analogous continuity corrections apply to the Poisson distribution, in which case $\mu = \theta, \sigma^2 = \theta$.

Example 23. 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,

$$X \sim \mathrm{Bin}\left(5, \frac{1}{6}\right)$$

Therefore,

$$\mathrm{P}(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$$
$$\therefore \mathrm{P}(X = 3) = \binom{5}{3} \left(\frac{1}{6}\right)^3 \left(\frac{5}{6}\right)^3$$

$\square$

Example 24. 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,

$$\begin{aligned}
\mathrm{P}(X=-1) &= \mathrm{P}(Y=0) \\
&= \binom{3}{0}\left(\frac{1}{6}\right)^0\left(\frac{5}{6}\right)^3 \\
&= \frac{125}{216} \\
\mathrm{P}(X=1) &= \mathrm{P}(Y=1) \\
&= \binom{3}{1}\left(\frac{1}{6}\right)^1\left(\frac{5}{6}\right)^2 \\
&= \frac{75}{216} \\
\mathrm{P}(X=2) &= \mathrm{P}(Y=2) \\
&= \binom{3}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^1 \\
&= \frac{15}{216} \\
\mathrm{P}(X=3) &= \mathrm{P}(Y=3) \\
&= \binom{3}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^0 \\
&= \frac{1}{216}
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{E}[X] &= (-1)\left(\frac{125}{216}\right) + (1)\left(\frac{75}{216}\right) + (2)\left(\frac{15}{216}\right) + (3)\left(\frac{1}{216}\right) \\
&= -\frac{17}{216}
\end{aligned}$$

Therefore, as the expected value of the winnings is less than 0, the game is not fair towards the player. $\square$

Example 25. 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

$$\begin{aligned}
\mathrm{P}(X=2) &= \binom{4}{2} 0.2^2 0.8^{4-2} = 0.1536\,, \\
\mathrm{P}(X<2) &= \mathrm{P}(X=0) + \mathrm{P}(X=1) = \binom{4}{0} 0.2^0 0.8^4 + \binom{4}{1} 0.2^1 0.8^3 = 0.8192\,, \\
\mathrm{P}(X \geq 2) &= \mathrm{P}(X=2) + \mathrm{P}(X=3) + \mathrm{P}(X=4) \\
&= 0.1536 + \binom{4}{3} 0.2^3 0.8^1 + \binom{4}{0} 0.2^4 0.8^0 = 0.1808\,.
\end{aligned}$$

□

Example 26. — 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

$$r = \text{P}(Y > 130) = 0.033$$

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:

$$\begin{aligned}
\mathrm{P}(D^c) &= \sum_{x=0}^{n} \mathrm{P}(D^c | X = x)\mathrm{P}(X = x) \\
&= \sum_{x=0}^{n} (1-r)^x \mathrm{P}(X = x) \\
&= \sum_{x=0}^{n} (1-r)^x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} ((1-r)p)^x (1-p)^{n-x} \\
&= (1-rp)^n = 0.944
\end{aligned}$$

and the desired probability is 1-0.944=0.0556. The last step follows from the binomial theorem (4.7) with $a = (1-r)p, b = 1-p$.

STOP! 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

$$Z \sim \mathrm{Bin}(n, pr)$$

and the desired probability is also $\mathrm{P}(Z > 0) = 1 - (1 - rp)^n = 0.0556$.

c) solved in class
d) solved in class
e) solved in class

$\square$

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 ith outcome, $i = 1, 2 \ldots k$

The (joint) distribution of the random vector $\mathbf{N} = \{N_1, ..., N_k\}$ is

$$\mathrm{P}(n_1, ..., n_k) = \frac{n!}{n_1! \cdots n_k!} \prod_{i=1}^{k} p_i^{n_i}$$

and:

$$\begin{aligned}\mathrm{E}(N_i) &= np_i \\ \mathrm{V}(N_i) &= np_i(1-p_i) \\ \mathrm{Cov}\left(N_i, N_j\right) &= -np_ip_j\end{aligned}$$

Fact 4.10 — Marginal distributions. $N_i \sim \mathrm{Bin}(n, p_i)$

Example 27. 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 $\mathrm{Cov}(N_1, N_2)$.

Example 28. 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 $\text{Cov}(N_1, N_2)$.

