Key distributions

Last modified: 2026-11-13 06:54:40 (PST)

> Remark
>
> *Remark.* Some distributions are typically used for outcome models (Table 1); other distributions are typically used for test statistics (Table 2).

Table 1: Distributions typically used for outcome models

Distribution	Uses
Bernoulli	Binary outcomes
Binomial	Sums of Bernoulli outcomes
Poisson	Unbounded count outcomes
Geometric	Counts of non-events before an event occurs
Negative binomial	Mixtures of Poisson distributions, counts of non-events until a given number of events occurs
Normal (Gaussian)	Continuous outcomes without a more specific distribution

Distribution	Uses
Multivariate normal	Vectors of continuous outcomes
Mixtures	Outcomes from a population of unobserved subpopulations
Exponential	Time to event outcomes
Gamma	Time to event outcomes
Weibull	Time to event outcomes
Log-normal	Time to event outcomes

Table 2: Distributions typically used for test statistics

Distribution	Uses
χ^2	Regression comparisons (asymptotic), contingency table independence tests, goodness-of-fit tests
F	Gaussian model comparisons (exact)
Z (standard normal)	Proportions, means, regression coefficients (asymptotic)
t (Student's)	Means, regression coefficients in Gaussian outcome models (exact)

The Bernoulli distribution

Example 0.1 (A coin flip)**.** The Bernoulli distribution describes a binary outcome. The indicator of heads on one flip of a fair coin is $\text{Ber}(1/2)$. By the expectation of the Bernoulli distribution, $X \sim \text{Ber}(\pi)$ has mean π, and by the variance of a Bernoulli random variable, it has variance $\pi(1-\pi)$.

The binomial distribution

Definition 0.1 (Binomial distribution)**.** A discrete random variable X has the **binomial distribution** with number of trials $n \in \{1, 2, \dots\}$ and event probability $\pi \in [0, 1]$, written $X \sim \text{Bin}(n, \pi)$ [1], if:

$$\mathbb{P}(X = x) \stackrel{\text{def}}{=} \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad x \in \{0, 1, \dots, n\} \tag{1}$$

where $\binom{n}{x} = \frac{n!}{x!\,(n-x)!}$ is the number of ways to choose x of the n trials.

Remark

Remark. Casella and Berger [2, Sec. 3.2] defines the binomial distribution the same way. With $n = 1$, Equation 1 is the Bernoulli PMF: $\text{Bin}(1, \pi)$ and $\text{Ber}(\pi)$ are the same distribution.

Example 0.1 (Two heads in three flips)**.** Let $X \sim \text{Bin}(3, 1/2)$ be the number of heads in three flips of a fair coin. By Equation 1:

$$\begin{aligned}
\mathbb{P}(X = 2) &= \binom{3}{2} \left(\frac{1}{2}\right)^2 \left(1 - \frac{1}{2}\right)^{3-2} && \text{(binomial PMF with } n = 3,\ \pi = 1/2,\ x = 2) \\
&= \frac{3!}{2!\,1!} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^1 && \text{(definition of } \tbinom{3}{2};\ 1 - 1/2 = 1/2) \\
&= 3 \cdot \frac{1}{4} \cdot \frac{1}{2} && \text{(evaluate the factorials and powers)} \\
&= \frac{3}{8} && \text{(multiply)}
\end{aligned}$$

The three outcomes with two heads are HHT, HTH, and THH, each with probability $1/8$.

Example 0.2 (Draws without replacement are not binomial)**.** An urn holds two red balls and two blue balls. Draw two balls without replacement, and let X be the number of red balls drawn. Each draw is red with probability $1/2$, but $X \nsim \text{Bin}(2, 1/2)$, because the draws are not independent:

$$\begin{aligned}
\mathbb{P}(X = 2) &= \mathbb{P}(\text{first red}) \cdot \mathbb{P}(\text{second red} \mid \text{first red}) && \text{(multiplication rule)} \\
&= \frac{2}{4} \cdot \frac{1}{3} && \text{(one red ball is left among three)} \\
&= \frac{1}{6} && \text{(multiply)}
\end{aligned}$$

whereas, by Equation 1, a $\text{Bin}(2, 1/2)$ random variable has

$$\begin{aligned}
\mathbb{P}(X = 2) &= \binom{2}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^0 && \text{(binomial PMF with } n = 2,\ \pi = 1/2,\ x = 2) \\
&= \frac{1}{4} && \text{(evaluate)}
\end{aligned}$$

Theorem 0.1 (A sum of IID Bernoulli variables is binomial)**.** *If $X_1, \ldots, X_n$ are IID $\text{Ber}(\pi)$ random variables, then their sum $S = \sum_{i=1}^{n} X_i$ has the $\text{Bin}(n, \pi)$ distribution.*

Proof

Proof. Each X_i is 0 or 1, so S takes values in $\{0, 1, \ldots, n\}$. Let $x \in \{0, 1, \ldots, n\}$ be a value of S, and let A_x be the set of sequences $(s_1, \ldots, s_n) \in \{0, 1\}^n$ with exactly x ones. The event $\{S = x\}$ is the disjoint union of the events $\{X_1 = s_1, \ldots, X_n = s_n\}$ over $(s_1, \ldots, s_n) \in A_x$. For one such sequence:

$$\begin{aligned}
\mathbb{P}(X_1 = s_1, \ldots, X_n = s_n) &= \prod_{i=1}^{n} \mathbb{P}(X_i = s_i) && \text{(independence)} \\
&= \prod_{i=1}^{n} \pi^{s_i} (1 - \pi)^{1 - s_i} && \text{(Bernoulli PMF)} \\
&= \pi^{\sum_{i=1}^{n} s_i} (1 - \pi)^{\sum_{i=1}^{n} (1 - s_i)} && \text{(collect exponents)} \\
&= \pi^x (1 - \pi)^{n - x} && \left(\textstyle\sum_{i=1}^{n} s_i = x \text{ for a sequence in } A_x\right)
\end{aligned}$$

Choosing which x of the n positions hold the ones determines the sequence, so A_x has $\binom{n}{x}$ elements. Therefore:

$$
\begin{aligned}
\mathbb{P}(S = x) &= \sum_{(s_1,\ldots,s_n)\in A_x} \mathbb{P}(X_1 = s_1, \ldots, X_n = s_n) && \text{(additivity over disjoint events)} \\
&= \sum_{(s_1,\ldots,s_n)\in A_x} \pi^x (1-\pi)^{n-x} && \text{(probability of one sequence)} \\
&= \binom{n}{x} \pi^x (1-\pi)^{n-x} && (A_x \text{ has } \tbinom{n}{x} \text{ elements})
\end{aligned}
$$

which is Equation 1. □

Example 0.3 (Counting treatment responders)**.** Suppose five patients each respond to a treatment with probability 0.2, independently of one another, and let $X_i = 1$ if patient i responds and $X_i = 0$ otherwise. By Theorem 0.1, the number of responders $S = \sum_{i=1}^{5} X_i$ is $\text{Bin}(5, 0.2)$, so the probability that nobody responds is:

$$
\begin{aligned}
\mathbb{P}(S = 0) &= \binom{5}{0} (0.2)^0 (1-0.2)^{5-0} && \text{(binomial PMF with } n = 5,\ \pi = 0.2,\ x = 0) \\
&= 1 \cdot 1 \cdot (0.8)^5 && (\tbinom{5}{0} = 1 \text{ and } (0.2)^0 = 1) \\
&= 0.32768 && \text{(evaluate the power)}
\end{aligned}
$$

Theorem 0.2 (Mean and variance of the binomial distribution)**.** *If $X \sim \text{Bin}(n, \pi)$, then:*

- $\mathbb{E}[X] = n\pi$
- $\text{Var}(X) = n\pi(1-\pi)$

Proof

Proof. The mean and variance of X depend only on its distribution, so by Theorem 0.1 we can compute them for $S = \sum_{i=1}^{n} X_i$, where $X_1, \ldots, X_n$ are IID $\text{Ber}(\pi)$. For the mean, apply linearity of expectation $n - 1$ times, then the Bernoulli mean:

$$
\begin{aligned}
\mathbb{E}[S] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] && \text{(definition of } S\text{)} \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{(linearity of expectation)} \\
&= \sum_{i=1}^{n} \pi && \text{(Bernoulli mean)} \\
&= n\pi && \text{(sum of } n \text{ copies of } \pi\text{)}
\end{aligned}
$$

For the variance, apply the variance of a linear combination with $a_1 = \cdots = a_n = 1$, then independent components give a diagonal variance matrix, then the Bernoulli variance:

$$
\begin{aligned}
\mathrm{Var}(S) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j) && \text{(variance of a linear combination with } a_i = 1\text{)} \\
&= \sum_{i=1}^{n} \mathrm{Cov}(X_i, X_i) && \text{(}\mathrm{Cov}(X_i, X_j) = 0 \text{ for } i \neq j\text{, by independence)} \\
&= \sum_{i=1}^{n} \mathrm{Var}(X_i) && \text{(}\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)\text{)} \\
&= \sum_{i=1}^{n} \pi(1-\pi) && \text{(Bernoulli variance)} \\
&= n\pi(1-\pi) && \text{(sum of } n \text{ copies of } \pi(1-\pi)\text{)}
\end{aligned}
$$

□

Example 0.4 (Expected number of responders)**.** In Example 0.3, $S \sim \mathrm{Bin}(5, 0.2)$. By Theorem 0.2:

$$
\begin{aligned}
\mathbb{E}[S] &= 5 \cdot 0.2 && (n\pi) \\
&= 1 && \text{(multiply)} \\
\mathrm{Var}(S) &= 5 \cdot 0.2 \cdot (1 - 0.2) && (n\pi(1-\pi)) \\
&= 0.8 && \text{(multiply)}
\end{aligned}
$$

On average, one of the five patients responds.

The Poisson distribution

Definition 0.1 (Poisson distribution)**.** A random variable Y has the **Poisson distribution** with mean parameter $\mu > 0$, written $Y \sim \text{Pois}(\mu)$ [3], if:

$$\mathbb{P}(Y = y) \overset{\text{def}}{=} \frac{\mu^y \mathrm{e}^{-\mu}}{y!} 1_{y \in \mathbb{N}}, \quad y \in \mathbb{R} \tag{2}$$

where $\mathbb{N} = \{0, 1, 2, \dots\}$ is the set of non-negative integers. The indicator makes the mass 0 outside $\mathbb{N}$, where $y!$ is undefined.

Remark

Remark. (see Figure 1)

Exercise 0.1. What is the range of possible values for a Poisson distribution?

Solution

Solution 0.1.

$$\begin{aligned} \mathcal{R}(Y) &= \{0, 1, 2, \dots\} \\ &= \mathbb{N} \end{aligned}$$

Theorem 0.1 (CDF of Poisson distribution)**.** *If $Y \sim \text{Pois}(\mu)$, then for every real number y:*

$$\mathbb{P}(Y \leq y) = e^{-\mu} \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j}{j!} \tag{3}$$

Proof

Proof. For $y < 0$, both sides are 0 (the sum is empty). For any $y \geq 0$, the event $\{Y \leq y\}$ is the disjoint union of events $\{Y = j\}$ for all non-negative integers $j \leq \lfloor y \rfloor$. Applying the Poisson PMF (Equation 2) and factoring out the common term $e^{-\mu}$:

$$
\begin{aligned}
\mathbb{P}(Y \leq y) &= \sum_{j=0}^{\lfloor y \rfloor} \mathbb{P}(Y = j) && \text{(disjoint union of events } Y = j\text{)} \\
&= \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j e^{-\mu}}{j!} && \text{(definition of Poisson PMF)} \\
&= e^{-\mu} \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j}{j!} && \text{(factoring out } e^{-\mu} \text{ constant wrt } j\text{)}
\end{aligned}
$$

□

Example 0.1 (Computing Poisson cumulative probabilities)**.** For a Poisson random variable $X \sim \text{Pois}(\mu = 2)$, the probability of observing at most 2 events is computed as:

$$
\begin{aligned}
\mathbb{P}(X \leq 2) &= e^{-2} \sum_{j=0}^{2} \frac{2^j}{j!} && \text{(apply CDF formula with } \mu = 2, y = 2\text{)} \\
&= e^{-2} \left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} \right) && \text{(expand terms for } j = 0, 1, 2\text{)} \\
&= e^{-2}(1 + 2 + 2) && \text{(simplify factorials and powers)} \\
&= 5e^{-2} \approx 0.677 && \text{(evaluate numerical value)}
\end{aligned}
$$

ℹ Remark

Remark. (see Figure 2)

```
pois_dists <-
  dplyr::tibble(mu = c(0.5, 1, 2, 5, 10, 20)) |>
  dplyr::reframe(.by = mu, x = 0:30) |>
  dplyr::mutate(
    `P(X = x)` = dpois(x, lambda = mu),
    `P(X <= x)` = ppois(x, lambda = mu),
    mu = factor(mu)
  )

plot0 <-
  pois_dists |>
  ggplot2::ggplot(
    ggplot2::aes(
      x = x,
      y = `P(X = x)`,
```

```
      fill = mu,
      col = mu
    )
  ) +
  ggplot2::theme(legend.position = "bottom") +
  ggplot2::labs(
    fill = latex2exp::TeX("$\\mu$"),
    col = latex2exp::TeX("$\\mu$"),
    y = latex2exp::TeX("$\\Pr_{\\mu}(X = x)$")
  )

plot1 <-
  plot0 +
  ggplot2::geom_segment(yend = 0) +
  ggplot2::facet_wrap(~mu)

print(plot1)
```

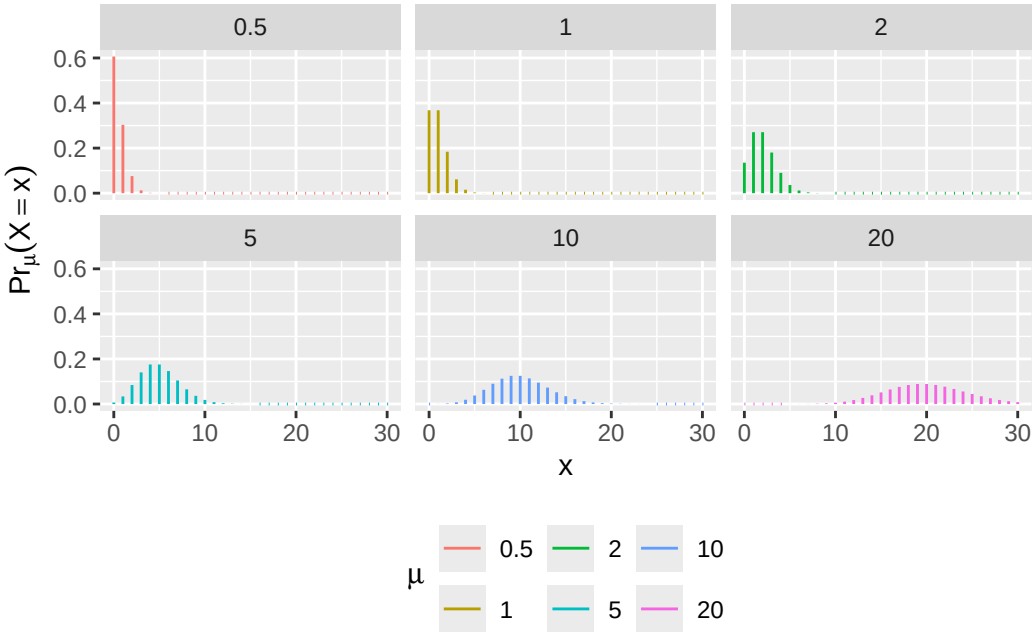


Figure 1: Poisson PMFs, by mean parameter μ

```r
plot2 <-
  plot0 +
  ggplot2::geom_step(alpha = 0.75) +
  ggplot2::aes(y = `P(X <= x)`) +
  ggplot2::labs(y = latex2exp::TeX("$\\Pr_{\\mu}(X \\leq x)$"))

print(plot2)
```

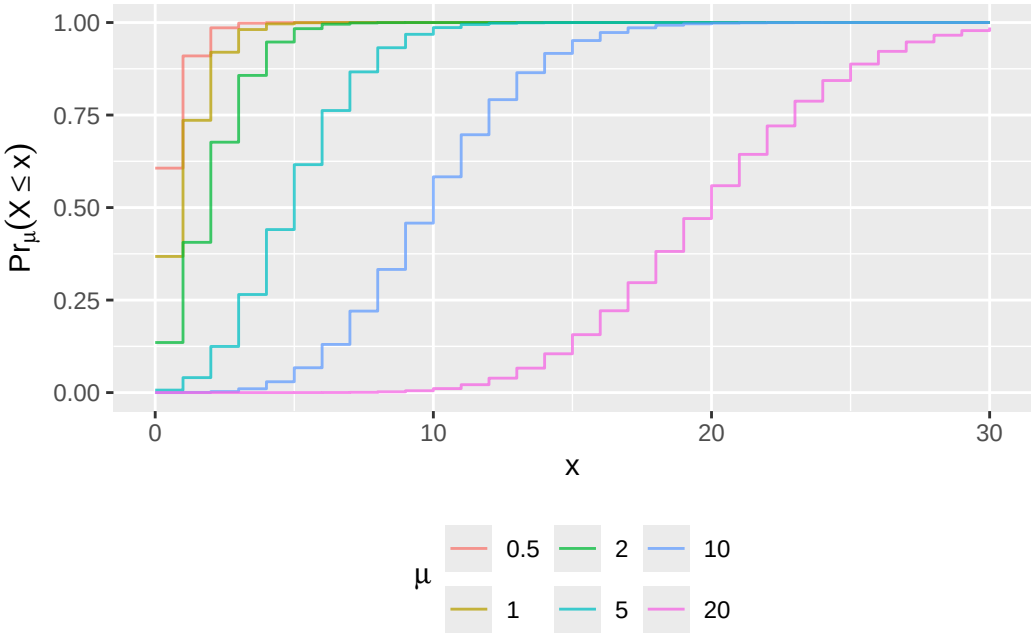


Figure 2: Poisson CDFs

Exercise 0.2 (Poisson distribution functions)**.** Let $X \sim \text{Pois}(\mu = 3.75)$. Compute:

- $\mathbb{P}(X = 4)$
- $\mathbb{P}(X \leq 7)$
- $\mathbb{P}(X > 5)$

Solution

Solution.