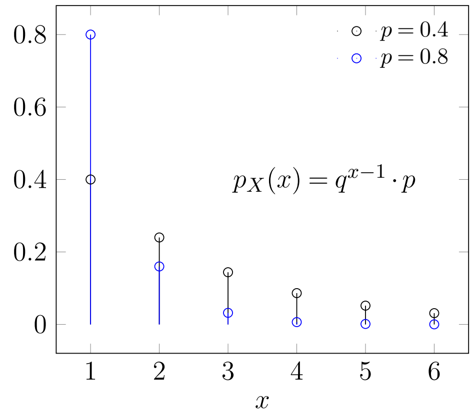
Solution: (a) According to Fact 4.10, $N_4 \sim \text{Bin}(16, 0.02)$. Thus, $P(N_4 = 1) = 16(0.02)(0.98)^{15} = 0.2363$.
(b) $P(N_1 = 10, N_2 = 4, N_3 = 2, N_4 = 0) = \frac{16!}{10!4!2!} 0.6^{10} 0.3^4 0.08^2 = 0.0377$.
c) Expect them to be negatively related: The larger N_1 is, the smaller N_2 is expected 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

$$X \sim \mathrm{Geo}(p)$$



The PMF $p_X(x) = \mathrm{P}(X = x)$ is

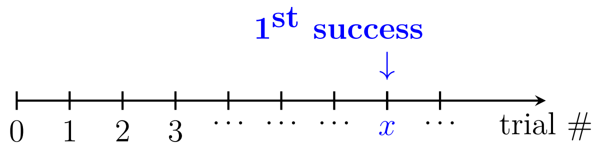
$$\boxed{p_X(x) = q^{x-1} p}$$

with $q = 1 - p$, and

$$\mathrm{E}(X) = 1/p$$
$$\mathrm{V}(X) = q/p^2$$

The CDF is given by:

$$\boxed{\mathrm{P}(X \leq x) = 1 - q^x}$$



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 $\mathrm{E}(X)$ is also called the return period T, and therefore

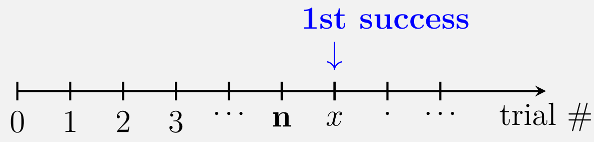
$$p = \frac{1}{T}$$

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 Geo(p) and $\mathrm{Bin}(n,p)$ distributions. If

$$Y \sim \mathrm{Bin}(n,p), \quad \#\text{ of successes in } n \text{ trails}$$
$$X \sim \mathrm{Geo}(p), \quad \#\text{ of trials until first success}$$

then, from the picture:



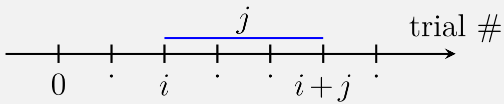
$$\begin{aligned} \mathrm{P}(X > n) &= \mathrm{P}(Y = 0) \\ &= \binom{n}{0} p^0 q^{n-0} \\ &= q^n \end{aligned}$$

Therefore, the CDF of the geometric distribution $F_X(x)$ is given by:

$$\boxed{\mathrm{P}(X \le x) = 1 - q^x} \tag{4.15}$$

Fact 4.12 — Memoryless Property of the Geometric Distribution. Let i, j be positive integers. $X \sim \text{Geo}(p)$. Then

$$P(X > i + j | X > i) \quad = \quad P(X > j)$$



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 Geo(p) distribution.

Proof.

$$\begin{aligned} P(X > i+j | X > i) &= \frac{P(\{X > i+j\} \cap \{X > i\})}{P(X > i)} \\ &= \frac{P(X > i+j)}{P(X > i)} \\ &= \frac{q^{(i+j)}}{q^i} = q^j \\ &= P(X > j) \end{aligned}$$

■

Example 29. — 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,

$$X \sim \text{Geo}(1/20) \rightarrow \text{P}(X \leq 20) = 1 - (1 - 1/20)^{20} = 0.6425$$

c) the probability that the first exceedance of the design wave height will occur after the third year?

$\text{P}(X > 3) = 1 - \text{P}(X \leq 3) = 0.95^3$

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?

$\text{P}(X = 5 | X > 3) = 0.048$

$\square$

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

$$\mathrm{P}(X \leq T) \approx 1 - e^{-1} = 0.6321$$

in the case of events with long return period T, by virtue of the identity

$$\lim_{T \to \infty} (1 - \frac{1}{T})^T = e^{-1}$$

Example 30. — * 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:

$$\begin{aligned} r_1 &= 0.5 \qquad \text{(houses)} \\ r_2 &= 0.2 \qquad \text{(buildings)} \end{aligned}$$

Probability of damage in one year:

$$\begin{aligned} p_1 &= r_1 p = 0.0166667 \qquad \text{(houses)} \\ p_2 &= r_2 p = 0.0066667 \qquad \text{(buildings)} \end{aligned}$$

Let X_1 and X_2 be the number of years between earthquakes that produced damage to a given house and building, respectively. Then,

$$X_i \sim Geo(p_i) \to \mathrm{P}(X_i \leq 100) = 1 - (1 - p_i)^{100}$$

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,

$$Y_i \sim Bin(100, p_i) \rightarrow \mathrm{P}(Y_i \geq 2) = 1 - \mathrm{P}(Y_i < 2)$$