- $\mathbb{P}(X = 4) = 0.1937803$

- $\mathbb{P}(X \leq 7) = 0.9623787$
- $\mathbb{P}(X > 5) = 0.1771172$

Theorem 0.2 (Mean and variance of the Poisson distribution)**.** *If* $X \sim \text{Pois}(\mu)$*, then:*

- $\mathbb{E}[X] = \mu$
- $\text{Var}(X) = \mu$

Proof

Proof. **Mean.**

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{x=0}^{\infty} x \cdot \mathbb{P}(X = x) && \text{(definition of expected value)} \\
&= 0 \cdot \mathbb{P}(X = 0) + \sum_{x=1}^{\infty} x \cdot \mathbb{P}(X = x) && \text{(separate } x = 0 \text{ term)} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\mu^x e^{-\mu}}{x!} && \text{(substitute Poisson PMF)} \\
&= \sum_{x=1}^{\infty} x \cdot \frac{\mu^x e^{-\mu}}{x \cdot (x-1)!} && \text{(definition of factorial } x!) \\
&= \sum_{x=1}^{\infty} \frac{\mu^x e^{-\mu}}{(x-1)!} && \text{(cancel factor of } x) \\
&= \mu \cdot \sum_{x=1}^{\infty} \frac{\mu^{x-1} e^{-\mu}}{(x-1)!} && \text{(factor out one power of } \mu) \\
&= \mu \cdot \sum_{y=0}^{\infty} \frac{\mu^y e^{-\mu}}{y!} && \text{(change index variable } y \stackrel{\text{def}}{=} x - 1) \\
&= \mu \cdot 1 && \text{(PMF sums to 1 over state space)} \\
&= \mu && \text{(simplify)}
\end{aligned}$$

Variance. The same steps, canceling two factors instead of one, give $\mathbb{E}[X(X-1)]$:

$$\begin{aligned}
\mathbb{E}[X(X-1)] &= \sum_{x=2}^{\infty} x(x-1) \cdot \frac{\mu^x e^{-\mu}}{x!} && \text{(LOTUS; the } x = 0, 1 \text{ terms are 0)} \\
&= \sum_{x=2}^{\infty} \frac{\mu^x e^{-\mu}}{(x-2)!} && \text{(cancel } x(x-1) \text{ against } x!) \\
&= \mu^2 \cdot \sum_{y=0}^{\infty} \frac{\mu^y e^{-\mu}}{y!} && \text{(factor out } \mu^2;\ y \stackrel{\text{def}}{=} x - 2) \\
&= \mu^2 && \text{(PMF sums to 1)}
\end{aligned}$$

Then, by the simplified expression for variance and linearity of expectation:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 && \text{(simplified expression for variance)} \\
&= \mathbb{E}[X(X-1)] + \mathbb{E}[X] - (\mathbb{E}[X])^2 && (X^2 = X(X-1) + X\text{; linearity}) \\
&= \mu^2 + \mu - \mu^2 && \text{(substitute)} \\
&= \mu && \text{(simplify)}
\end{aligned}$$

See also https://statproofbook.github.io/P/poiss-mean and https://statproofbook.github.io/P/poiss-var. □

Theorem 0.3 (Ratio of consecutive Poisson probabilities, and the mode)**.** *If* $X \sim \mathrm{Pois}(\mu)$*, then:*

- $\mathbb{P}(X = x) = \frac{\mu}{x}\mathbb{P}(X = x-1)$ *for* $x \in \{1, 2, \dots\}$
- *For* $x \in \{1, 2, \dots\}$ *with* $x < \mu$, $\mathbb{P}(X = x) > \mathbb{P}(X = x-1)$
- *For* $x = \mu$ *(possible only when* μ *is an integer),* $\mathbb{P}(X = x) = \mathbb{P}(X = x-1)$
- *For* $x \in \{1, 2, \dots\}$ *with* $x > \mu$, $\mathbb{P}(X = x) < \mathbb{P}(X = x-1)$
- *If* μ *is not an integer,* $\arg\max_x \mathbb{P}(X = x) = \lfloor \mu \rfloor$*; if* μ *is an integer, the maximum is attained at both* $x = \mu - 1$ *and* $x = \mu$

Proof

Proof. **Ratio of consecutive probabilities.** For $x \in \{1, 2, \dots\}$:

$$\begin{aligned}
\frac{\mathbb{P}(X = x)}{\mathbb{P}(X = x-1)} &= \frac{\mu^x e^{-\mu}/x!}{\mu^{x-1}e^{-\mu}/(x-1)!} && \text{(substitute Poisson PMF)} \\
&= \frac{\mu}{x} && \text{(cancel } \mu^{x-1}e^{-\mu} \text{ and } (x-1)!\text{)}
\end{aligned}$$

This ratio is greater than, equal to, or less than 1 as x is less than, equal to, or greater than μ, which gives the three comparisons. So $\mathbb{P}(X = x)$ increases while $x < \mu$ and decreases once $x > \mu$. If μ is not an integer, the last increase is at $x = \lfloor \mu \rfloor$, which is therefore the unique mode. If μ is an integer, $\mathbb{P}(X = \mu) = \mathbb{P}(X = \mu - 1)$, and both are modes. □

Example 0.2 (Mode of a Poisson distribution)**.** For $X \sim \mathrm{Pois}(2.5)$, $\mathbb{P}(X = 2) = \frac{2.5}{2}\mathbb{P}(X = 1) > \mathbb{P}(X = 1)$ and $\mathbb{P}(X = 3) = \frac{2.5}{3}\mathbb{P}(X = 2) < \mathbb{P}(X = 2)$, so the mode is $\lfloor 2.5 \rfloor = 2$. For $X \sim \mathrm{Pois}(2)$, $\mathbb{P}(X = 2) = \frac{2}{2}\mathbb{P}(X = 1)$, so 1 and 2 are both modes, each with probability $2e^{-2} \approx 0.271$.

Definition 0.2 (Exposure magnitude)**.** For many count outcomes, there is some sense of an **exposure magnitude**, such as population size or duration of observation, which multiplicatively rescales the expected (mean) count.

Exercise 0.3. What are some examples of exposure magnitudes?

Solution

Solution.

Table 3: Examples of exposure units

outcome	exposure units
disease incidence	number of individuals exposed; time at risk
car accidents	miles driven
worksite accidents	person-hours worked
population size	size of habitat

Remark

Remark. Exposure units are similar to the number of trials in a binomial distribution, but **in non-binomial count outcomes, there can be more than one event per unit of exposure**.
We can use t to represent continuous-valued exposures/observation durations, and n to represent discrete-valued exposures.

Definition 0.3 (Event rate)**.** For a count Y observed over a fixed, known exposure magnitude $t > 0$, the **event rate**, denoted λ, is the mean of Y divided by the exposure magnitude:

$$\lambda \stackrel{\text{def}}{=} \frac{\mathbb{E}[Y]}{t} \tag{4}$$

Remark

Remark. The event rate is a transformation of the mean: it removes the exposure magnitude from the mean, so counts observed over different exposures can be compared on one scale. Regression models for counts use the same decomposition, with the rate depending on covariates and the exposure magnitude known (see rme's count-regression chapter).

Theorem 0.4 (Transformation function from event rate to mean)**.** *If a count Y is observed over exposure magnitude $t > 0$ with event rate λ, then its mean $\mu \overset{def}{=} \mathbb{E}[Y]$ is:*

$$\mu = \lambda \cdot t \tag{5}$$

Proof

Proof. By Definition 0.3:

$$\begin{aligned} \lambda &\overset{\text{def}}{=} \frac{\mathbb{E}[Y]}{t} && \text{(definition of event rate)} \\ \mathbb{E}[Y] &= \lambda \cdot t && \text{(multiply both sides by } t > 0\text{)} \\ \mu &= \lambda \cdot t && (\mu \overset{\text{def}}{=} \mathbb{E}[Y]) \end{aligned}$$

□

Example 0.3 (Calculating expected counts from event rates)**.** Suppose a city records a disease event rate of $\lambda = 0.05$ cases per person-year. For a subpopulation with an exposure magnitude of $t = 100$ person-years, the expected count of cases is, by Theorem 0.4:

$$\begin{aligned} \mu &= \lambda \cdot t && \text{(transformation from event rate to mean)} \\ &= 0.05 \times 100 && \text{(substitute } \lambda = 0.05 \text{ and } t = 100\text{)} \\ &= 5 \text{ cases} && \text{(evaluate expected count)} \end{aligned}$$

Theorem 0.5 (No exposure means no expected events)**.** *For each exposure magnitude $t \geq 0$, let Y_t be the count observed over exposure t. If the mean count is proportional to the exposure magnitude, $\mathbb{E}[Y_t] = \lambda \cdot t$ for all $t \geq 0$ with one finite rate λ, then there are no expected events without exposure:*

$$\mathbb{E}[Y_0] = 0$$

Proof

Proof.

$$\begin{aligned} \mathbb{E}[Y_0] &= \lambda \cdot 0 && \text{(evaluate } \mathbb{E}[Y_t] = \lambda t \text{ at } t = 0\text{)} \\ &= 0 && \text{(multiplication by zero; } \lambda \text{ is finite)} \end{aligned}$$

□

Remark

Remark. The hypothesis carries the content here. Definition 0.3 alone cannot give $\mathbb{E}[Y_0] = 0$, since $\lambda = \mathbb{E}[Y]/t$ is undefined at $t = 0$; the result holds for a model that assumes one rate λ shared across exposure magnitudes, including $t = 0$.

Example 0.4 (Zero exposure time)**.** If a subject is observed for $t = 0$ person-years, no follow-up time has elapsed, so under a constant-rate model the expected number of incident events is $\mathbb{E}[Y_0] = 0$.

Important

A mean proportional to exposure, $\mathbb{E}[Y_t] = \lambda t$, has no term that stays nonzero at $t = 0$: a model of this form says that with no exposure, no events are expected. A mean with an added constant, such as $\mathbb{E}[Y_t] = c + \lambda t$ with $c > 0$, would expect c events even with no exposure. Regression models for counts keep the proportional form when they add covariates (see rme's count-regression chapter).

Theorem 0.6 (Exposure is additive on the log scale)**.** *If $\mu = \lambda \cdot t$ with $\lambda > 0$ and $t > 0$, then:*

$$\log \mu = \log \lambda + \log t$$

Proof

Proof.

$$\begin{aligned} \log \mu &= \log(\lambda \cdot t) && \text{(substitute } \mu = \lambda \cdot t\text{)} \\ &= \log \lambda + \log t && \text{(logarithm product rule)} \end{aligned}$$

□

Example 0.5 (Log-linear representation of expected counts)**.** If a clinic sees an event rate of $\lambda = 0.02$ events/day and $t = 30$ days of observation:

$$\begin{aligned} \log \mu &= \log(0.02) + \log(30) && \text{(apply log-scale formula)} \\ &= -3.912 + 3.401 && \text{(evaluate natural logarithms)} \\ &= -0.511 && \text{(sum terms)} \end{aligned}$$

Exponentiating yields $\mu = \exp\{-0.511\} \approx 0.60$ expected events.

Definition 0.4 (Offset)**.** In a model whose linear component contains a term with no unknown coefficient, that term is called an **offset**: it shifts the linear component by a known amount. In a count model with exposure magnitude t, the offset is the known term $\log t$ in the log-scale decomposition of the mean (0.6),

$$\log \mu = \log \lambda + \log t.$$

Remark

Remark. The offset needs no covariates: with a single exposure t and an unknown rate λ, $\log t$ is already an offset. Regression models for counts keep the same term and add covariate terms beside it (see rme's count-regression chapter).

Example 0.6 (The offset for a clinic's event count)**.** In Example 0.5, the clinic was observed for $t = 30$ days, so the offset is $\log 30 \approx 3.401$. Only $\log \lambda$ is unknown before the data are seen; the offset is fixed by the length of observation. A clinic observed for $t = 60$ days has offset $\log 60 \approx 4.094$, which raises $\log \mu$ by $\log 2 \approx 0.693$ at the same event rate.

Theorem 0.7 (Sum of independent Poisson random variables)**.** *If X and Y are independent Poisson random variables with means μ_X and μ_Y, their sum, $Z = X + Y$, is also a Poisson random variable, with mean $\mu_Z = \mu_X + \mu_Y$.*

Proof

Proof. The event $\{Z = z\}$ is the disjoint union of the events $\{X = k, Y = z - k\}$ for $k = 0, 1, \ldots, z$ (since X and Y are non-negative integers), so:

$$
\begin{aligned}
\mathbb{P}(Z=z) &= \sum_{k=0}^{z} \mathbb{P}(X=k,\, Y=z-k) && \text{(countable additivity)} \\
&= \sum_{k=0}^{z} \mathbb{P}(X=k)\mathbb{P}(Y=z-k) && \text{(independence)} \\
&= \sum_{k=0}^{z} \frac{\mu_X^k e^{-\mu_X}}{k!} \frac{\mu_Y^{z-k} e^{-\mu_Y}}{(z-k)!} && \text{(substitute Poisson PMFs)} \\
&= e^{-(\mu_X+\mu_Y)} \sum_{k=0}^{z} \frac{\mu_X^k \mu_Y^{z-k}}{k!(z-k)!} && \text{(factor out } e^{-\mu_X} e^{-\mu_Y}\text{)} \\
&= \frac{e^{-(\mu_X+\mu_Y)}}{z!} \sum_{k=0}^{z} \frac{z!}{k!(z-k)!} \mu_X^k \mu_Y^{z-k} && \text{(multiply and divide by } z!\text{)} \\
&= \frac{e^{-(\mu_X+\mu_Y)}}{z!} (\mu_X+\mu_Y)^z && \text{(binomial theorem)}
\end{aligned}
$$

This probability expression matches the PMF of a $\text{Pois}(\mu_X + \mu_Y)$ random variable (see also https://web.stanford.edu/class/archive/cs/cs109/cs109.1206/lectureNotes/LN12_independent_rvs.pdf, Example 3). □

Example 0.7 (Aggregating independent region counts)**.** Suppose Region A records $X \sim \text{Pois}(\mu_X = 12)$ cases and Region B records $Y \sim \text{Pois}(\mu_Y = 18)$ cases independently. By Theorem 0.7, the combined total count $Z = X + Y$ follows a Poisson distribution:

$$
\begin{aligned}
Z &\sim \text{Pois}(\mu_X + \mu_Y) && \text{(sum of independent Poissons)} \\
&= \text{Pois}(12 + 18) && \text{(substitute region means)} \\
&= \text{Pois}(30) && \text{(evaluate sum)}
\end{aligned}
$$

The Negative-Binomial distribution

Definition 0.1 (Negative binomial distribution)**.** A random variable Y has the **negative binomial distribution** with mean $\mu > 0$ and overdispersion parameter $\rho > 0$, written $Y \sim \text{NegBin}(\mu, \rho)$ [4], if, for $y \in \{0, 1, 2, \dots\}$:

$$
\mathbb{P}(Y=y) \stackrel{\text{def}}{=} \frac{\mu^y}{y!} \cdot \frac{\Gamma(\rho+y)}{\Gamma(\rho)\cdot(\rho+\mu)^y} \cdot \left(1+\frac{\mu}{\rho}\right)^{-\rho}
$$

where Γ is the gamma function, which satisfies $\Gamma(x) = (x-1)!$ for positive integers x.

Theorem 0.1 (The negative binomial converges to the Poisson)**.** *Let the population mean be $\mu > 0$, and let $Y_\rho \sim \text{NegBin}(\mu, \rho)$. Then for each $y \in \{0, 1, 2, \dots\}$, as $\rho \to \infty$, the negative binomial PMF converges to the Poisson PMF (Equation 2):*

$$\lim_{\rho \to \infty} \mathbb{P}(Y_\rho = y) = \frac{\mu^y e^{-\mu}}{y!}$$

Proof

Proof. Choose a value y of Y_ρ. In Definition 0.1, the first factor $\mu^y/y!$ does not depend on ρ. For the second factor, $\Gamma(\rho+y) = \Gamma(\rho)\prod_{k=0}^{y-1}(\rho+k)$, by applying $\Gamma(x+1) = x\,\Gamma(x)$ y times (for $y = 0$, the product is empty and equals 1), so:

$$\begin{aligned}
\frac{\Gamma(\rho+y)}{\Gamma(\rho)\cdot(\rho+\mu)^y} &= \frac{\Gamma(\rho)\prod_{k=0}^{y-1}(\rho+k)}{\Gamma(\rho)\cdot(\rho+\mu)^y} && (\Gamma(x+1) = x\,\Gamma(x), \text{ applied } y \text{ times}) \\
&= \prod_{k=0}^{y-1}\frac{\rho+k}{\rho+\mu} && (\text{cancel } \Gamma(\rho); \text{ one factor of } \rho+\mu \text{ per } k) \\
&= \prod_{k=0}^{y-1}\frac{1+k/\rho}{1+\mu/\rho} && (\text{divide each numerator and denominator by } \rho) \\
&\to \prod_{k=0}^{y-1}\frac{1+0}{1+0} && (k/\rho \to 0 \text{ and } \mu/\rho \to 0; \text{ finitely many factors}) \\
&= 1 && (\text{simplify})
\end{aligned}$$

For the third factor, write $x \overset{\text{def}}{=} \mu/\rho$, so $x \to 0$ as $\rho \to \infty$, and $\rho = \mu/x$:

$$\begin{aligned}
\log\left(\left(1+\frac{\mu}{\rho}\right)^{-\rho}\right) &= -\rho\log\left(1+\frac{\mu}{\rho}\right) && (\text{log of a power}) \\
&= -\mu\cdot\frac{\log(1+x)}{x} && (\text{substitute } \rho = \mu/x) \\
&\to -\mu\cdot 1 && (\lim_{x\to 0}\log(1+x)/x = 1, \text{ the derivative of } \log(1+x) \text{ at } x = 0) \\
&= -\mu && (\text{simplify})
\end{aligned}$$

so, since the exponential function is continuous, $(1+\mu/\rho)^{-\rho} \to \exp\{-\mu\}$. Each of the three factors has a limit, so their product does too:

$$\begin{aligned}
\lim_{\rho\to\infty}\mathbb{P}(Y_\rho = y) &= \frac{\mu^y}{y!}\cdot 1\cdot\exp\{-\mu\} && (\text{limit of a product of convergent factors}) \\
&= \frac{\mu^y e^{-\mu}}{y!} && (\text{rearrange})
\end{aligned}$$

which is the Pois(μ) PMF. □

Theorem 0.2 (Mean and variance of the negative binomial distribution)**.** *If* $Y \sim \text{NegBin}(\mu, \rho)$*, then:*

- $\mathbb{E}[Y] = \mu$
- $\text{Var}(Y) = \mu + \frac{\mu^2}{\rho} > \mu$

Proof

Proof. The negative binomial distribution is a gamma mixture of Poisson distributions: if Λ has the gamma density $g(\lambda) = \frac{(\rho/\mu)^\rho}{\Gamma(\rho)} \lambda^{\rho-1} e^{-\rho\lambda/\mu}$ for $\lambda > 0$, which has mean μ and variance μ^2/ρ [2], and $Y \mid \Lambda = \lambda \sim \text{Pois}(\lambda)$, then $Y \sim \text{NegBin}(\mu, \rho)$. To check this, integrate the joint density over λ:

$$
\begin{aligned}
\mathbb{P}(Y = y) &= \int_0^\infty \frac{\lambda^y e^{-\lambda}}{y!} \cdot \frac{(\rho/\mu)^\rho}{\Gamma(\rho)} \lambda^{\rho-1} e^{-\rho\lambda/\mu}\, d\lambda && \text{(marginalize over } \Lambda\text{)} \\
&= \frac{(\rho/\mu)^\rho}{y!\,\Gamma(\rho)} \int_0^\infty \lambda^{y+\rho-1} e^{-\lambda(1+\rho/\mu)}\, d\lambda && \text{(collect powers of } \lambda \text{ and exponents)} \\
&= \frac{(\rho/\mu)^\rho}{y!\,\Gamma(\rho)} \cdot \frac{\Gamma(y+\rho)}{(1+\rho/\mu)^{y+\rho}} && \left(\textstyle\int_0^\infty \lambda^{a-1} e^{-b\lambda}\, d\lambda = \Gamma(a)/b^a\right) \\
&= \frac{\Gamma(y+\rho)}{y!\,\Gamma(\rho)} \left(\frac{\rho}{\mu}\right)^\rho \left(\frac{\mu}{\mu+\rho}\right)^{y+\rho} && (1+\rho/\mu = (\mu+\rho)/\mu) \\
&= \frac{\Gamma(y+\rho)}{y!\,\Gamma(\rho)} \left(\frac{\rho}{\mu+\rho}\right)^\rho \left(\frac{\mu}{\mu+\rho}\right)^{y} && \text{(combine the } \mu^\rho \text{ factors)} \\
&= \frac{\mu^y}{y!} \cdot \frac{\Gamma(\rho+y)}{\Gamma(\rho)\,(\rho+\mu)^y} \cdot \left(1+\frac{\mu}{\rho}\right)^{-\rho} && \text{(rearrange into the form of the definition)}
\end{aligned}
$$

Then, by the law of iterated expectations and the law of total variance, using

$$
\begin{aligned}
\mathbb{E}[Y \mid \Lambda] &= \text{Var}(Y \mid \Lambda) \\
&= \Lambda
\end{aligned}
$$

(Theorem 0.2):

$$
\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[\mathbb{E}[Y \mid \Lambda]] && \text{(law of iterated expectations)} \\
&= \mathbb{E}[\Lambda] && (\mathbb{E}[Y \mid \Lambda] = \Lambda) \\
&= \mu && \text{(mean of the gamma distribution)} \\
\operatorname{Var}(Y) &= \mathbb{E}[\operatorname{Var}(Y \mid \Lambda)] + \operatorname{Var}(\mathbb{E}[Y \mid \Lambda]) && \text{(law of total variance)} \\
&= \mathbb{E}[\Lambda] + \operatorname{Var}(\Lambda) && (\operatorname{Var}(Y \mid \Lambda) = \Lambda \text{ and } \mathbb{E}[Y \mid \Lambda] = \Lambda) \\
&= \mu + \frac{\mu^2}{\rho} && \text{(mean and variance of the gamma distribution)}
\end{aligned}
$$

and $\mu^2/\rho > 0$ gives $\operatorname{Var}(Y) > \mu$. □

Example 0.1 (Overdispersion relative to the Poisson)**.** With $\mu = 4$ and $\rho = 2$,

$$
\begin{aligned}
\operatorname{Var}(Y) &= 4 + 16/2 \\
&= 12,
\end{aligned}
$$

three times the variance of a Pois(4) count with the same mean.

The Laplace distribution

Definition 0.1 (Laplace distribution)**.** A random variable Y has the **Laplace distribution** with location parameter $\mu \in \mathbb{R}$ and scale parameter $b > 0$, written $Y \sim \text{Laplace}(\mu, b)$ [5], if Y is continuous with density:

$$
\mathbb{p}(Y = y) \stackrel{\text{def}}{=} \frac{1}{2b} \mathrm{e}^{-\frac{|y-\mu|}{b}}, \quad y \in \mathbb{R} \tag{6}
$$

Remark

Remark. The normal density has the squared distance $(y - \mu)^2$ in its exponent. The Laplace density has the absolute distance $|y - \mu|$ instead. Both densities are symmetric about μ, but the Laplace density has a sharp peak at μ and decays more slowly in the tails.

Example 0.1 (Laplace density values)**.** Let $Y \sim \text{Laplace}(3, 2)$, so $\mu = 3$ and $b = 2$ (Definition 0.1). At the location, $|y - \mu| = 0$, so the density is $\frac{1}{2b}$. At a point $y = 5$, the distance is

$$
\begin{aligned}
|5 - 3| &= 2 \\
&= b,
\end{aligned}
$$

so the density is $\frac{1}{2b}\mathrm{e}^{-1}$. Here are both values:

```
b <- 2
c(
  at_mu = 1 / (2 * b),
  at_mu_plus_2 = exp(-1) / (2 * b)
) |>
  round(3)
#>        at_mu at_mu_plus_2
#>        0.250        0.092
```

The density at $\mu + b$ is the density at μ times e^{-1}, and the same holds at $\mu - b$, because the density depends only on the distance from μ.

Theorem 0.1 (Mean and variance of the Laplace distribution)**.** *If* $Y \sim \text{Laplace}(\mu, b)$ *(Definition 0.1), then*

$$
\mathbb{E}[Y] = \mu \quad \textit{and} \quad \text{Var}(Y) = 2b^2. \tag{7}
$$

Proof

Proof. Write $U = Y - \mu$. By Equation 6, U has density $\frac{1}{2b}\mathrm{e}^{-|u|/b}$, which is symmetric about 0.
For the mean, the integrand $u \cdot \frac{1}{2b}\mathrm{e}^{-|u|/b}$ is an odd function, so its integral over $\mathbb{R}$ is 0. Therefore $\mathbb{E}[U] = 0$, and

$$
\begin{aligned}
\mathbb{E}[Y] &= \mu + \mathbb{E}[U] \\
&= \mu.
\end{aligned}
$$

For the variance,

$$
\begin{aligned}
\text{Var}(Y) &= \text{Var}(U) \\
&= \mathbb{E}\big[U^2\big]
\end{aligned}
$$

because $\mathbb{E}[U] = 0$. The integrand $u^2 \cdot \frac{1}{2b}\mathrm{e}^{-|u|/b}$ is even, so

$$
\begin{aligned}
\mathbb{E}[U^2] &= 2\int_0^\infty u^2 \frac{1}{2b} \mathrm{e}^{-u/b}\, du && \text{(even integrand)} \\
&= \frac{1}{b}\int_0^\infty u^2 \mathrm{e}^{-u/b}\, du && \text{(multiplying the constants)} \\
&= \frac{1}{b}\left(2b\int_0^\infty u \mathrm{e}^{-u/b}\, du\right) && \text{(integration by parts; the boundary terms are 0)} \\
&= \frac{1}{b}(2b \cdot b^2) && \text{(integration by parts again; the boundary terms are 0)} \\
&= 2b^2. && \text{(multiplying the constants)}
\end{aligned}
$$

□

Example 0.2 (Laplace and normal tails with the same mean and variance)**.** A Laplace variable $Y \sim \text{Laplace}(0, b)$ has variance $2b^2$ (Theorem 0.1). Choosing $b = 1/\sqrt{2}$ gives variance 1, the same mean and variance as a standard normal. Figure 3 compares the two densities. The Laplace density is higher at the center and in the far tails, and lower in between.

```
b <- 1 / sqrt(2)
grid <- seq(-4, 4, length.out = 801)
dlaplace <- function(y, mu = 0, b = 1) exp(-abs(y - mu) / b) / (2 * b)
densities <- tibble::tibble(
  y = rep(grid, times = 2),
  distribution = rep(c("Laplace", "Normal"), each = length(grid)),
  density = c(dlaplace(grid, b = b), dnorm(grid))
)
ggplot2::ggplot(densities) +
  ggplot2::aes(x = y, y = density, colour = distribution) +
  ggplot2::geom_line() +
  ggplot2::labs(colour = NULL)
```

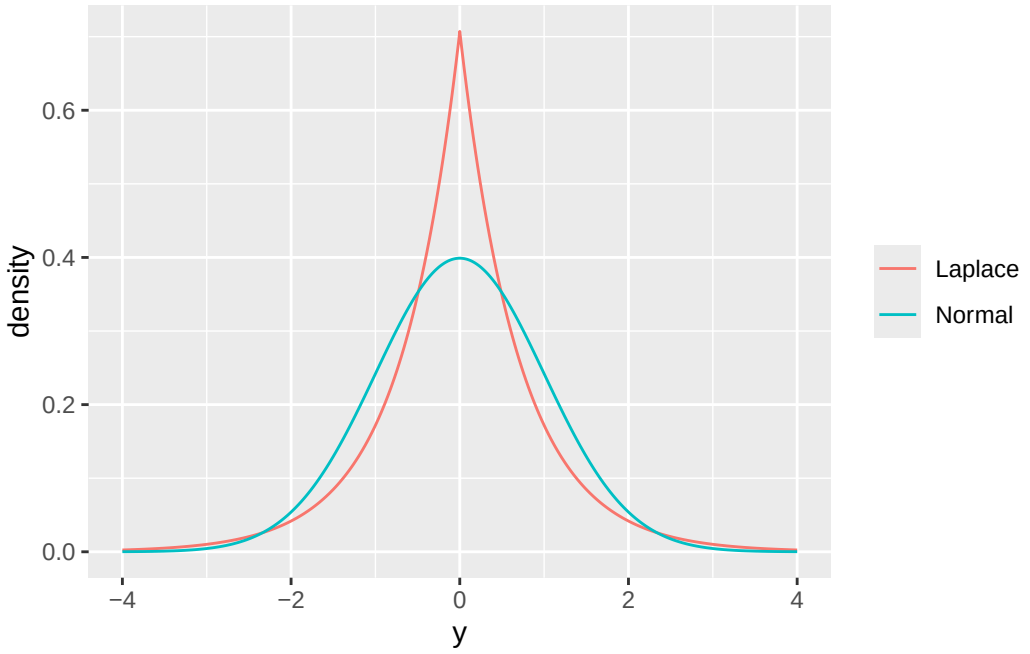


Figure 3: Laplace and standard normal densities with the same mean (0) and variance (1)

The probability of landing more than 3 standard deviations from the mean shows the difference in the tails. For the Laplace distribution, $\mathbb{P}(|Y| > t) = \mathrm{e}^{-t/b}$, because each tail has probability $\frac{1}{2}\mathrm{e}^{-t/b}$ by integrating Equation 6.

```
t <- 3
c(
  Laplace = exp(-t / b),
  Normal = 2 * pnorm(-t)
)
#>     Laplace      Normal
#> 0.014369596 0.002699796
```

The Laplace distribution puts about 5.3 times as much probability beyond 3 standard deviations as the normal distribution does.

The gamma distribution

Definition 0.1 (Gamma function)**.** The **gamma function** Γ is defined for each $s > 0$ by:

$$\Gamma(s) \stackrel{\text{def}}{=} \int_0^\infty u^{s-1} \mathrm{e}^{-u}\, du$$

Definition 0.2 (Gamma distribution)**.** A non-negative random variable T has the **gamma distribution** with shape $\alpha > 0$ and rate $\lambda > 0$, written $T \sim \mathrm{Gam}(\alpha, \lambda)$, if T is continuous with density:

$$f(t) \stackrel{\text{def}}{=} \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)}\, t^{\alpha-1} \mathrm{e}^{-\lambda t}, & t > 0 \\ 0, & t \le 0 \end{cases}$$

where Γ is the gamma function.