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,

$$\begin{aligned} W_i &= \sum_{j=1}^{n_i} Y_{i,j} \\ &\sim N(n_i \mathrm{E}(Y_i), n_i \mathrm{V}(Y_i)) \qquad \text{by the CLT} \end{aligned}$$

where

$$\mathrm{E}(Y_i) = 100 p_i$$
$$\mathrm{V}(Y_i) = 100 p_i q_i$$

Finally, $\mathrm{P}(W_2 > W_1) = \mathrm{P}(W_2 - W_1 > 0)$, which gives ≈ 0 in this case. We used

$$W_2 - W_1 \sim N\left(\underbrace{\mathrm{E}(W_2) - \mathrm{E}(W_1)}_{=-33.3}, \underbrace{\mathrm{V}(W_2) + \mathrm{V}(W_1)}_{=1284.7}\right)$$

(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

$$U \sim N\left(\underbrace{\mathrm{E}(U)}_{1{,}433}, \underbrace{\mathrm{V}(U)}_{253{,}283}\right)$$

in thousands of dollars. The final answer is 0.80539. $\square$

Example 31. Alice eats cookies one after another until she finds and 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.

$$
\begin{aligned}
\mathrm{P}(X>3) &= \sum_{k=4}^{\infty} \mathrm{P}(X=k) \\
&= \sum_{k=4}^{\infty} \left(1-\frac{1}{10}\right)^{k-1} \left(\frac{1}{10}\right) \\
&= \left(1-\frac{1}{10}\right)^{3} \sum_{j=1}^{\infty} \left(1-\frac{1}{10}\right)^{j-1} \left(\frac{1}{10}\right) \\
&= \left(\frac{9}{10}\right)^{3} \left(\frac{1}{10}\right) \left(\frac{1}{1-\frac{9}{10}}\right) \\
&= \left(\frac{9}{10}\right)^{3}
\end{aligned}
$$

2.

$$
\begin{aligned}
\mathrm{P}(X \geq 13 | X > 5) &= \mathrm{P}(X > 12 | X > 5) \\
&= \frac{\mathrm{P}(X > 12 \cap X > 5)}{\mathrm{P}(X > 5)} \\
&= \frac{\mathrm{P}(X > 12)}{\mathrm{P}(X > 5)} \\
&= \frac{\left(\frac{9}{10}\right)^{12}}{\left(\frac{9}{10}\right)^{5}} \\
&= \left(\frac{9}{10}\right)^{7} \\
&= \mathrm{P}(X > 7)
\end{aligned}
$$

Therefore, the fact that Alice has already eaten 5 cookies does not affect the probability of her eating at least 8 more cookies.

$\square$

Example 32. 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: $X \sim Geo(0.2)$
$E(X) = \frac{1}{0.2} = 5$ $V(X) = \frac{1-0.2}{0.2^2} = 20$

$\square$

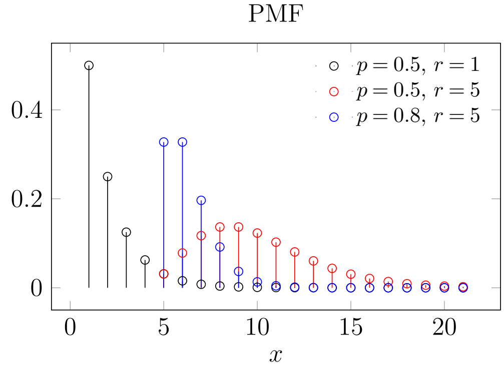
4.4.4 Negative Binomial Random Variable

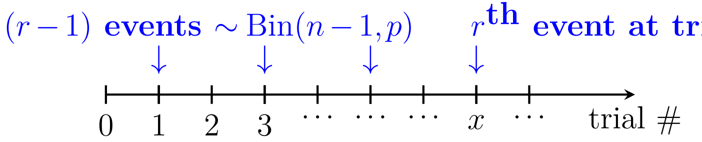
Here X is **the number of Bernoulli trials required until the r-th success occurs**.

$$X \sim \mathrm{NB}(r,p)$$

The PMF of X is

$$\mathrm{P}(X=n) = \binom{n-1}{r-1} p^r (1-p)^{n-r}$$
$$\mathrm{E}(X) = r/p$$
$$\mathrm{V}(X) = rq/p^2$$





$$\begin{aligned}\to \mathrm{P}(X = n) &= \binom{n-1}{r-1} p^{r-1}(1-p)^{n-r} \cdot p \\ &= \binom{n-1}{r-1} p^{r}(1-p)^{n-r}\end{aligned}$$

Example 33. Find the expected value and variance of the number of time 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,

$$X \sim \mathrm{NB}\left(4, \frac{1}{6}\right)$$

Therefore,

$$
\begin{aligned}
\mathrm{E}[X] &= \frac{r}{p} \\
&= \frac{4}{\frac{1}{6}} \\
&= 24 \\
\mathrm{V}(X) &= \frac{r(1-p)}{p^2} \\
&= \frac{4\left(1-\frac{1}{6}\right)}{\left(\frac{1}{6}\right)^2} \\
&= 120
\end{aligned}
$$