Example 0.1 (Gamma density with shape 2)**.** If $T \sim \mathrm{Gam}(2, \lambda)$, then $\Gamma(2) = 1$, so the density is:

$$f(t) = \lambda^2\, t\, \mathrm{e}^{-\lambda t}, \quad t > 0$$

The density is 0 at $t = 0$, rises to a peak at $t = 1/\lambda$, and then decays (Figure 4).

```r
rates <- c(0.5, 1, 2)
t_grid <- seq(0, 8, length.out = 400)
rate_label <- latex2exp::TeX("$\\lambda$")
density_label <- latex2exp::TeX("$f(t)$")

gamma_densities <-
  dplyr::tibble(rate = rates) |>
  dplyr::reframe(.by = rate, t = t_grid) |>
  dplyr::mutate(
    density = dgamma(t, shape = 2, rate = rate),
    rate = factor(rate)
  )

gamma_plot <-
  gamma_densities |>
  ggplot2::ggplot() +
  ggplot2::aes(x = t, y = density, col = rate) +
  ggplot2::geom_line() +
  ggplot2::labs(col = rate_label, y = density_label) +
  ggplot2::theme(legend.position = "bottom")

print(gamma_plot)
```

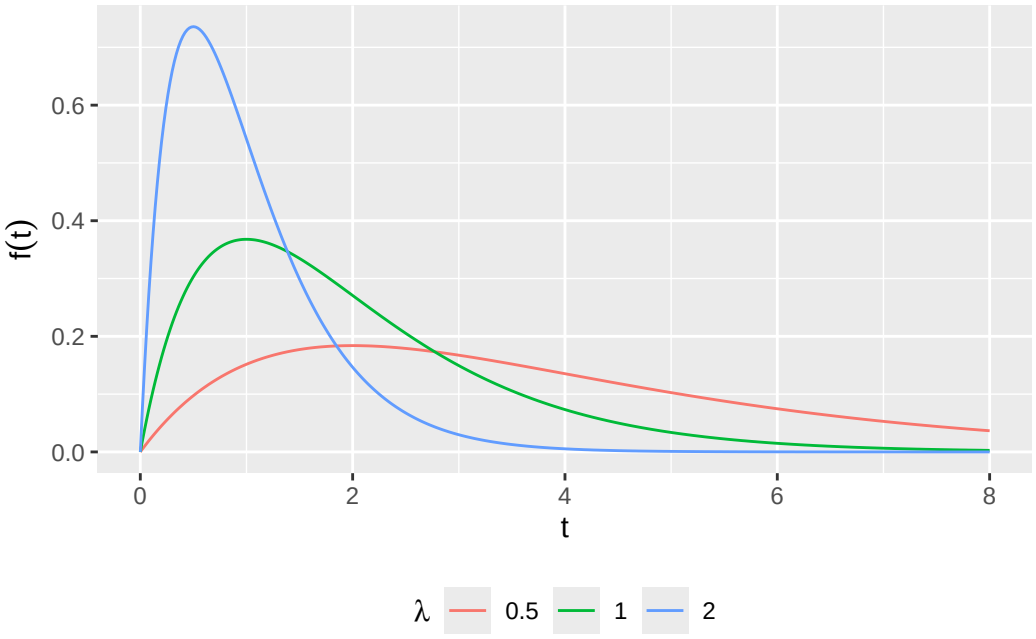


Figure 4: Gamma densities with shape 2, by rate λ

Theorem 0.1 (Moments of the gamma distribution)**.** *If* $T \sim \text{Gam}(\alpha, \lambda)$*, then for every integer* $k \geq 0$*:*

$$\mathbb{E}[T^k] = \frac{\Gamma(\alpha + k)}{\Gamma(\alpha)\,\lambda^k}$$

Proof

Proof. The substitution below reduces the expectation to the gamma function:

$$
\begin{aligned}
\mathbb{E}[T^k] &= \int_0^\infty t^k \, \frac{\lambda^\alpha}{\Gamma(\alpha)} \, t^{\alpha-1} \mathrm{e}^{-\lambda t} \, dt && \text{(definition of expectation, with the gamma density)} \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty t^{\alpha+k-1} \mathrm{e}^{-\lambda t} \, dt && (t^k \cdot t^{\alpha-1} = t^{\alpha+k-1}) \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{1}{\lambda^{\alpha+k}} \int_0^\infty u^{\alpha+k-1} \mathrm{e}^{-u} \, du && \text{(substitute } u = \lambda t) \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \cdot \frac{\Gamma(\alpha+k)}{\lambda^{\alpha+k}} && \text{(definition of } \Gamma) \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)\,\lambda^k} && (\lambda^\alpha / \lambda^{\alpha+k} = \lambda^{-k})
\end{aligned}
$$

□

Corollary 0.1 (Mean of the gamma distribution)**.** *If* $T \sim \mathrm{Gam}(\alpha, \lambda)$*, then* $\mathbb{E}[T] = \alpha/\lambda$*.*

Proof

Proof. By Theorem 0.1 with $k = 1$:

$$
\begin{aligned}
\mathbb{E}[T] &= \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)\,\lambda} && \text{(moment formula with } k = 1) \\
&= \frac{\alpha\,\Gamma(\alpha)}{\Gamma(\alpha)\,\lambda} && (\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)) \\
&= \frac{\alpha}{\lambda} && \text{(cancel } \Gamma(\alpha))
\end{aligned}
$$

□

Corollary 0.2 (Variance of the gamma distribution)**.** *If* $T \sim \mathrm{Gam}(\alpha, \lambda)$*, then* $\mathrm{Var}(T) = \alpha/\lambda^2$*.*

Proof

Proof. First, the second moment, by Theorem 0.1 with $k = 2$:

$$
\begin{aligned}
\mathbb{E}[T^2] &= \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)\,\lambda^2} && \text{(moment formula with } k=2\text{)} \\
&= \frac{(\alpha+1)\,\Gamma(\alpha+1)}{\Gamma(\alpha)\,\lambda^2} && (\Gamma(\alpha+2) = (\alpha+1)\,\Gamma(\alpha+1)) \\
&= \frac{(\alpha+1)\,\alpha\,\Gamma(\alpha)}{\Gamma(\alpha)\,\lambda^2} && (\Gamma(\alpha+1) = \alpha\,\Gamma(\alpha)) \\
&= \frac{\alpha(\alpha+1)}{\lambda^2} && \text{(cancel } \Gamma(\alpha)\text{)}
\end{aligned}
$$

Then, by the simplified expression for variance and Corollary 0.1:

$$
\begin{aligned}
\operatorname{Var}(T) &= \mathbb{E}[T^2] - (\mathbb{E}[T])^2 && \text{(simplified expression for variance)} \\
&= \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} && \text{(mean of the gamma distribution)} \\
&= \frac{\alpha^2+\alpha-\alpha^2}{\lambda^2} && \text{(expand } \alpha(\alpha+1) \text{ and combine fractions)} \\
&= \frac{\alpha}{\lambda^2} && (\alpha^2 - \alpha^2 = 0)
\end{aligned}
$$

□

Corollary 0.3 (The gamma distribution with shape 1 is the exponential)**.** *If* $T \sim \operatorname{Gam}(1,\lambda)$*, then* T *has the exponential distribution with rate* λ*.*

Proof

Proof. For $t > 0$:

$$
\begin{aligned}
f(t) &= \frac{\lambda^1}{\Gamma(1)}\, t^{1-1} \mathrm{e}^{-\lambda t} && \text{(gamma density with } \alpha = 1\text{)} \\
&= \lambda\, t^0 \mathrm{e}^{-\lambda t} && (\Gamma(1) = 1 \text{ and } 1-1=0) \\
&= \lambda \mathrm{e}^{-\lambda t} && (t^0 = 1)
\end{aligned}
$$

and $f(t) = 0$ for $t \le 0$. This density is the exponential density with rate λ. □

Weibull distribution

Definition 0.1 (Weibull distribution)**.** A non-negative random variable T has the **Weibull distribution** with shape $\alpha > 0$ and rate $\lambda > 0$ [6] if its survival function is:

$$\mathbb{S}(t) \stackrel{\text{def}}{=} \mathrm{e}^{-\lambda t^\alpha}, \quad t \geq 0$$

Theorem 0.1 (Weibull density and hazard)**.** *If T has the Weibull distribution with shape α and rate λ, then for $t > 0$:*

$$f(t) = \alpha\lambda t^{\alpha-1} e^{-\lambda t^\alpha}$$
$$\mathbb{h}(t) = \alpha\lambda t^{\alpha-1}$$

Proof

Proof. The CDF of T is $F(t) = 1 - \mathbb{S}(t)$ (survival function and CDF), which is 0 for $t < 0$ and $1 - \mathrm{e}^{-\lambda t^\alpha}$ for $t \geq 0$. This F is continuous everywhere, with a continuous derivative everywhere except possibly $t = 0$, so

$$\begin{aligned} f &= F' \\ &= -\mathbb{S}' \end{aligned}$$

is a density of T (a piecewise-smooth CDF has its derivative as a density). This density is continuous at every $t > 0$, so there the hazard is $f(t)/\mathbb{S}(t)$ (hazard equals density over survival):

$$\begin{aligned} f(t) &= -\frac{d}{dt}\mathrm{e}^{-\lambda t^\alpha} && (f = -\mathbb{S}') \\ &= \alpha\lambda t^{\alpha-1}\mathrm{e}^{-\lambda t^\alpha} && \text{(chain rule)} \\ \mathbb{h}(t) &= \frac{\alpha\lambda t^{\alpha-1}\mathrm{e}^{-\lambda t^\alpha}}{\mathrm{e}^{-\lambda t^\alpha}} && \text{(hazard is density over survival)} \\ &= \alpha\lambda t^{\alpha-1} && \text{(cancel)} \end{aligned}$$

□

Remark

Remark. The hazard is written $\mathbb{h}(t)$ here, rather than $\lambda(t)$, because the Weibull rate parameter is also called λ.

Theorem 0.2 (Weibull mean)**.** *If T has the Weibull distribution with shape α and rate λ, then:*

$$\mathbb{E}[T] = \Gamma(1 + 1/\alpha) \cdot \lambda^{-1/\alpha}$$

Proof

Proof. For the mean, use the survival-function formula for the mean and substitute $u = \lambda t^\alpha$, so that $t = (u/\lambda)^{1/\alpha}$ and $dt = \frac{1}{\alpha}\lambda^{-1/\alpha}u^{1/\alpha-1}\,du$:

$$\begin{aligned}
\mathbb{E}[T] &= \int_0^\infty \mathrm{e}^{-\lambda t^\alpha}\,dt && \text{(expectation via the survival function)} \\
&= \frac{1}{\alpha}\lambda^{-1/\alpha}\int_0^\infty u^{1/\alpha-1}\mathrm{e}^{-u}\,du && \text{(substitute } u = \lambda t^\alpha\text{)} \\
&= \frac{1}{\alpha}\lambda^{-1/\alpha}\,\Gamma(1/\alpha) && \text{(definition of the gamma function)} \\
&= \Gamma(1 + 1/\alpha)\,\lambda^{-1/\alpha} && (\Gamma(1+a) = a\,\Gamma(a))
\end{aligned}$$

□

Corollary 0.1 (The Weibull with shape 1 is the exponential)**.** *If T has the Weibull distribution with shape $\alpha = 1$ and rate λ, then T has the exponential distribution with rate λ.*

Proof

Proof. By Theorem 0.1 with $\alpha = 1$, for $t > 0$:

$$\begin{aligned}
f(t) &= \alpha\lambda t^{\alpha-1}\mathrm{e}^{-\lambda t^\alpha} && \text{(Weibull density)} \\
&= 1 \cdot \lambda t^0 \mathrm{e}^{-\lambda t^1} && \text{(substitute } \alpha = 1\text{)} \\
&= \lambda\mathrm{e}^{-\lambda t} && (t^0 = 1 \text{ and } t^1 = t)
\end{aligned}$$

and $f(t) = 0$ for $t < 0$, since T is non-negative. This density matches the exponential density at every $t \neq 0$. Changing a density at the single point $t = 0$ does not change its integral over any set, so T has the exponential distribution with rate λ. □

Corollary 0.2 (The Weibull shape sets the direction of the hazard)**.** *If T has the Weibull distribution with shape α and rate λ, then on $t > 0$ its hazard $\mathbb{h}(t)$ is:*

- *strictly increasing if $\alpha > 1$*
- *constant if $\alpha = 1$*

- *strictly decreasing if* $\alpha < 1$

Proof

Proof. By Theorem 0.1, $\mathbb{h}(t) = \alpha\lambda t^{\alpha-1}$ for $t > 0$, so:

$$\begin{aligned}\frac{d}{dt}\mathbb{h}(t) &= \frac{d}{dt}\alpha\lambda t^{\alpha-1} && \text{(Weibull hazard)}\\ &= \alpha(\alpha-1)\lambda t^{\alpha-2} && \text{(power rule)}\end{aligned}$$

For $t > 0$, the factors α, λ, and $t^{\alpha-2}$ are all positive, so $\frac{d}{dt}\mathbb{h}(t)$ has the sign of $\alpha - 1$: positive for $\alpha > 1$, zero for $\alpha = 1$, and negative for $\alpha < 1$. A function with a positive (negative) derivative on an interval is strictly increasing (decreasing) there, and one with a zero derivative is constant. □

Remark

Remark. The exponential distribution's hazard is constant, so the Weibull's choice of an increasing, constant, or decreasing hazard (Corollary 0.2) provides more flexibility than the exponential.

Example 0.1 (Exponential as a special case)**.** With $\alpha = 1$, Theorem 0.1 gives $\mathbb{h}(t) = \lambda$ and

$$\begin{aligned}\mathbb{E}[T] &= \Gamma(2)\lambda^{-1}\\ &= 1/\lambda,\end{aligned}$$

matching the exponential distribution's constant hazard and mean. With $\alpha = 2$ and $\lambda = 1$, $\mathbb{h}(t) = 2t$ increases with t, and

$$\begin{aligned}\mathbb{E}[T] &= \Gamma(3/2)\\ &= \sqrt{\pi}/2\\ &\approx 0.886.\end{aligned}$$

The multivariate normal distribution

Definition 0.1 (Multivariate normal distribution (multivariate Gaussian distribution))**.** A $p \times 1$ random vector $\underset{\sim}{X} = (X_1, \ldots, X_p)^\top$ has the **multivariate normal distribution** (or **multivariate Gaussian distribution**) with mean parameter $\mu \in \mathbb{R}^p$ and variance parameter $\mathbf{\Sigma}$, a $p \times p$ positive definite matrix, written $\underset{\sim}{X} \sim \mathrm{N}_p(\underset{\sim}{\mu}, \mathbf{\Sigma})$, if $\underset{\sim}{X}$ has joint density:

$$\mathbb{p}(\underset{\sim}{X} = \underset{\sim}{x}) \stackrel{\text{def}}{=} \frac{1}{(2\pi)^{p/2} \det(\mathbf{\Sigma})^{1/2}} \mathrm{e}^{-\frac{1}{2}(\underset{\sim}{x}-\underset{\sim}{\mu})^\top \mathbf{\Sigma}^{-1}(\underset{\sim}{x}-\underset{\sim}{\mu})}, \quad \underset{\sim}{x} \in \mathbb{R}^p$$

where $\det(\mathbf{\Sigma})$ is the determinant of $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$ is its inverse.

Example 0.1 (A bivariate Gaussian density)**.** Let $p = 2$, $\underset{\sim}{\mu} = \underset{\sim}{0}$, and $\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}$. Then $\det(\mathbf{\Sigma}) = 1 - 0.8^2 = 0.36$, so the density of $\underset{\sim}{X} \sim \mathrm{N}_2(\underset{\sim}{0}, \mathbf{\Sigma})$ (multivariate Gaussian) at the mean, which is also its peak, is:

$$\begin{aligned} \mathbb{p}(\underset{\sim}{0}) &= \frac{1}{2\pi \cdot 0.36^{1/2}} \mathrm{e}^0 && \text{(substitute into the multivariate normal density)} \\ &= \frac{1}{0.6 \cdot 2\pi} && \text{(simplify)} \end{aligned}$$

which is about 0.265. The density is a bell-shaped surface over the (x_1, x_2) plane (Figure 5). Because $\mathbf{\Sigma}_{12} = 0.8 > 0$, the surface is stretched along the line $x_2 = x_1$.

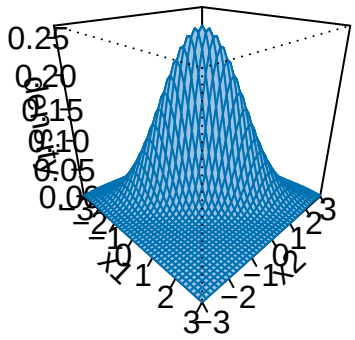


Figure 5: The density of $\mathrm{N}_2(\underset{\sim}{0}, \mathbf{\Sigma})$ in the example, with variances 1 and covariance 0.8. Its horizontal slices are ellipses stretched along $x_2 = x_1$.

Remark

Remark. A joint density of p random variables is the p-variable analogue of a joint density of two: a non-negative function on $\mathbb{R}^p$ whose integral over any region $A \subseteq \mathbb{R}^p$ is $\Pr(\underset{\sim}{X} \in A)$.
The density is well defined because $\mathbf{\Sigma}$ is positive definite:

- $\mathbf{\Sigma}^{-1}$ exists (positive definite matrices have positive definite inverses), and
- $\det(\mathbf{\Sigma}) > 0$ (positive definite matrices have positive determinants), so its square root is a positive number.

Theorem 0.1 (The multivariate normal density integrates to 1)**.** *If $\underset{\sim}{X} \sim \mathrm{N}_p(\underset{\sim}{\mu}, \mathbf{\Sigma})$ (Definition 0.1), then the density in Definition 0.1 integrates to 1 over $\mathbb{R}^p$.*

Theorem 0.2 (The mean and variance of the multivariate normal distribution)**.** *If $\underset{\sim}{X} \sim \mathrm{N}_p(\mu, \mathbf{\Sigma})$ (Definition 0.1), then $\mathbb{E}\underset{\sim}{X} = \mu$ (the expectation of a random vector), and $\mathrm{Var}(\underset{\sim}{X}) = \mathbf{\Sigma}$ (the variance of a random vector).*

Remark

Remark. The proof changes variables in a p-dimensional integral, which these notes do not develop; see [7]. As with the univariate normal distribution, these facts are where the parameters' names come from. They also show why $\mathbf{\Sigma}$ must at least be positive semidefinite; requiring it to be positive definite rules out cases like a variance matrix that is not positive definite, which has no joint density.

Example 0.2 (The univariate normal is the case $p = 1$)**.** With $p = 1$, $\mu = (\mu)$, and $\mathbf{\Sigma} = (\sigma^2)$ for $\sigma^2 > 0$, $\det(\mathbf{\Sigma}) = \sigma^2$ (determinant, $p = 1$) and $\mathbf{\Sigma}^{-1} = (1/\sigma^2)$, so:

$$\begin{aligned} \mathbb{p}(\underset{\sim}{X} = \underset{\sim}{x}) &= \frac{1}{(2\pi)^{1/2}(\sigma^2)^{1/2}} \mathrm{e}^{-\frac{1}{2}(x-\mu)\frac{1}{\sigma^2}(x-\mu)} && \text{(substitute into the multivariate normal density)} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} && \text{(simplify)} \end{aligned}$$

which is the normal density $\mathrm{N}(\mu, \sigma^2)$.

Lemma 0.1 (The quadratic form in the multivariate normal density is non-negative)**.** *If $\mathbf{\Sigma}$ is a $p \times p$ positive definite matrix and $\underset{\sim}{x}, \underset{\sim}{\mu} \in \mathbb{R}^p$, then*

$$(\underset{\sim}{x} - \underset{\sim}{\mu})^\top \mathbf{\Sigma}^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu}) \geq 0,$$

with equality if and only if $\underset{\sim}{x} = \underset{\sim}{\mu}$.

Proof

Proof. $\mathbf{\Sigma}^{-1}$ is positive definite (positive definite inverse). If $\underset{\sim}{x} \neq \mu$, then $\underset{\sim}{x} - \underset{\sim}{\mu} \neq \underset{\sim}{0}$, so the quadratic form is positive by the definition of positive definite. If $\underset{\sim}{x} = \underset{\sim}{\mu}$, then $\underset{\sim}{x} - \underset{\sim}{\mu} = \underset{\sim}{0}$, and the quadratic form is 0. □

Definition 0.2 (Mahalanobis distance)**.** For a $p \times p$ positive definite matrix $\mathbf{\Sigma}$ and $\underset{\sim}{x}, \underset{\sim}{\mu} \in \mathbb{R}^p$, the **Mahalanobis distance** from $\underset{\sim}{x}$ to $\underset{\sim}{\mu}$ with respect to $\mathbf{\Sigma}$ is:

$$\Delta(\underset{\sim}{x}) \stackrel{\text{def}}{=} \sqrt{(\underset{\sim}{x} - \underset{\sim}{\mu})^\top \mathbf{\Sigma}^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu})}$$

Remark

Remark. The square root is defined by Lemma 0.1. See also [8].

Corollary 0.1 (The multivariate normal density decreases with Mahalanobis distance)**.** *If* $\underset{\sim}{X} \sim \mathrm{N}_p(\underset{\sim}{\mu}, \mathbf{\Sigma})$*, then:*

$$\mathbb{p}(\underset{\sim}{X} = \underset{\sim}{x}) = \frac{1}{(2\pi)^{p/2} \det(\mathbf{\Sigma})^{1/2}} e^{-\frac{\Delta(\underset{\sim}{x})^2}{2}}$$

where Δ *is the Mahalanobis distance to* μ *with respect to* $\mathbf{\Sigma}$ *(Definition 0.2). So the density is highest at* $\underset{\sim}{x} = \underset{\sim}{\mu}$*, decreases as* $\Delta(\underset{\sim}{x})$ *increases, and is the same at every* $\underset{\sim}{x}$ *with the same* $\Delta(\underset{\sim}{x})$*.*

Proof

Proof. Substitute Definition 0.2 into Definition 0.1. The function $\delta \mapsto \mathrm{e}^{-\delta^2/2}$ decreases on $\delta \geq 0$, and by Lemma 0.1, $\Delta(\underset{\sim}{x}) = 0$ only at $\underset{\sim}{x} = \underset{\sim}{\mu}$. □

Example 0.3 (Mahalanobis distance for diagonal variance matrices)**.** If $\mathbf{\Sigma}$ is diagonal with diagonal elements $\sigma_1^2, \ldots, \sigma_p^2$, all positive, then $\mathbf{\Sigma}^{-1}$ is diagonal with diagonal elements $1/\sigma_1^2, \ldots, 1/\sigma_p^2$ (multiplying the two gives $\mathbf{I}_p$; matrix inverse), so:

$$\Delta(\underset{\sim}{x})^2 = \sum_{i=1}^{p} \frac{(x_i - \mu_i)^2}{\sigma_i^2}$$

Two special cases:

- If $\mathbf{\Sigma} = \mathbf{I}_p$, then $\Delta(\underset{\sim}{x})^2 = \sum_{i=1}^{p} (x_i - \mu_i)^2$, so the Mahalanobis distance is the ordinary (Euclidean) distance.
- If $\mathbf{\Sigma} = \sigma^2 \mathbf{I}_p$, then $\Delta(\underset{\sim}{x})$ is the Euclidean distance divided by σ.

In general, each coordinate's difference is measured in units of that coordinate's standard deviation.

Theorem 0.3 (A multivariate normal vector with diagonal variance has independent normal components)**.** *If* $\underset{\sim}{X} \sim \mathrm{N}_p(\underset{\sim}{\mu}, \mathbf{\Sigma})$ *and* $\mathbf{\Sigma}$ *is diagonal with diagonal elements* $\sigma_1^2, \ldots, \sigma_p^2$*, then:*

- *the joint density of* $\underset{\sim}{X}$ *is the product of the normal densities* $\mathrm{N}(\mu_i, \sigma_i^2)$*,* $i = 1, \ldots, p$*,*
- *each* $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$*, and*
- $X_1, \ldots, X_p$ *are independent.*

Proof

Proof. By the determinant of a diagonal matrix, $\det(\mathbf{\Sigma}) = \prod_{i=1}^{p} \sigma_i^2$, so $\det(\mathbf{\Sigma})^{1/2} = \prod_{i=1}^{p} \sigma_i$. Using Example 0.3 for the quadratic form:

$$\begin{aligned} \mathbb{p}(\underset{\sim}{X} = \underset{\sim}{x}) &= \frac{1}{(2\pi)^{p/2} \prod_{i=1}^{p} \sigma_i} \mathrm{e}^{-\frac{1}{2} \sum_{i=1}^{p} \frac{(x_i - \mu_i)^2}{\sigma_i^2}} && \text{(substitute)} \\ &= \prod_{i=1}^{p} \frac{1}{\sigma_i \sqrt{2\pi}} \mathrm{e}^{-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}} && (\mathrm{e}^{a+b} = \mathrm{e}^a \mathrm{e}^b) \end{aligned}$$

which is the product of the $\mathrm{N}(\mu_i, \sigma_i^2)$ densities. Integrating out every coordinate except x_i, each other factor integrates to 1 (the normal density integrates to 1), leaving the $\mathrm{N}(\mu_i, \sigma_i^2)$ density as the density of X_i (marginal density from a joint density, with p variables). So the joint density is the product of the marginal densities, and the components are independent by the factorization theorem for densities, extended to p variables as its remark describes. □

Corollary 0.2 (For a multivariate normal vector, uncorrelated components are independent)**.** *If $\underset{\sim}{X} \sim \mathrm{N}_p(\mu, \mathbf{\Sigma})$, then $X_1, \ldots, X_p$ are independent if and only if $\mathbf{\Sigma}$ is diagonal, that is, if and only if $\mathrm{Cov}(X_i, X_j) = 0$ for every $i \neq j$.*

Proof

Proof. By Theorem 0.2, $\mathrm{Var}(\underset{\sim}{X}) = \mathbf{\Sigma}$, whose (i, j)-th element is $\mathrm{Cov}(X_i, X_j)$ (elements of the variance matrix). If $\mathbf{\Sigma}$ is diagonal, the components are independent by Theorem 0.3. If the components are independent, then:

- each pair X_i, X_j with $i \neq j$ is independent,
- each X_i is continuous, with the density left after integrating out the other coordinates of the joint density (marginal density from a joint density, with p variables), and
- $\mathbb{E}[X_i^2] < \infty$, because $\mathrm{Var}(X_i)$, the (i, i)-th element of $\mathbf{\Sigma}$, is defined, and a variance is defined only when $\mathbb{E}[X_i^2] < \infty$.

So $\mathbf{\Sigma}$ is diagonal by independent components give a diagonal variance matrix. □

Remark

Remark. In general, uncorrelated random variables need not be independent (an example). Corollary 0.2 needs the whole vector to be multivariate normal: two random vari-

ables that are each normal and uncorrelated can still be dependent (https://en.wikipedia.org/wiki/Normally_distributed_and_uncorrelated_does_not_imply_independent).

Theorem 0.4 (The conditional distribution of a block of a multivariate Gaussian vector)**.** *Let* $\underset{\sim}{X}$ *have the multivariate normal distribution* $\mathrm{N}_p(\underset{\sim}{\mu}, \mathbf{\Sigma})$*, cut into* $\underset{\sim}{X} = \begin{bmatrix} \underset{\sim}{X}_1 \\ \underset{\sim}{X}_2 \end{bmatrix}$ *with* k *and* m *elements, and cut* $\underset{\sim}{\mu}$ *and* $\mathbf{\Sigma}$ *to match:*

$$\underset{\sim}{\mu} = \begin{bmatrix} \underset{\sim}{\mu}_1 \\ \underset{\sim}{\mu}_2 \end{bmatrix}, \quad \mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{bmatrix}.$$

Let $\mathbf{S} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}\mathbf{\Sigma}_{12}$*. Then, given* $\underset{\sim}{X}_1 = \underset{\sim}{x}_1$*,*

$$\underset{\sim}{X}_2 \mid \underset{\sim}{X}_1 = \underset{\sim}{x}_1 \sim \mathrm{N}_m\big(\underset{\sim}{\mu}_2 + \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}(\underset{\sim}{x}_1 - \underset{\sim}{\mu}_1), \mathbf{S}\big).$$

Proof

Proof. $\mathbf{\Sigma}$ is positive definite, so by the inverse of a positive definite block matrix, $\mathbf{S}$ is positive definite, and $(\mathbf{\Sigma})^{-1}$ has blocks $\mathbf{\Omega}_{ij}$ with

$$\mathbf{\Omega}_{22} = (\mathbf{S})^{-1}, \quad \mathbf{\Omega}_{21} = -(\mathbf{S})^{-1}\mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}.$$

Write $\underset{\sim}{d}_1 = \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1$ and $\underset{\sim}{d}_2 = \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2$. The conditional density of $\underset{\sim}{X}_2$ given $\underset{\sim}{X}_1 = \underset{\sim}{x}_1$ is the joint density divided by the marginal density of $\underset{\sim}{x}_1$, and the marginal density does not depend on $\underset{\sim}{x}_2$. So as a function of $\underset{\sim}{x}_2$, the conditional density equals the joint density times a factor that does not change with $\underset{\sim}{x}_2$, and the joint density is a factor that does not change with $\underset{\sim}{x}_2$ times e raised to

$$-\frac{1}{2}(\underset{\sim}{d}_1{}^{\top}\mathbf{\Omega}_{11}\underset{\sim}{d}_1 + 2\underset{\sim}{d}_2{}^{\top}\mathbf{\Omega}_{21}\underset{\sim}{d}_1 + \underset{\sim}{d}_2{}^{\top}\mathbf{\Omega}_{22}\underset{\sim}{d}_2).$$

The first summand does not involve $\underset{\sim}{x}_2$, so its contribution to the joint density is a factor that does not change with $\underset{\sim}{x}_2$. Substituting $\mathbf{\Omega}_{21}$ and $\mathbf{\Omega}_{22}$ into the other two summands:

$$\begin{aligned} 2\underset{\sim}{d}_2{}^{\top}\mathbf{\Omega}_{21}\underset{\sim}{d}_1 + \underset{\sim}{d}_2{}^{\top}\mathbf{\Omega}_{22}\underset{\sim}{d}_2 &= \underset{\sim}{d}_2{}^{\top}(\mathbf{S})^{-1}\underset{\sim}{d}_2 - 2\underset{\sim}{d}_2{}^{\top}(\mathbf{S})^{-1}\mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}\underset{\sim}{d}_1 && \text{(substitute)} \\ &= (\underset{\sim}{d}_2 - \underset{\sim}{c})^{\top}(\mathbf{S})^{-1}(\underset{\sim}{d}_2 - \underset{\sim}{c}) - \underset{\sim}{c}^{\top}(\mathbf{S})^{-1}\underset{\sim}{c} && \text{(complete the square, with } \underset{\sim}{c} = \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11}) \end{aligned}$$

The last summand does not involve $\underset{\sim}{x}_2$ either. Since $\underset{\sim}{d}_2 - \underset{\sim}{c} = \underset{\sim}{x}_2 - \big(\underset{\sim}{\mu}_2 + \underset{\sim}{c}\big)$, the conditional density of $\underset{\sim}{X}_2$ equals a factor that does not change with $\underset{\sim}{x}_2$, times

$$\mathrm{e}^{-\frac{1}{2}\left(\underset{\sim}{x}_2 - \left(\underset{\sim}{\mu}_2 + \underset{\sim}{c}\right)\right)^{\top}(\mathbf{S})^{-1}\left(\underset{\sim}{x}_2 - \left(\underset{\sim}{\mu}_2 + \underset{\sim}{c}\right)\right)},$$

which has the form of the $\mathrm{N}_m\big(\mu_2 + \underset{\sim}{c}, \mathbf{S}\big)$ density. A conditional density integrates to 1, so the factor that does not change with $\underset{\sim}{x}_2$ must be the normalizing factor of that density. □

Example 0.4 (The conditional distribution of one element of a bivariate Gaussian vector)**.**

Let $\underset{\sim}{X} = (X_1, X_2)^\top \sim \mathrm{N}_2\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $|\rho| < 1$. Here $\mathbf{\Sigma}_{11} = [1]$, $\mathbf{\Sigma}_{21} = [\rho]$ and $\mathbf{\Sigma}_{22} = [1]$, so $\mathbf{S} = [1 - \rho^2]$. By Theorem 0.4:

$$X_2 \mid X_1 = x_1 \sim \mathrm{N}(\rho x_1, 1 - \rho^2).$$

At $\rho = 0$ the conditional law of X_2 is $\mathrm{N}(0, 1)$ for every x_1, which is the marginal distribution of X_2, as it must be for independent elements (Theorem 0.3).

Theorem 0.5 (The marginal distribution of a subvector of a multivariate Gaussian vector)**.** *Let $\underset{\sim}{X} \sim \mathrm{N}_p(\mu, \mathbf{\Sigma})$, and let $\underset{\sim}{X}_J$ be the subvector of k elements of $\underset{\sim}{X}$ whose indices are in J. Then $\underset{\sim}{X}_J \sim \mathrm{N}_k\big(\underset{\sim}{\mu}_J, \mathbf{\Sigma}_{JJ}\big)$, where $\underset{\sim}{\mu}_J$ holds the elements of $\underset{\sim}{\mu}$ with indices in J, and $\mathbf{\Sigma}_{JJ}$ holds the rows and columns of $\mathbf{\Sigma}$ with indices in J.*

Proof

Proof. Reordering the elements of $\underset{\sim}{X}$ reorders the rows and columns of $\mathbf{\Sigma}$ in the same way, so we may take $\underset{\sim}{X}_J$ to be the first k elements. Cut $\underset{\sim}{X}$, μ and $\mathbf{\Sigma}$ into blocks as in Theorem 0.4, with $\underset{\sim}{X}_1 = \underset{\sim}{X}_J$ and $m = p - k$ elements in $\underset{\sim}{X}_2$, and let $\mathbf{S} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}\mathbf{\Sigma}_{12}$. Write $\mathbf{\Omega} = (\mathbf{\Sigma})^{-1}$ with blocks $\mathbf{\Omega}_{ij}$. By the inverse of a positive definite block matrix,

$$\mathbf{\Omega}_{22} = (\mathbf{S})^{-1}, \quad \mathbf{\Omega}_{21} = -(\mathbf{S})^{-1}\mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}.$$

The first block row of $\mathbf{\Sigma}\mathbf{\Omega} = \mathbf{I}$ gives $\mathbf{\Sigma}_{11}\mathbf{\Omega}_{11} + \mathbf{\Sigma}_{12}\mathbf{\Omega}_{21} = \mathbf{I}$, so, using $\mathbf{\Sigma}_{12} = \mathbf{\Sigma}_{21}{}^\top$,

$$\mathbf{\Omega}_{11} = (\mathbf{\Sigma}_{11})^{-1} - (\mathbf{\Sigma}_{11})^{-1}\mathbf{\Sigma}_{12}\mathbf{\Omega}_{21} = (\mathbf{\Sigma}_{11})^{-1} + (\mathbf{\Sigma}_{11})^{-1}\mathbf{\Sigma}_{12}(\mathbf{S})^{-1}\mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}.$$

Write $\underset{\sim}{d}_1 = \underset{\sim}{x}_1 - \underset{\sim}{\mu}_1$, $\underset{\sim}{d}_2 = \underset{\sim}{x}_2 - \underset{\sim}{\mu}_2$ and $\underset{\sim}{c} = \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}\underset{\sim}{d}_1$. The joint density in Definition 0.1 is a normalizing factor times e raised to $-1/2$ times

$$\begin{aligned}
&\underset{\sim}{d}_1{}^\top\mathbf{\Omega}_{11}\underset{\sim}{d}_1 + 2\underset{\sim}{d}_2{}^\top\mathbf{\Omega}_{21}\underset{\sim}{d}_1 + \underset{\sim}{d}_2{}^\top\mathbf{\Omega}_{22}\underset{\sim}{d}_2 \\
&= \underset{\sim}{d}_1{}^\top(\mathbf{\Sigma}_{11})^{-1}\underset{\sim}{d}_1 + \underset{\sim}{c}{}^\top(\mathbf{S})^{-1}\underset{\sim}{c} - 2\underset{\sim}{d}_2{}^\top(\mathbf{S})^{-1}\underset{\sim}{c} + \underset{\sim}{d}_2{}^\top(\mathbf{S})^{-1}\underset{\sim}{d}_2 \quad \text{(substitute } \mathbf{\Omega}_{11}, \mathbf{\Omega}_{21}, \mathbf{\Omega}_{22}\text{)} \\
&= \underset{\sim}{d}_1{}^\top(\mathbf{\Sigma}_{11})^{-1}\underset{\sim}{d}_1 + (\underset{\sim}{d}_2 - \underset{\sim}{c})^\top(\mathbf{S})^{-1}(\underset{\sim}{d}_2 - \underset{\sim}{c}). \quad \text{(complete the square)}
\end{aligned}$$