□

Fact 4.13 Let

$$X_i \sim \mathrm{Geo}(p)$$

be independent random variables, for $i \in \mathbb{N}$. Then,

$$\sum_{i=1}^{n} X_i \sim \mathrm{NB}(n, p)$$

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 trails:

$$X \sim \mathrm{HG}(n, N, k)$$

The probability distribution of X is

$$\begin{aligned} \mathrm{P}(X=i) &= \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}} \\ \mathrm{E}[X] &= \frac{nk}{N} \\ \mathrm{V}(X) &= n\frac{k}{N}\left(1-\frac{k}{N}\right)\left(\frac{N-n}{N-1}\right) \end{aligned}$$

Example 34. 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.

$$\mathrm{P}(X = i) = \frac{\binom{k}{i}\binom{N-k}{n-i}}{\binom{N}{n}} = 0.32513$$

□

Example 35. 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) $\mathrm{Cov}(N_1, N_2)$.
e) $\mathrm{Cov}(N_1, N_2 + N_3)$.

Solution: office hours :)

□

4.5 Poisson Random Variables

A discrete random variable X, taking one of the values $0, 1, 2, \ldots$, is said to be a **Poisson random variable** with parameter $\theta > 0$ if:

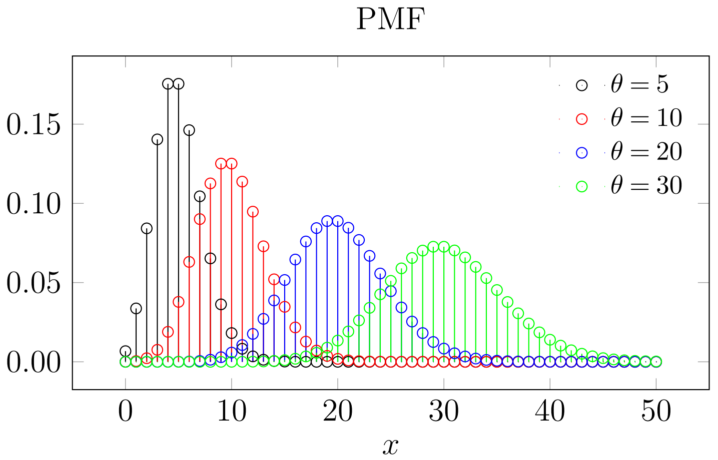
$$X \sim \text{Poi}(\theta)$$

The PMF of X is

$$p_X(x) = \frac{e^{-\theta}\theta^x}{x!}$$

and

$$\text{E}(X) = \theta$$
$$\text{V}(X) = \theta$$



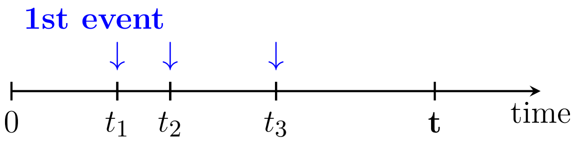
In a Poisson Process events occur at a given rate

$$\lambda = \text{number of events per unit of time}$$

(or distance t, or area, or volume, or population, etc.). Then,

$$\begin{aligned} X &= \# \text{ of events in } (0, t) \\ &\sim \text{Poi}(\lambda t), \end{aligned}$$

Note that the Poisson parameter $\theta = \lambda t$ has no units.



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,

$$\lim_{n\to\infty} \binom{n}{x} p^x (1-p)^{n-x} = \frac{e^{-x}\theta^x}{x!}$$

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 it is large enough (by the CLT).
2. The **sum of two Poisson random variables** with parameters θ_1 and θ_2 is also Poisson:

   if $X_1 \sim \mathrm{Poi}(\theta_1)$ and $X_2 \sim \mathrm{Poi}(\theta_2) \to X_1 + X_2 \sim \mathrm{Poi}(\theta_1+\theta_2)$

Proof. Express the binomial probability in terms of the parameter θ:

$$\begin{aligned}
\lim_{n\to\infty}\binom{n}{x}p^x(1-p)^{n-x} &= \lim_{n\to\infty}\binom{n}{x}\left(\frac{\theta}{n}\right)^x\left(1-\frac{\theta}{n}\right)^{n-x} \\
&= \lim_{n\to\infty}\frac{n!}{x!(n-x)!}\theta^x\left(\frac{1}{n}\right)^x\left(1-\frac{\theta}{n}\right)^{-x}\left(1-\frac{\theta}{n}\right)^n \\
&= \frac{\theta^x}{x!}\lim_{n\to\infty}\frac{n!}{(n-x)!}\frac{1}{(n-\theta)^x}\left(1-\frac{\theta}{n}\right)^n \\
&= \frac{\theta^x}{x!}\lim_{n\to\infty}\frac{n!}{(n-x)!}\frac{1}{(n-\theta)^x}\left(1-\frac{\theta}{n}\right)^n
\end{aligned}$$