The marginal density of $\underset{\sim}{X}_1$ is the joint density integrated over $\underset{\sim}{x}_2$. The first summand does not involve $\underset{\sim}{x}_2$, and $\mathbf{S}$ is positive definite, so by Theorem 0.1 integrating the factor from the second summand over $\underset{\sim}{x}_2$ gives a number that does not depend on $\underset{\sim}{x}_1$. So the marginal density of $\underset{\sim}{X}_1$ equals a factor free of $\underset{\sim}{x}_1$ times

$$\mathrm{e}^{-\frac{1}{2}\underset{\sim}{d}_1{}^\top(\mathbf{\Sigma}_{11})^{-1}\underset{\sim}{d}_1},$$

which has the form of the $\mathrm{N}_k\left(\underset{\sim}{\mu}_1, \mathbf{\Sigma}_{11}\right)$ density. A marginal density integrates to 1, so that factor must be the normalizing factor of that density. □

Example 0.5 (The marginal distribution of the two end elements of a chain of three Gaussian variables)**.**
Let $\underset{\sim}{X} = (X_1, X_2, X_3)^\top$ be the Gaussian vector of Example 0.9, with $\mathbb{E}[\underset{\sim}{X}] = \mathbf{0}$ and

$$\mathrm{Var}(\underset{\sim}{X}) = \frac{1}{4}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix}.$$

Take $J = \{1, 3\}$. By Theorem 0.5, keeping the entries in positions 1 and 3,

$$(X_1, X_3)^\top \sim \mathrm{N}_2\left(\mathbf{0}, \frac{1}{4}\begin{pmatrix} 3 & 1 \\ 1 & 3 \end{pmatrix}\right).$$

So X_1 and X_3 each have variance $3/4$ and covariance $1/4$, and they are not independent, although they are conditionally independent given X_2.

Definition 0.3 (Conditional precision)**.** Let $\underset{\sim}{X}$ be a $p \times 1$ multivariate Gaussian vector, and let $\underset{\sim}{X}_{-i}$ be $\underset{\sim}{X}$ without its ith element. The **conditional precision** of X_i given $\underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}$ is the precision computed from the conditional variance of X_i given $\underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}$:

$$\tau(X_i \mid \underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}) \stackrel{\text{def}}{=} (\mathrm{Var}(X_i \mid \underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}))^{-1}.$$

Example 0.6 (The conditional precision of one element of a bivariate Gaussian vector)**.**

Let $\underset{\sim}{X} = (X_1, X_2)^\top \sim \mathrm{N}_2\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $|\rho| < 1$. By Example 0.4, X_1 given $X_2 = x_2$ has variance $1 - \rho^2$, so the conditional precision of X_1 given $X_2 = x_2$ is

$$\frac{1}{1 - \rho^2}.$$

Definition 0.4 (Coprecision)**.** Let $\underset{\sim}{X}$ be a $p \times 1$ multivariate Gaussian vector with $p \geq 2$, let $i \neq j$, and let $\underset{\sim}{X}_{-ij}$ be $\underset{\sim}{X}$ without its ith and jth elements. The **coprecision** of X_i and X_j given $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$ is defined through their conditional covariance and their conditional variances as

$$\frac{-\mathrm{Cov}(X_i, X_j \mid \underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij})}{\mathrm{Var}(X_i \mid \underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij})\mathrm{Var}(X_j \mid \underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}) - \left(\mathrm{Cov}(X_i, X_j \mid \underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij})\right)^2}.$$

Example 0.7 (The coprecision of a bivariate Gaussian vector)**.**
Let $\underset{\sim}{X} = (X_1, X_2)^\top \sim \mathrm{N}_2\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $|\rho| < 1$. With no other elements to condition on, the conditional variances of X_1 and X_2 are both 1 and their conditional covariance is ρ, so the coprecision of X_1 and X_2 is

$$\frac{-\rho}{1 \cdot 1 - \rho^2} = \frac{-\rho}{1 - \rho^2}.$$

Definition 0.5 (Precision matrix)**.** Let $\underset{\sim}{X}$ be a $p \times 1$ multivariate Gaussian vector. The **precision matrix** $\tau(\underset{\sim}{X})$ of $\underset{\sim}{X}$ is the $p \times p$ matrix whose ith diagonal element is the conditional precision of X_i given $\underset{\sim}{X}_{-i}$, and whose (i, j) element, for $i \neq j$, is the coprecision of X_i and X_j given $\underset{\sim}{X}_{-ij}$.

Theorem 0.6 (The precision matrix is the inverse of the variance matrix)**.** *If $\underset{\sim}{X}$ has the multivariate normal distribution $\mathrm{N}_p(\underset{\sim}{\mu}, \mathbf{\Sigma})$, then $\tau(\underset{\sim}{X}) = (\mathbf{\Sigma})^{-1}$.*

Proof

Proof. Let $\mathbf{\Omega} = (\mathbf{\Sigma})^{-1}$. Reordering the elements of $\underset{\sim}{X}$ reorders the rows and columns of $\mathbf{\Sigma}$ and of $\mathbf{\Omega}$ in the same way, so for each i and j we may take X_i and X_j to be the last two elements.
Cut $\mathbf{\Sigma}$ into blocks with the first $p - 2$ elements in the first block. By the inverse of a positive definite block matrix, the last 2×2 block of $\mathbf{\Omega}$ is $(\mathbf{S})^{-1}$, where $\mathbf{S}$ is the Schur complement of the first block of $\mathbf{\Sigma}$. By Theorem 0.4, $\mathbf{S}$ is the variance matrix of $(X_i, X_j)^\top$ given $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$, so the diagonal elements v_i and v_j of $\mathbf{S}$ are the conditional variances of X_i and X_j, and its off-diagonal element c_{ij} is their conditional covariance. The inverse of a 2×2 matrix gives the off-diagonal element of $\mathbf{\Omega}$:

$$\begin{aligned}\boldsymbol{\Omega}_{ij} &= \left((\mathbf{S})^{-1}\right)_{12} && \text{(last block of } \boldsymbol{\Omega}\text{)} \\ &= \frac{-c_{ij}}{v_i v_j - c_{ij}^2} && \text{(inverse of a } 2\times 2 \text{ matrix)}\end{aligned}$$

which is the coprecision of X_i and X_j.

Taking X_i alone as the last element, the last 1×1 block of $\boldsymbol{\Omega}$ is $(\mathbf{S})^{-1}$, where by Theorem 0.4 $\mathbf{S}$ is the conditional variance of X_i given $\underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}$. So $\boldsymbol{\Omega}_{ii}$ is the inverse of that conditional variance, which is the conditional precision of X_i. □

Example 0.8 (The precision matrix of a bivariate Gaussian vector)**.**

Let $\underset{\sim}{X} = (X_1, X_2)^\top \sim \mathrm{N}_2\left(\mathbf{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$ with $|\rho| < 1$. By Example 0.6 and Example 0.7, the conditional precision of each element is $1/(1-\rho^2)$, and the coprecision of X_1 and X_2 is $-\rho/(1-\rho^2)$, so

$$\tau(\underset{\sim}{X}) = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

This matrix is the inverse of $\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, as the precision matrix is the inverse of the variance matrix says, because

$$\begin{aligned}\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \cdot \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} &= \frac{1}{1-\rho^2}\begin{pmatrix} 1-\rho^2 & -\rho+\rho \\ \rho-\rho & -\rho^2+1 \end{pmatrix} \\ &= \mathbf{I}_2.\end{aligned}$$

The off-diagonal entries have the opposite sign of ρ. Each diagonal entry, $1/(1-\rho^2)$, is at least the precision $1/\mathrm{Var}(X_1) = 1$ of X_1 alone, and it grows indefinitely as $|\rho| \to 1$. If $\rho = 0$, then $\tau(\underset{\sim}{X})$ is the identity matrix, which is diagonal like $\mathrm{Var}(\underset{\sim}{X})$ (compare independent components give a diagonal variance matrix).

Theorem 0.7 (A zero off-diagonal element of the precision matrix means conditional independence)**.** *Let $\underset{\sim}{X} \sim \mathrm{N}_p(\mu, \boldsymbol{\Sigma})$ with $p \geq 2$, and let $i \neq j$. The (i,j) element of the precision matrix $\tau(\underset{\sim}{X})$ is zero if and only if X_i and X_j are conditionally independent given the continuous random vector $\underset{\sim}{X}_{-ij}$, that is, $\underset{\sim}{X}$ without its ith and jth elements.*

Proof

Proof. By Theorem 0.6, the (i,j) element of $\tau(\underset{\sim}{X})$ is the coprecision of X_i and X_j given $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$. Write v_i and v_j for the conditional variances of X_i and X_j given

$\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$, and c_{ij} for their conditional covariance. By Theorem 0.4, given $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$, $(X_i, X_j)^\top$ has a bivariate Gaussian distribution whose variance matrix $\mathbf{S}$ is positive definite, with diagonal elements v_i and v_j and off-diagonal element c_{ij}, and $\mathbf{S}$ does not depend on $\underset{\sim}{x}_{-ij}$. A positive definite matrix has a positive determinant, so $v_i v_j - c_{ij}^2 > 0$, and the coprecision $-c_{ij}/(v_i v_j - c_{ij}^2)$ is zero exactly when $c_{ij} = 0$.
If $c_{ij} = 0$, then $\mathbf{S}$ is diagonal, so by Corollary 0.2 X_i and X_j are independent given $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$, for every $\underset{\sim}{x}_{-ij}$. That is, the conditional density of $(X_i, X_j)^\top$ equals the conditional density of X_i times the conditional density of X_j, which is the factorization in the definition of conditional independence. Conversely, if X_i and X_j are conditionally independent, then they are independent given each $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$, so by Corollary 0.2 $c_{ij} = 0$. □

Example 0.9 (A chain of three Gaussian variables)**.**
Let $\underset{\sim}{X} = (X_1, X_2, X_3)^\top$ have $\mathbb{E}[\underset{\sim}{X}] = \mathbf{0}$ and precision matrix

$$\tau(\underset{\sim}{X}) = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

By Theorem 0.6, the variance matrix is the inverse of this matrix,

$$\mathrm{Var}(\underset{\sim}{X}) = \frac{1}{4}\begin{pmatrix} 3 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 3 \end{pmatrix},$$

which one can check by multiplying the two matrices to get $\mathbf{I}_3$. So $\mathrm{Cov}(X_1, X_3) = 1/4 \neq 0$, and X_1 and X_3 are not independent. But the $(1,3)$ element of the precision matrix is 0, so by Theorem 0.7, X_1 and X_3 are conditionally independent given X_2: once X_2 is known, X_1 carries no further information about X_3.

Theorem 0.8 (A zero off-diagonal element of the variance matrix means independence)**.** *Let $\underset{\sim}{X} \sim \mathrm{N}_p(\mu, \mathbf{\Sigma})$ with $p \geq 2$, and let $i \neq j$. The (i,j) element of $\mathbf{\Sigma}$ is zero if and only if X_i and X_j are independent.*

Proof

Proof. By Theorem 0.5, $(X_i, X_j)^\top$ has a bivariate Gaussian distribution, whose variance matrix has off-diagonal element $\mathrm{Cov}(X_i, X_j)$, which by the elements of the variance matrix is the (i,j) element of $\mathbf{\Sigma}$. By Corollary 0.2 with $p = 2$, X_i and X_j are independent if and only if this element is zero. □

Example 0.10 (Independent but not conditionally independent Gaussian variables)**.** Let $\underset{\sim}{X} = (X_1, X_2, X_3)^\top$ be a Gaussian vector with $\mathbb{E}[\underset{\sim}{X}] = \mathbf{0}$ and

$$\mathrm{Var}(\underset{\sim}{X}) = \begin{pmatrix} 1 & 0 & 1/2 \\ 0 & 1 & 1/2 \\ 1/2 & 1/2 & 1 \end{pmatrix}.$$

The $(1,2)$ element is 0, so by Theorem 0.8, X_1 and X_2 are independent. By Theorem 0.6, the precision matrix is the inverse of this matrix,

$$\tau(\underset{\sim}{X}) = \begin{pmatrix} 3/2 & 1/2 & -1 \\ 1/2 & 3/2 & -1 \\ -1 & -1 & 2 \end{pmatrix},$$

which one can check by multiplying the two matrices to get $\mathbf{I}_3$. Its $(1,2)$ element is $1/2 \neq 0$, so by Theorem 0.7, X_1 and X_2 are not conditionally independent given X_3. So a zero in the variance matrix does not give a zero in the precision matrix. Example 0.9 shows the reverse case: a zero in the precision matrix, with a nonzero covariance.

Definition 0.6 (Partial correlation)**.** Let $\underset{\sim}{X}$ be a $p \times 1$ multivariate Gaussian vector with $p \geq 2$, let $i \neq j$, and let $\underset{\sim}{X}_{-ij}$ be $\underset{\sim}{X}$ without its ith and jth elements. The **partial correlation** of X_i and X_j given $\underset{\sim}{X}_{-ij}$ is their conditional covariance given $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$ for a fixed value $\underset{\sim}{x}_{-ij}$, divided by the square root of the conditional variance of X_i times that of X_j:

$$\frac{\mathrm{Cov}\left(X_i, X_j \mid \underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}\right)}{\sqrt{\mathrm{Var}\left(X_i \mid \underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}\right)\mathrm{Var}\left(X_j \mid \underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}\right)}}.$$

Theorem 0.9 (The partial correlation from the precision matrix)**.** *Let $\underset{\sim}{X} \sim \mathrm{N}_p(\mu, \mathbf{\Sigma})$ with $p \geq 2$, and let $i \neq j$. Then the partial correlation of X_i and X_j given $\underset{\sim}{X}_{-ij}$ is*

$$\frac{-\tau(\underset{\sim}{X})_{ij}}{\sqrt{\tau(\underset{\sim}{X})_{ii}\tau(\underset{\sim}{X})_{jj}}},$$

where $\tau(\underset{\sim}{X})$ is the precision matrix of $\underset{\sim}{X}$.

Proof

Proof. Reordering the elements of $\underset{\sim}{X}$ reorders the rows and columns of $\mathbf{\Sigma}$ and of $\tau(\underset{\sim}{X})$ in the same way, so we may take X_i and X_j to be the last two elements. Cut $\mathbf{\Sigma}$ into blocks with the first $p-2$ elements in the first block. By Theorem 0.6, $\tau(\underset{\sim}{X}) = (\mathbf{\Sigma})^{-1}$,

and by the inverse of a positive definite block matrix, the last 2×2 block of $\tau(\underset{\sim}{X})$ is $(\mathbf{S})^{-1}$, where $\mathbf{S}$ is the Schur complement of the first block of $\mathbf{\Sigma}$. By Theorem 0.4, $\mathbf{S}$ is the variance matrix of $(X_i, X_j)^\top$ given $\underset{\sim}{X}_{-ij} = \underset{\sim}{x}_{-ij}$, so its diagonal elements v_i and v_j are the conditional variances of X_i and X_j, and its off-diagonal element c_{ij} is their conditional covariance. Let $d = v_i v_j - c_{ij}^2$, the determinant of $\mathbf{S}$. The inverse of a 2×2 matrix gives

$$(\mathbf{S})^{-1} = \frac{1}{d} \begin{pmatrix} v_j & -c_{ij} \\ -c_{ij} & v_i \end{pmatrix},$$

so $\tau(\underset{\sim}{X})_{ii} = v_j/d$, $\tau(\underset{\sim}{X})_{jj} = v_i/d$ and $\tau(\underset{\sim}{X})_{ij} = -c_{ij}/d$. Since $\mathbf{S}$ is positive definite, $d > 0$, so

$$\begin{aligned} \frac{-\tau(\underset{\sim}{X})_{ij}}{\sqrt{\tau(\underset{\sim}{X})_{ii}\tau(\underset{\sim}{X})_{jj}}} &= \frac{c_{ij}/d}{\sqrt{v_i v_j}/d} && \text{(substitute)} \\ &= \frac{c_{ij}}{\sqrt{v_i v_j}}, && \text{(cancel } d > 0\text{)} \end{aligned}$$

which is the partial correlation of X_i and X_j. □

Example 0.11 (The partial correlations of a chain of three Gaussian variables)**.** Let $\underset{\sim}{X} = (X_1, X_2, X_3)^\top$ be the Gaussian vector of Example 0.9, with

$$\tau(\underset{\sim}{X}) = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

By Theorem 0.9, the partial correlation of X_1 and X_2 given X_3 is

$$\frac{-\tau(\underset{\sim}{X})_{12}}{\sqrt{\tau(\underset{\sim}{X})_{11}\tau(\underset{\sim}{X})_{22}}} = \frac{1}{\sqrt{2 \cdot 2}} = \frac{1}{2},$$

and the partial correlation of X_1 and X_3 given X_2 is $-0/\sqrt{2 \cdot 2} = 0$, as Theorem 0.7 says for a zero element, although their correlation is not zero.