From calculus, we know that

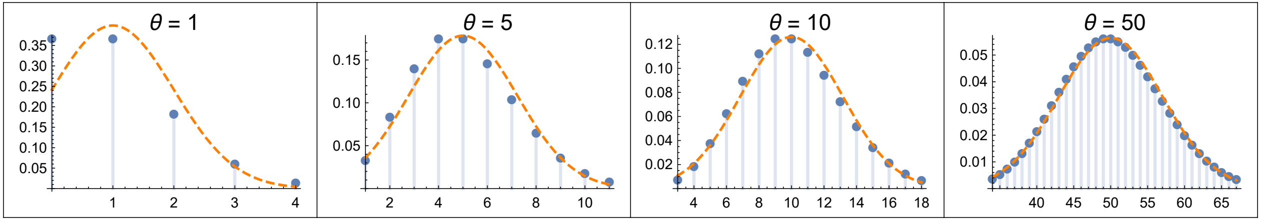
$$\lim_{n\to\infty}\left(1-\frac{\theta}{n}\right)^n = e^{-\theta}$$

and

$$\lim_{n\to\infty}\frac{n!}{(n-x)!}\frac{1}{(n-\theta)^x} = \lim_{n\to\infty}\frac{n(n-1)\dots(n-x+1)}{(n-\theta)(n-\theta)\dots(n-\theta)} = 1$$

■

The normal approximation as a function of θ:



Example 36. 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 $t = 2$s, so the Poisson parameter

$$\theta = \lambda t = 3.2$$

Therefore,

$$X \sim \text{Poi}(3.2)$$

and,

$$\begin{aligned} \mathrm{P}(X \leq 2) &= \mathrm{P}(X=0) + \mathrm{P}(X=1) + \mathrm{P}(X=2) \\ &= \frac{e^{-3.2}\theta^0}{0!} + \frac{e^{-3.2}\theta^1}{1!} + \frac{e^{-3.2}\theta^2}{2!} \\ &\approx 0.3799 \end{aligned}$$

$\square$

Example 37. 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)

$$\begin{aligned} P(Accepted) &= P(X \leq 15) \\ &= \sum_{x=0}^{15} \frac{\theta^x e^{-\theta}}{x!} \\ &= \sum_{x=0}^{15} \frac{(103.68)^x e^{-103.68}}{x!} \\ &= 1.44 \times 10^{-27} \end{aligned}$$

(b) Assume λ is the pixel failure rate required:

$$\begin{aligned}
P(Accepted) &= P(X \leq 15) \\
&= \sum_{x=0}^{15} \frac{(\lambda t)^x e^{-\lambda t}}{x!} \\
&= \sum_{x=0}^{15} \frac{(\lambda \times 4000 \times 2000)^x e^{-\lambda \times 4000 \times 2000}}{x!} \\
&= 0.9 \\
\lambda &= 1.39 \times 10^{-6}
\end{aligned}$$

$\square$

Example 38. 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,

$$\begin{aligned} \text{P}(X \geq 3) &= 1 - (\text{P}(X = 0) + \text{P}(X = 1) + \text{P}(X = 2)) \\ &= 1 - \left(\frac{e^{-4}4^0}{0!} + \frac{e^{-4}4^1}{1!} + \frac{e^{-4}4^2}{21} \right) \\ &= 1 - 13e^{-4} \end{aligned}$$

$\square$

Example 39. 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. $\square$

Example 40. 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

$$\theta = np = 20 \qquad \text{(avg faulty solutions in 1 book)}.$$

Hence,

$$P(X \geq 15) = \sum_{k=15}^{1000} \frac{\theta^k e^{-\theta}}{k!} = 0.89513 = 89.5\%.$$

$\square$

Example 41. — **** 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:

$$\begin{aligned}
\lambda &= \frac{1}{10} \times 5\% \\
&= 0.005 \\
t &= 50 \\
\theta &= \lambda t \\
&= 0.005 \times 50 \\
&= 0.25 \\
Y &\sim Pois(0.25)
\end{aligned}$$

(b)

$$\begin{aligned}
P(Y \geq 1) &= 1 - P(Y = 0) \\
&= 1 - e^{-\theta} \\
&= 1 - e^{-0.25} \\
&= 0.2212
\end{aligned}$$

(c) $Z \sim Bin(100000, 0.2212)$
Use normal approximation:

$$
\begin{aligned}
E(Z) &= np = 100000 \times 0.2212 = 22120 \\
V(Z) &= npq = 100000 \times 0.2212 \times (1 - 0.2212) \\
&= 17727.056 \\
\sigma_Z &= \sqrt{V(Z)} = 131.252 \\
P(Z < 22000) &= \Phi(\frac{22000 - E(Z)}{\sigma_Z}) \\
&= \Phi(-0.91) \\
&= 0.1814
\end{aligned}
$$

□

4.6 Exponential Random Variable

A random variable X is said to be a exponential random variable over the interval $(0, \infty)$,

$$X \sim \text{Expo}(\lambda)$$

The PDF of X is

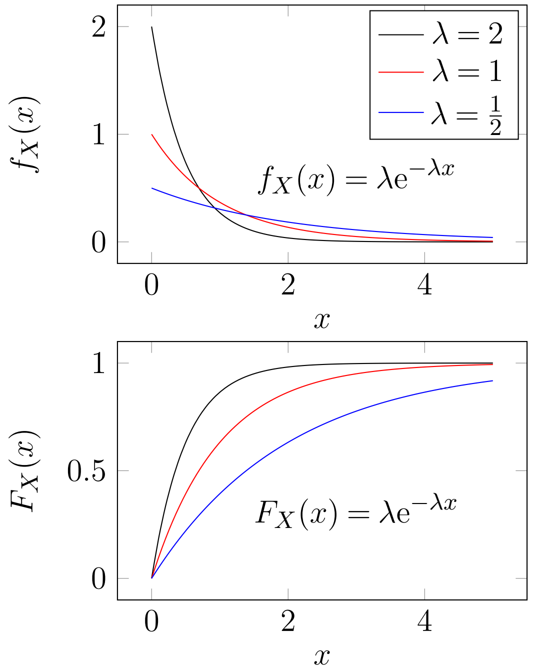
$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & ; \quad x \geq 0 \\ 0 & ; \quad x < 0 \end{cases}$$

The CDF:

$$\text{F}_X(x) = \begin{cases} 1 - e^{-\lambda x} & ; \quad x \geq 0 \\ 0 & ; \quad x < 0 \end{cases}$$

and

$$\text{E}(X) = 1/\lambda$$
$$\text{V}(X) = 1/\lambda^2$$



Example 42. 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,

$$X \sim \text{Expo}(\lambda = 1/5)$$

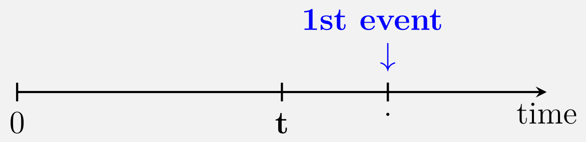
Therefore,

$$\text{P}(X \leq 10) = 1 - e^{-10/5} = 0.864665$$

$\square$

Fact 4.15 — Connection between the $\mathrm{Expo}(\lambda)$ and $\mathrm{Poi}(\theta = \lambda t)$ distributions. If

$$Y \sim \mathrm{Poi}(\lambda t), \quad \# \text{ of events in } (0, t)$$
$$X = \text{time of first event}$$



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,

$$\boxed{X \sim \mathrm{Expo}(\lambda)}$$

Fact 4.16 — The time between arrivals. of a Poisson process, $\mathrm{Poi}(\lambda t)$ 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 a exponential rv is $Y =$ **the remaining time s until the next event**, which also has the $\text{Expo}(\lambda)$ distribution. Also,

$$\mathrm{E}(X|X > t) = E(Y) = \mathrm{E}(X) = 1/\lambda$$

Proof.

$$\begin{aligned} P(X > t + s | X > t) &= \frac{P(\{X > t + s\} \cap \{X > t\})}{P(X > t)} = \frac{P(X > t + s)}{P(X > t)} \\ &= \frac{e^{-\lambda(t+s)}}{e^{-\lambda t}} = e^{-\lambda s} = P(X > s) \end{aligned}$$

■

STOP! **The** $\mathrm{Expo}(\lambda)$ **and** $\mathrm{Geo}(p)$ **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 $0 \le x \le 1$ we have:

$$P(X > t + s | X > t) = \frac{P(X > t + s)}{P(X > t)} = \frac{1 - (t + s)}{1 - t} \tag{4.16}$$

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:

$$\mathrm{E}(X | X > t) = \frac{1 - t}{2} \neq \mathrm{E}(X)$$

Example 43. A battery has a lifespan that is exponentially distributed with rate parameter $1/3000$ per hour.

a) Find the probability that a random battery has a lifespan of more than 2500 hours.
b) 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.

$$X \sim \text{Expo}(\lambda = 1/3000)$$

a) Find the probability that a random battery has a lifespan of more than 2500 hours.