Theorem 0.10 (The conditional mean of one element from the precision matrix)**.** *Let $\underset{\sim}{X} \sim \mathrm{N}_p(\mu, \mathbf{\Sigma})$ with $p \geq 2$, let $\underset{\sim}{X}_{-i}$ be $\underset{\sim}{X}$ without its ith element, and let $\underset{\sim}{x}_{-i}$ be a fixed value of $\underset{\sim}{X}_{-i}$. Then*

$$\mathbb{E}[X_i \mid \underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}] = \mu_i - \sum_{j \neq i} \frac{\tau(\underset{\sim}{X})_{ij}}{\tau(\underset{\sim}{X})_{ii}} (x_j - \mu_j),$$

and $\mathrm{Var}(X_i \mid \underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}) = 1/\tau(\underset{\sim}{X})_{ii}$*, where $\tau(\underset{\sim}{X})$ is the precision matrix of $\underset{\sim}{X}$.*

Proof

Proof. Reordering the elements of $\underset{\sim}{X}$ reorders the rows and columns of $\mathbf{\Sigma}$ and of $\tau(\underset{\sim}{X})$ in the same way, so we may take X_i to be the last element. Cut $\underset{\sim}{X}$, μ and $\mathbf{\Sigma}$ into blocks with $\underset{\sim}{X}_{-i}$ first and X_i last, as in Theorem 0.4 with $\underset{\sim}{X}_1 = \underset{\sim}{X}_{-i}$ and $\underset{\sim}{X}_2 = X_i$, and let $\mathbf{S} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}\mathbf{\Sigma}_{12}$. By Theorem 0.4, the conditional mean and conditional variance of X_i given $\underset{\sim}{X}_{-i} = \underset{\sim}{x}_{-i}$ are

$$\mu_i + \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}\big(\underset{\sim}{x}_{-i} - \underset{\sim}{\mu}_{-i}\big) \quad \text{and} \quad \mathbf{S}.$$

By Theorem 0.6, $\tau(\underset{\sim}{X}) = (\mathbf{\Sigma})^{-1}$, and by the inverse of a positive definite block matrix, its blocks $\mathbf{\Omega}_{21}$ and $\mathbf{\Omega}_{22}$ satisfy

$$\mathbf{\Omega}_{22} = (\mathbf{S})^{-1}, \quad \mathbf{\Omega}_{21} = -(\mathbf{S})^{-1}\mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1}.$$

Here $\mathbf{\Omega}_{22} = \tau(\underset{\sim}{X})_{ii}$ is a scalar, so the conditional variance is $\mathbf{S} = 1/\tau(\underset{\sim}{X})_{ii}$, and multiplying the second equation by $\mathbf{S}$ gives

$$\begin{aligned} \mathbf{\Sigma}_{21}(\mathbf{\Sigma}_{11})^{-1} &= -\mathbf{S}\mathbf{\Omega}_{21} && \text{(multiply by } \mathbf{S}) \\ &= -\frac{\mathbf{\Omega}_{21}}{\tau(\underset{\sim}{X})_{ii}}. && (\mathbf{S} = 1/\tau(\underset{\sim}{X})_{ii}) \end{aligned}$$

The elements of $\mathbf{\Omega}_{21}$ are $\tau(\underset{\sim}{X})_{ij}$ for $j \neq i$, so the conditional mean is

$$\mu_i - \sum_{j \neq i} \frac{\tau(\underset{\sim}{X})_{ij}}{\tau(\underset{\sim}{X})_{ii}}(x_j - \mu_j).$$

□

Example 0.12 (The conditional mean and variance of the middle element of a chain of three Gaussian variables)**.**
Let $\underset{\sim}{X} = (X_1, X_2, X_3)^\top$ be the Gaussian vector of Example 0.9, with $\mathbb{E}[\underset{\sim}{X}] = \mathbf{0}$ and

$$\tau(\underset{\sim}{X}) = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

By Theorem 0.10 with $i = 2$, using the second row of the precision matrix,

$$\mathbb{E}[X_2 \mid X_1 = x_1, X_3 = x_3] = -\frac{-1}{2}x_1 - \frac{-1}{2}x_3 = \frac{x_1 + x_3}{2},$$

and $\mathrm{Var}(X_2 \mid X_1 = x_1, X_3 = x_3) = 1/2$. So the conditional mean of X_2 is the midpoint of its two neighbors' values. The coefficient of x_1 is $-\tau(\underset{\sim}{X})_{21}/\tau(\underset{\sim}{X})_{22} = 1/2$, and the coefficient of a neighbor with a zero off-diagonal element would be zero, as Theorem 0.7

says.

Theorem 0.11 (Mahalanobis distance in eigenvector coordinates)**.** *Let* $\mathbf{\Sigma}$ *be positive definite, with eigendecomposition* $\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$*, where* $\mathbf{Q}$ *has columns* $\underset{\sim}{q}_1, \ldots, \underset{\sim}{q}_p$ *and* $\mathbf{\Lambda}$ *has diagonal elements* $\lambda_1, \ldots, \lambda_p$*. Then:*

$$\Delta(\underset{\sim}{x})^2 = \sum_{i=1}^{p} \frac{(\underset{\sim}{q}_i^\top(\underset{\sim}{x} - \underset{\sim}{\mu}))^2}{\lambda_i}$$

Proof

Proof. Let $\underset{\sim}{y} = \mathbf{Q}^\top(\underset{\sim}{x} - \underset{\sim}{\mu})$, whose i-th element is $y_i = \underset{\sim}{q}_i^\top(\underset{\sim}{x} - \underset{\sim}{\mu})$. By the inverse of a positive definite matrix, $\mathbf{\Sigma}^{-1} = \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top$, with every $\lambda_i > 0$. So:

$$\begin{aligned}
\Delta(\underset{\sim}{x})^2 &= (\underset{\sim}{x} - \underset{\sim}{\mu})^\top \mathbf{Q}\mathbf{\Lambda}^{-1}\mathbf{Q}^\top(\underset{\sim}{x} - \underset{\sim}{\mu}) && \text{(definition; substitute } \mathbf{\Sigma}^{-1}\text{)} \\
&= \underset{\sim}{y}^\top \mathbf{\Lambda}^{-1}\underset{\sim}{y} && ((\underset{\sim}{x} - \underset{\sim}{\mu})^\top \mathbf{Q} = \underset{\sim}{y}^\top) \\
&= \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i} && (\mathbf{\Lambda}^{-1} \text{ is diagonal})
\end{aligned}$$

□

Remark

Remark. y_i is the coordinate of $\underset{\sim}{x} - \underset{\sim}{\mu}$ along the eigenvector $\underset{\sim}{q}_i$. So the set of points at Mahalanobis distance δ from $\underset{\sim}{\mu}$, where the multivariate normal density is constant (Corollary 0.1), is an ellipse ($p = 2$) or ellipsoid centered at $\underset{\sim}{\mu}$, with axes along the eigenvectors of $\mathbf{\Sigma}$ and half-length $\delta\sqrt{\lambda_i}$ along $\underset{\sim}{q}_i$.

Example 0.13 (Elliptical contours of a bivariate normal density)**.** Let $\underset{\sim}{\mu} = \underset{\sim}{0}$ and $\mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, which has eigenvalues 3 and 1, with eigenvectors $\underset{\sim}{q}_1 = \frac{1}{\sqrt{2}}(1,1)^\top$ and $\underset{\sim}{q}_2 = \frac{1}{\sqrt{2}}(1,-1)^\top$ (an eigendecomposition example). By Theorem 0.11:

$$\Delta(\underset{\sim}{x})^2 = \frac{(x_1 + x_2)^2}{2 \cdot 3} + \frac{(x_1 - x_2)^2}{2 \cdot 1}$$

The point

$$\begin{aligned}
\underset{\sim}{x} &= \sqrt{3}\,\underset{\sim}{q}_1 \\
&= \sqrt{3/2}\,(1,1)^\top
\end{aligned}$$

has $x_1 + x_2 = \sqrt{6}$ and $x_1 - x_2 = 0$, so

$$\begin{aligned}\Delta(\underset{\sim}{x})^2 &= 6/6 \\ &= 1;\end{aligned}$$

the point $\underset{\sim}{x} = \underset{\sim}{q}_2$ has $x_1 + x_2 = 0$ and $x_1 - x_2 = \sqrt{2}$, so

$$\begin{aligned}\Delta(\underset{\sim}{x})^2 &= 2/2 \\ &= 1.\end{aligned}$$

So the contour $\Delta(\underset{\sim}{x}) = 1$ of the $\mathrm{N}_2(\underset{\sim}{0}, \mathbf{\Sigma})$ density is an ellipse stretched along $(1, 1)^\top$, with half-length $\sqrt{3}$, and squeezed along $(1, -1)^\top$, with half-length 1: X_1 and X_2 are positively correlated (Figure 6).

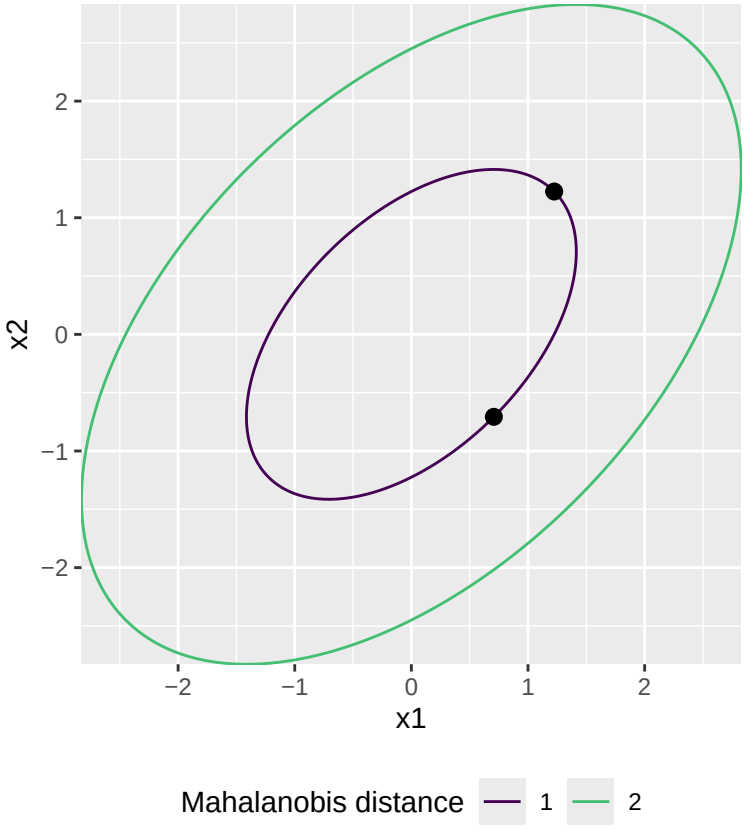


Figure 6: The contours of the Mahalanobis distance at 1 and 2 for the example. The dots are the points $\sqrt{3}\,\underset{\sim}{q}_1$ and $\underset{\sim}{q}_2$ of the example on the inner ellipse.

(This 3D surface is interactive; see the HTML or slide version of this page to view it.)

Figure 7: The Mahalanobis distance is a surface over the plane, a cone whose horizontal slices are the ellipses of Figure 6. The red lines are the contours at 1 and 2, and the dots are the points $\sqrt{3}\,\underset{\sim}{q}_1$ and $\underset{\sim}{q}_2$ of the example on the inner one.

Theorem 0.12 (The maximum of the Gaussian density)**.** *If $X \sim \mathrm{N}(\mu, \sigma^2)$ (Gaussian distribution), then its density is largest at $x = \mu$, where it equals:*

$$\max_x \mathbb{p}(X = x) = \frac{1}{\sigma\sqrt{2\pi}}$$

Proof

Proof. For each $x \in \mathbb{R}$, the quantity $u(x) \stackrel{\text{def}}{=} (x-\mu)^2/(2\sigma^2)$ is at least 0, and it equals 0 only at $x = \mu$. The factor $\mathrm{e}^{-u(x)}$ of the density is at most 1:

$$\begin{aligned} -u(x) &\leq 0 && \text{(a square is non-negative, and } 2\sigma^2 > 0) \\ \mathrm{e}^{-u(x)} &\leq \mathrm{e}^0 && (\mathrm{e}^t \text{ is increasing in } t) \\ &= 1 && (\mathrm{e}^0 = 1) \end{aligned}$$

Multiplying both sides by $1/(\sigma\sqrt{2\pi}) > 0$ keeps their order:

$$\begin{aligned} \mathbb{p}(X = x) &= \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-u(x)} && \text{(Gaussian density)} \\ &\leq \frac{1}{\sigma\sqrt{2\pi}} && (\mathrm{e}^{-u(x)} \leq 1) \end{aligned}$$

Equality holds at $x = \mu$, where $u(x) = 0$, so the density peaks there. □

Example 0.14 (The peak of the Gaussian density grows as the standard deviation shrinks)**.** By Theorem 0.12, the peak of the density of $\mathrm{N}(0, \sigma^2)$ is $1/(\sigma\sqrt{2\pi})$. Figure 8 marks each peak for standard deviations 0.5, 1 and 2.

```r
sds <- c(0.5, 1, 2)
x_grid <- seq(-6, 6, length.out = 400)
sd_label <- latex2exp::TeX("$\\sigma$")
density_label <- latex2exp::TeX("$f(x)$")

normal_curves <-
  dplyr::tibble(sd = sds) |>
  dplyr::reframe(.by = sd, x = x_grid) |>
  dplyr::mutate(
    density = dnorm(x, mean = 0, sd = sd),
    sd = factor(sd)
  )

normal_peaks <-
  dplyr::tibble(sd = sds) |>
  dplyr::mutate(
    peak = 1 / (sd * sqrt(2 * pi)),
    sd = factor(sd)
  )

normal_plot <-
  normal_curves |>
  ggplot2::ggplot() +
  ggplot2::aes(x = x, y = density, col = sd) +
  ggplot2::geom_line() +
  ggplot2::geom_point(
    data = normal_peaks,
    ggplot2::aes(x = 0, y = peak)
  ) +
  ggplot2::labs(col = sd_label, y = density_label) +
  ggplot2::theme(legend.position = "bottom")

print(normal_plot)
```

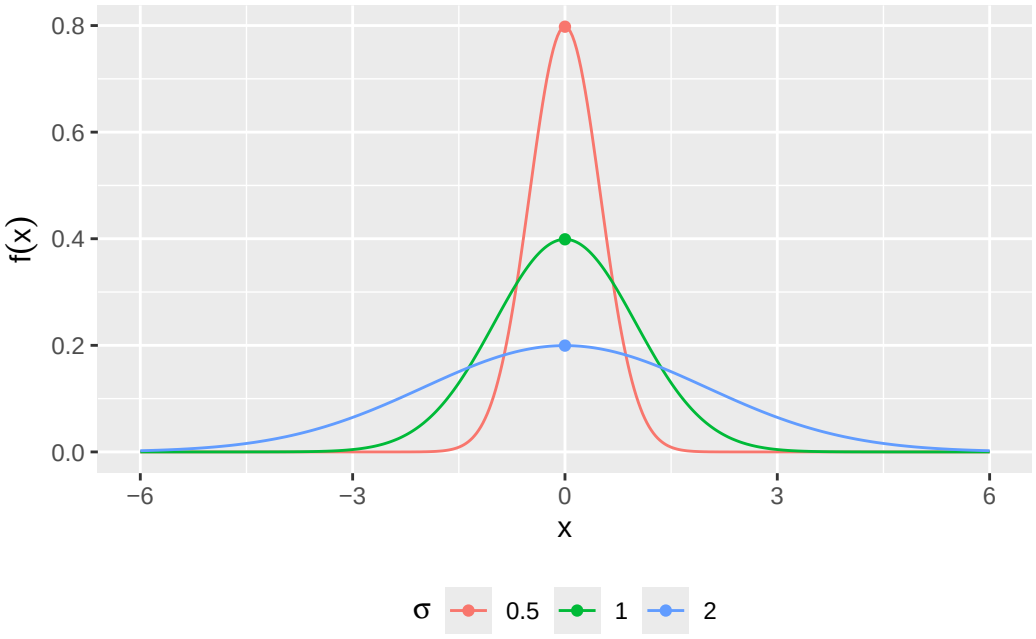


Figure 8: Gaussian densities, with each peak marked

As $\sigma \to 0$, the peak $1/(\sigma\sqrt{2\pi})$ grows indefinitely, so a density can exceed 1.