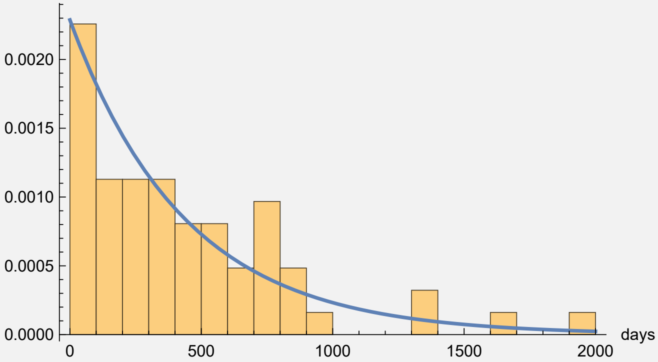
$$\mathrm{P}(X \geq 2500) = 1 - F_X(2500) = e^{-2500/3000} = 0.565$$

b) 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 a exponential rv is **the remaining time s until the first success**, which also has the Expo(λ) distribution. So,

$$\begin{aligned}\mathrm{P}(X \geq 2500 | X > 2000) = \mathrm{P}(X > 500) &= 1 - F_X(500) \\ &= e^{-500/3000} = 0.846\end{aligned}$$

$\square$

Example 44. 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:

a) if the last earthquake occurred four years ago, what is the expected time until the next earthquake?
b) what is the probability of having 2 earthquakes in the next year?
c) if the last earthquake occurred four years ago, what is the probability of having 2 earthquakes in the next year?

Solution: in class $\square$

4.7 Gamma (Erlang) distributions are sums of exponentials

If $X_1, X_2, \ldots, X_n$ are distributed as Expo(λ), independently, and

$$X = X_1 + \cdots + X_n$$

then $X \sim \text{Gamma}(n, \lambda)$. In general, a random variable X is said to have a Gamma(α, β) 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 45. In Example 44:

a) what is the probability of having 2 earthquakes in the next year? (using Gamma)
b) what is the probability that the third earthquake happens after 10 years from now?

Solution: in class $\square$

4.8 The beta distribution: finite interval sample space

A random variable X is said to have a beta distributionbeta 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 $X \sim Beta(\alpha, \beta)$, and

$$\begin{aligned} E(X) &= \alpha/(\alpha+\beta) \\ V(X) &= \alpha+\beta+1 \end{aligned}$$

More on Wikipedia

4.9 The Bivariate Normal Distribution

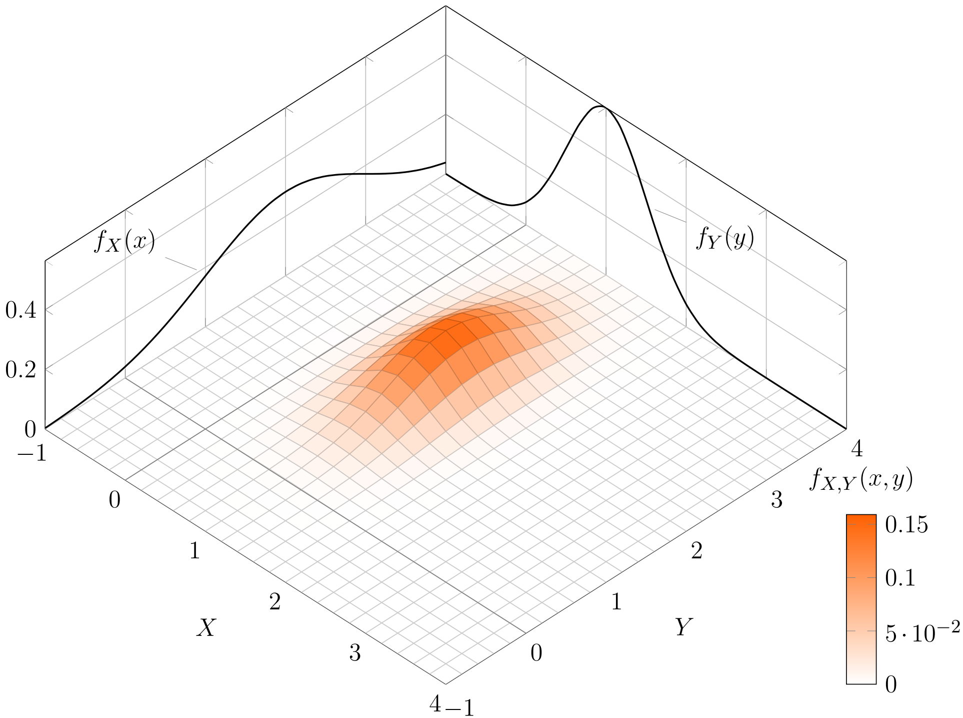
Let X_1 and X_2 have the bivariate normal joint distribution. Then, the joint pdf of (X_1, X_2) is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{2\pi \left|\mathbf{V}\right|^{1/2}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}-\mu\right)^T \mathbf{V}^{-1}\left(\mathbf{x}-\mu\right)\right\}$$

where $\mathbf{X}^T = (X_1, X_2)$, $\mu^T = (\mu_1, \mu_2)$ and $\mathbf{V}$ is a full rank variance-covariance matrix, i.e.,

$$\mathbf{V}_{ij} = \mathrm{Cov}\left(X_i, X_j\right)$$

$\mathbf{V}^{-1}$ is the inverse of $\mathbf{V}$, $|\mathbf{V}|$ is the determinant of $\mathbf{V}$.