Theorem 0.13 (Gaussian distributions are closed under affine transformations)**.** *If* $X \sim \mathrm{N}(\mu, \sigma^2)$ *and* $a \neq 0$*, then* $aX + b \sim \mathrm{N}(a\mu + b, a^2\sigma^2)$.

Proof

Proof. Let $Y = aX+b$. Let f be the density of X. The CDF of Y is $\mathbb{F}_Y(y) = \Pr(aX+b \leq y)$. The density f is continuous everywhere. So at every y, the derivative of $\mathbb{F}_Y(y)$ is the density of Y, by the derivative of the CDF is the density. To differentiate $\mathbb{F}_Y(y)$, write it through the CDF $\mathbb{F}_X$ of X. There are two cases, by the sign of a.

- **Case** $a > 0$**.** Dividing by $a > 0$ leaves $\leq$ as $\leq$, so:

$$\begin{aligned} \mathbb{F}_Y(y) &= \Pr\left(X \leq \frac{y-b}{a}\right) && (aX + b \leq y \iff X \leq (y-b)/a) \\ &= \mathbb{F}_X\left(\frac{y-b}{a}\right) && \text{(definition of the CDF)} \end{aligned}$$

  The chain rule gives the derivative:

$$\frac{\partial}{\partial y}\mathbb{F}_Y(y) = \frac{1}{a}f\left(\frac{y-b}{a}\right) = \frac{1}{|a|}f\left(\frac{y-b}{a}\right) \quad (|a| = a)$$

- **Case** $a < 0$**.** Dividing by $a < 0$ turns $\leq$ into $\geq$, and $\Pr(X = x) = 0$ for a continuous random variable X, so:

$$\begin{aligned}\mathbb{F}_Y(y) &= \Pr\left(X \geq \frac{y-b}{a}\right) && (aX + b \leq y \iff X \geq (y-b)/a)\\ &= 1 - \mathbb{F}_X\left(\frac{y-b}{a}\right) && (\text{complement, and } \Pr(X = (y-b)/a) = 0)\end{aligned}$$

  The chain rule gives the derivative:

$$\frac{\partial}{\partial y}\mathbb{F}_Y(y) = -\frac{1}{a}f\left(\frac{y-b}{a}\right) = \frac{1}{|a|}f\left(\frac{y-b}{a}\right) \quad (|a| = -a)$$

In both cases, Y has density $\mathbb{p}(Y = y) = f((y-b)/a)/|a|$. Substitute the Gaussian density of X:

$$\begin{aligned}\mathbb{p}(Y = y) &= \frac{1}{|a|} \cdot \frac{1}{\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{((y-b)/a-\mu)^2}{2\sigma^2}} && (\text{Gaussian density of } X)\\ &= \frac{1}{|a|\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{((y-b-a\mu)/a)^2}{2\sigma^2}} && (\text{common denominator } a)\\ &= \frac{1}{|a|\sigma\sqrt{2\pi}}\mathrm{e}^{-\frac{(y-(a\mu+b))^2}{2a^2\sigma^2}} && ((u/a)^2 = u^2/a^2)\\ &= \frac{1}{\sqrt{a^2\sigma^2}\sqrt{2\pi}}\mathrm{e}^{-\frac{(y-(a\mu+b))^2}{2a^2\sigma^2}} && (|a|\sigma = \sqrt{a^2\sigma^2})\end{aligned}$$

The last line is the density of $\mathrm{N}(a\mu + b, a^2\sigma^2)$ because $a^2\sigma^2 > 0$. □

Example 0.15 (Standardizing a Gaussian random variable)**.**
Let $X \sim \mathrm{N}(\mu, \sigma^2)$ with $\sigma > 0$, and let $Z = (X - \mu)/\sigma$. Then $Z = aX + b$ with $a = 1/\sigma$ and $b = -\mu/\sigma$, and $a \neq 0$. By Theorem 0.13, Z is Gaussian with parameters:

$$\begin{aligned}a\mu + b &= \frac{\mu}{\sigma} - \frac{\mu}{\sigma} && (\text{substitute } a \text{ and } b)\\ &= 0 && (u - u = 0)\\ a^2\sigma^2 &= \frac{\sigma^2}{\sigma^2} && (a^2 = 1/\sigma^2)\\ &= 1 && (u/u = 1)\end{aligned}$$

So $Z \sim \mathrm{N}(0, 1)$.

Mixture distributions

Definition 0.1 (Mixture density)**.** Let $\mathbb{p}_1, \dots, \mathbb{p}_K$ be densities on $\mathbb{R}$, and let $w_1, \dots, w_K$ be numbers with:

- $w_c \geq 0$ for every c, and
- $\sum_{c=1}^{K} w_c = 1$.

The **mixture** of $\mathbb{p}_1, \dots, \mathbb{p}_K$ with **mixing weights** $w_1, \dots, w_K$ is the function [9]:

$$\mathbb{p}_{\text{mix}}(x) \stackrel{\text{def}}{=} \sum_{c=1}^{K} w_c \, \mathbb{p}_c(x), \quad x \in \mathbb{R}$$

Definition 0.2 (Mixture component)**.** The densities $\mathbb{p}_1, \dots, \mathbb{p}_K$ in a mixture (Definition 0.1) are its **components**.

Theorem 0.1 (A mixture is the density of a two-stage random variable)**.** *Let C be a discrete random variable and X a continuous random variable with joint density-mass function*

$$\mathbb{p}(C = c, X = x) = w_c \, \mathbb{p}_c(x), \quad c \in \{1, \dots, K\},$$

with $\mathbb{p}_c$ and w_c as in Definition 0.1. Then:

- $\mathbb{P}(C = c) = w_c$,
- *for each c with $w_c > 0$, the conditional density of X given $C = c$ is $\mathbb{p}_c$, and*
- *the mixture $\mathbb{p}_{mix}$ is a density of X.*

Proof

Proof. By the marginal PMF from a joint density-mass function,

$$\begin{aligned} \mathbb{P}(C = c) &= \int_{-\infty}^{\infty} w_c \, \mathbb{p}_c(x) \, dx \\ &= w_c, \end{aligned}$$

because $\mathbb{p}_c$ is a density and so integrates to 1. Then, by the definition of the conditional density,

$$\begin{aligned} \mathbb{p}(X = x \mid C = c) &= w_c \, \mathbb{p}_c(x) / w_c \\ &= \mathbb{p}_c(x). \end{aligned}$$

For the last claim, the events $\{C = 1\}, \ldots, \{C = K\}$ are mutually exclusive, and their union is the whole sample space. So for any interval B:

$$\begin{aligned}
\Pr(X \in B) &= \sum_{c=1}^{K} \Pr(C = c,\, X \in B) && \text{(additivity over the events } \{C = c\}) \\
&= \sum_{c=1}^{K} \int_B w_c\, \mathbb{p}_c(x)\, dx && \text{(definition of a joint density-mass function)} \\
&= \int_B \sum_{c=1}^{K} w_c\, \mathbb{p}_c(x)\, dx && \text{(a finite sum of integrals is the integral of the sum)} \\
&= \int_B \mathbb{p}_{\text{mix}}(x)\, dx && \text{(definition of a mixture)}
\end{aligned}$$

And $\mathbb{p}_{\text{mix}} \geq 0$, as a sum of non-negative terms, so $\mathbb{p}_{\text{mix}}$ is a density of X. □

i Remark

Remark. Theorem 0.1 describes how to generate a draw from a mixture: first draw a component label C with $\mathbb{P}(C = c) = w_c$, then draw X from the component density $\mathbb{p}_C$. In applications, C is often an unobserved subpopulation, such as a disease subtype. Mixtures of multivariate densities, such as mixtures of multivariate normal densities (Gaussian mixture models), work the same way, with p-variable densities in place of densities on $\mathbb{R}$.

Theorem 0.2 (The mean of a mixture is the weighted mean of the component means)**.** *If X has density $\mathbb{p}_{mix}$ (Definition 0.1), and each component $\mathbb{p}_c$ has a defined mean $\mu_c = \int_{-\infty}^{\infty} x\, \mathbb{p}_c(x)\, dx$, then:*

$$\mathbb{E}[X] = \sum_{c=1}^{K} w_c\, \mu_c$$

i Proof

Proof. The integral defining $\mathbb{E}[X]$ converges absolutely, because $\int_{-\infty}^{\infty} |x|\, \mathbb{p}_{\text{mix}}(x)\, dx = \sum_{c=1}^{K} w_c \int_{-\infty}^{\infty} |x|\, \mathbb{p}_c(x)\, dx$, a finite sum of finite numbers. Then:

$$
\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x\, \mathbb{p}_{\text{mix}}(x)\, dx && \text{(definition of expectation)} \\
&= \int_{-\infty}^{\infty} x \sum_{c=1}^{K} w_c\, \mathbb{p}_c(x)\, dx && \text{(definition of a mixture)} \\
&= \sum_{c=1}^{K} w_c \int_{-\infty}^{\infty} x\, \mathbb{p}_c(x)\, dx && \text{(a finite sum of integrals is the integral of the sum)} \\
&= \sum_{c=1}^{K} w_c\, \mu_c && \text{(definition of } \mu_c\text{)}
\end{aligned}
$$

The first step is the definition of expectation. □

Theorem 0.3 (Which component did an observation come from?)**.** *Under the conditions of Theorem 0.1, for each x with $\mathbb{p}_{mix}(x) > 0$:*

$$
\mathbb{P}(C = c \mid X = x) = \frac{w_c\, \mathbb{p}_c(x)}{\sum_{k=1}^{K} w_k\, \mathbb{p}_k(x)}
$$

Proof

Proof. By Theorem 0.1, $\mathbb{p}_{\text{mix}}$ is a density of X. By the definition of the conditional PMF (the case X continuous, C discrete):

$$
\begin{aligned}
\mathbb{P}(C = c \mid X = x) &= \frac{\mathbb{p}(C = c,\, X = x)}{\mathbb{p}(X = x)} && \text{(definition of the conditional PMF)} \\
&= \frac{w_c\, \mathbb{p}_c(x)}{\sum_{k=1}^{K} w_k\, \mathbb{p}_k(x)} && \text{(substitute the joint density-mass function and } \mathbb{p}_{\text{mix}}\text{)}
\end{aligned}
$$

□

Remark

Remark. Theorem 0.3 has the form of Bayes' theorem, with densities in place of some probabilities: the prior probability w_c of component c is multiplied by the component density $\mathbb{p}_c(x)$ at the observation and divided by the overall density $\mathbb{p}_{\text{mix}}(x)$, which plays the role of the law of total probability.

Definition 0.3 (Responsibility of a mixture component)**.** In the setting of Theorem 0.3, the conditional probability $\mathbb{P}(C = c \mid X = x)$ is called, in machine learning, the **responsibility** of component c for x.

Example 0.1 (A two-component normal mixture)**.** Let $\mathbb{p}_1$ be the $\mathrm{N}(0, 1)$ density and $\mathbb{p}_2$ the $\mathrm{N}(6, 1)$ density, with mixing weights $w_1 = 0.3$ and $w_2 = 0.7$, and let X have density $\mathbb{p}_{\text{mix}} = 0.3\,\mathbb{p}_1 + 0.7\,\mathbb{p}_2$.

- **Mean.** By Theorem 0.2 and the normal means,

$$\begin{aligned}\mathbb{E}[X] &= 0.3 \cdot 0 + 0.7 \cdot 6 \\ &= 4.2.\end{aligned}$$

- **Density.** $\mathbb{p}_{\text{mix}}$ has peaks near the two component means, $\mathbb{p}_{\text{mix}}(0) \approx 0.120$ and $\mathbb{p}_{\text{mix}}(6) \approx 0.279$, with a trough between them, $\mathbb{p}_{\text{mix}}(3) \approx 0.004$. The density at the mean, $\mathbb{p}_{\text{mix}}(4.2) \approx 0.055$, is less than half its value at either peak.
- **Responsibilities.** By Theorem 0.3, at $x = 3$, halfway between the component means, $\mathbb{p}_1(3) = \mathbb{p}_2(3)$ by the symmetry of the normal density, so:

$$\begin{aligned}\mathbb{P}(C = 2 \mid X = 3) &= \frac{0.7\,\mathbb{p}_2(3)}{0.3\,\mathbb{p}_1(3) + 0.7\,\mathbb{p}_2(3)} \\ &= \frac{0.7}{0.3 + 0.7} \\ &= 0.7,\end{aligned}$$

  the prior weight: an observation equally far from both components carries no information about which component it came from. At $x = 4.2$, the same formula gives $\mathbb{P}(C = 2 \mid X = 4.2) \approx 0.9997$.

Remark

Remark. Example 0.1 shows that the mean of a mixture can fall where observations are rare, so a single normal distribution fitted to such data would describe them poorly.

Example 0.2 (The negative binomial as a continuous mixture)**.** The negative binomial distribution is a mixture of Poisson distributions with a continuum of components, one for each mean $\lambda > 0$: its PMF is $\mathbb{P}(Y = y) = \int_0^\infty g(\lambda)\,\frac{\lambda^y e^{-\lambda}}{y!}\,d\lambda$, where the gamma density g plays the role of the mixing weights and the integral replaces the sum in Definition 0.1 (see the proof of Theorem 0.2).

Exercise 0.1 (Choosing a family for a research question)**.** A clinic wants to describe the number of emergency visits it receives each week. Which of the families named on this page fits this research question, and why?

Solution

Solution 0.1. A weekly visit count is a non-negative whole number with no largest value. The Poisson distribution is a natural first choice for such a count, because its values are $0, 1, 2, \ldots$. A Bernoulli variable takes only the values 0 and 1, and a Gaussian count can be negative or fractional, so neither matches the weekly count.
If the weekly counts vary more than a Poisson count allows, the negative binomial distribution is the next candidate.

Distribution functions in R

R names its functions by a prefix that says what the function computes, followed by a short name for the family (Table 4 and Table 5).

Table 4: Prefixes of R's function names, with the Gaussian family as the example.

Prefix	Computes	Example
`d`	the mass function or density	`dnorm()`
`p`	the cumulative distribution function	`pnorm()`
`q`	the quantile function	`qnorm()`
`r`	random draws	`rnorm()`

Table 5: R names for some families in this chapter.

Family	R name	Parameters, in R's order
Bernoulli	`binom` with `size = 1`	`prob`
Binomial	`binom`	`size`, `prob`
Poisson	`pois`	`lambda`
Uniform	`unif`	`min`, `max`
Gaussian	`norm`	`mean`, `sd`
Exponential	`exp`	`rate`

R has no separate Bernoulli family: use the binomial functions with `size = 1`. The next chunk

evaluates each prefix for the standard Gaussian distribution and for a binomial distribution with three trials (Table 6).

```
set.seed(1)
gaussian_density <- dnorm(0)
gaussian_cdf <- pnorm(1.96)
gaussian_quantile <- qnorm(0.975)
gaussian_draw <- rnorm(1)
binomial_mass <- dbinom(2, size = 3, prob = 1 / 2)
binomial_cdf <- pbinom(2, size = 3, prob = 1 / 2)
binomial_quantile <- qbinom(0.5, size = 3, prob = 1 / 2)
binomial_draw <- rbinom(1, size = 3, prob = 1 / 2)

call_text <- c(
  "dnorm(0)",
  "pnorm(1.96)",
  "qnorm(0.975)",
  "rnorm(1)",
  "dbinom(2, size = 3, prob = 1 / 2)",
  "pbinom(2, size = 3, prob = 1 / 2)",
  "qbinom(0.5, size = 3, prob = 1 / 2)",
  "rbinom(1, size = 3, prob = 1 / 2)"
)

call_values <- c(
  gaussian_density,
  gaussian_cdf,
  gaussian_quantile,
  gaussian_draw,
  binomial_mass,
  binomial_cdf,
  binomial_quantile,
  binomial_draw
)

calls <- dplyr::tibble(call = call_text, value = call_values)

knitr::kable(calls, digits = 4)
```

Table 6: The four prefixes applied to two families

call	value
dnorm(0)	0.3989
pnorm(1.96)	0.9750
qnorm(0.975)	1.9600
rnorm(1)	-0.6265
dbinom(2, size = 3, prob = 1 / 2)	0.3750
pbinom(2, size = 3, prob = 1 / 2)	0.8750
qbinom(0.5, size = 3, prob = 1 / 2)	1.0000
rbinom(1, size = 3, prob = 1 / 2)	2.0000

The value of `dbinom(2, size = 3, prob = 1 / 2)` is $3/8 = 0.375$, which matches the two heads in three flips example.

References

[1] Wikipedia contributors, "Binomial distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Binomial_distribution

[2] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[3] Wikipedia contributors, "Poisson distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Poisson_distribution

[4] Wikipedia contributors, "Negative binomial distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Negative_binomial_distribution

[5] Wikipedia contributors, "Laplace distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Laplace_distribution

[6] Wikipedia contributors, "Weibull distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Weibull_distribution

[7] Wikipedia contributors, "Multivariate normal distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Multivariate_normal_distribution

[8] Wikipedia contributors, "Mahalanobis distance — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Mahalanobis_distance

[9] Wikipedia contributors, "Mixture distribution — Wikipedia, the free encyclopedia." 2026. Available: https://en.wikipedia.org/wiki/Mixture_distribution