$f_X(x)$
$f_Y(y)$
0.4
0.2
0
−1
0
1
2
3
4
X
−1
0
1
2
3
4
Y
$f_{X,Y}(x,y)$
0.15
0.1
$5 \cdot 10^{-2}$
0

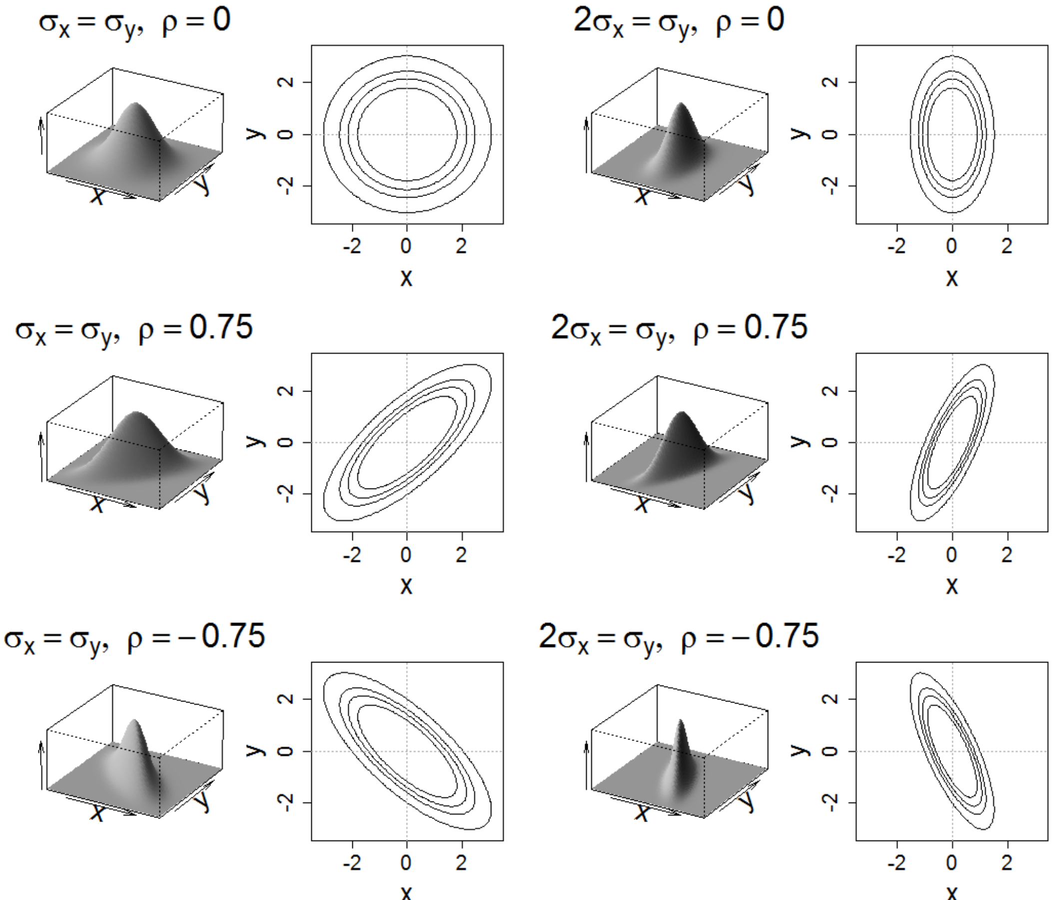
Fact 4.19 — Marginal distributions are normal. If (X_1, X_2) have a bivariate normal distribution, then the marginal distribution of X_2 is also normal with mean μ_2 and variance σ_2^2.

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

$$\mathrm{E}(X_2|X_1 = x_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1). \tag{4.17}$$

$$\mathrm{V}(X_2|X_1 = x_1) = (1-\rho^2)\sigma_2^2. \tag{4.18}$$

Fact 4.21 If (X_1, X_2) have a bivariate normal distribution with $\rho = 0$, X_1 and X_2 are independent.


σx = σy, ρ = 0
2σx = σy, ρ = 0
σx = σy, ρ = 0.75
2σx = σy, ρ = 0.75
σx = σy, ρ = −0.75
2σx = σy, ρ = −0.75
x
y
-2
0
2

Multi-variate normal distribution

The random vector $\mathbf{X}^T = (\mathbf{X}_1, ..., \mathbf{X}_n)$ has the multivariate normal distribution if its joint density function is given by

$$f(\mathbf{X}) = \frac{1}{(2\pi)^{n/2} |\mathbf{V}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{V}^{-1} (\mathbf{x}-\mu)\right\}$$

where $\mu^T = (\mu_1, ..., \mu_n)$ and $\mathbf{V}$ is a full rank variance-covariance matrix. Note that for $n = 2$ we get the density of the bivariate normal distribution.