Key distributions

Last modified: 2026-09-28 23:45:35 (PDT)

> *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
Multivariate normal	Vectors of continuous outcomes

Distribution	Uses
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 Poisson distribution

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

$$\text{P}(Y = y) \stackrel{\text{def}}{=} \frac{\mu^y e^{-\mu}}{y!}, \quad y \in \mathbb{N} = \{0, 1, 2, \dots\} \tag{1}$$

Remark. (see Figure 1)

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

Solution

Solution 0.1.

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

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

$$\text{P}(Y \leq y) = e^{-\mu} \sum_{j=0}^{\lfloor y \rfloor} \frac{\mu^j}{j!} \tag{2}$$

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 1) and factoring out the common term $e^{-\mu}$:

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

□

Example 0.2 (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:

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

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)
```

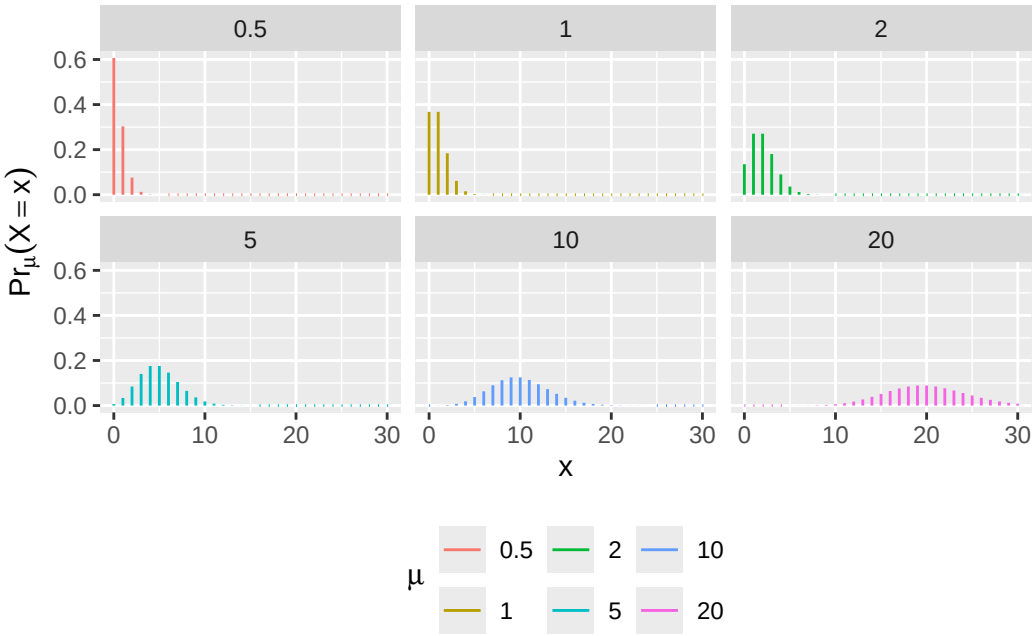

Figure 1: Poisson PMFs, by mean parameter μ

```
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)
```

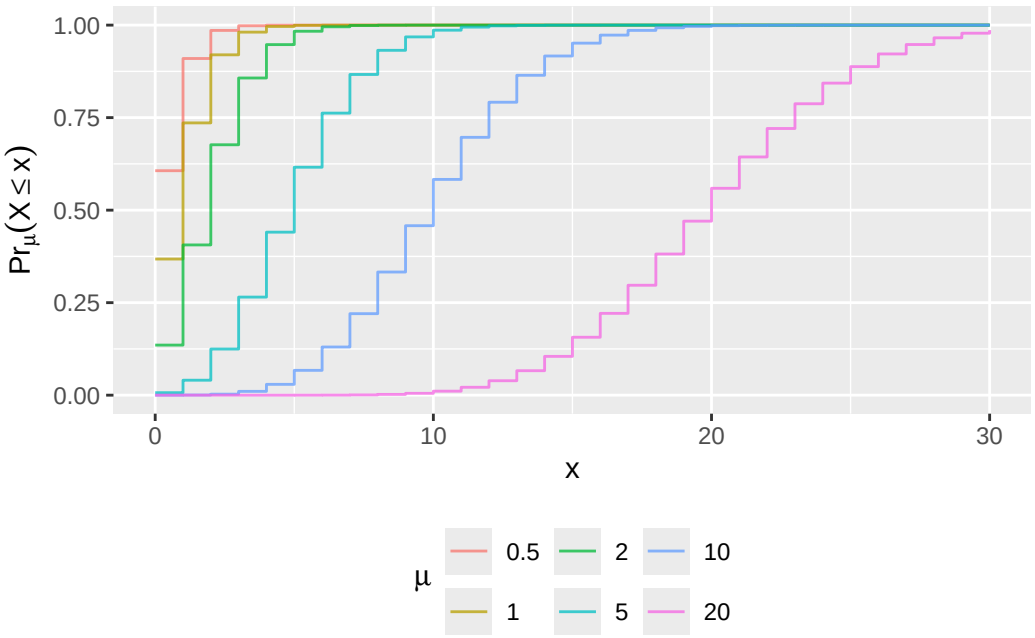

Figure 2: Poisson CDFs

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

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

Solution

Solution.

- $\text{P}(X = 4) = 0.1937803$
- $\text{P}(X \leq 7) = 0.9623787$
- $\text{P}(X > 5) = 0.1771172$

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

- $\text{E}[X] = \mu$
- $\text{Var}(X) = \mu$
- $\text{P}(X = x) = \frac{\mu}{x}\,\text{P}(X = x - 1)$ *for* $x \in \{1, 2, \dots\}$

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

Proof

Proof. **Mean.**

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

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

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

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

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

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

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

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 $\mathrm{P}(X = x)$ 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, $\mathrm{P}(X = \mu) = \mathrm{P}(X = \mu - 1)$, and both are modes.

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

Example 0.3 (Mode of a Poisson distribution)**.** For $X \sim \mathrm{Pois}(2.5)$, $\mathrm{P}(X = 2) = \frac{2.5}{2}\,\mathrm{P}(X = 1) > \mathrm{P}(X = 1)$ and $\mathrm{P}(X = 3) = \frac{2.5}{3}\,\mathrm{P}(X = 2) < \mathrm{P}(X = 2)$, so the mode is $\lfloor 2.5 \rfloor = 2$. For $X \sim \mathrm{Pois}(2)$, $\mathrm{P}(X = 2) = \frac{2}{2}\,\mathrm{P}(X = 1)$, 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. 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:

$$\lambda \overset{\text{def}}{=} \frac{\mathrm{E}[Y]}{t} \tag{3}$$

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 0.3 (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 $\mu \overset{def}{=} \mathrm{E}[Y]$ is:*

$$\mu = \lambda \cdot t \tag{4}$$

Proof

Proof. By Definition 0.3:

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

□

Example 0.4 (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 0.3:

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

Theorem 0.4 (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, $\mathrm{E}[Y_t] = \lambda \cdot t$ for all $t \geq 0$ with one finite rate λ, then there are no expected events without exposure:*

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

Proof

Proof.

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

□

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

Example 0.5 (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 $\mathrm{E}[Y_0] = 0$.

Important

A mean proportional to exposure, $\mathrm{E}[Y_t] = \lambda t$, 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 $\mathrm{E}[Y_t] = c + \lambda t$ 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 0.5 (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 0.6 (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)**.** For a count outcome with exposure magnitude t, the known term $\log t$ in the log-scale decomposition of the mean (Theorem 0.5),

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

is called the **offset**: it shifts $\log \mu$ by a known amount, with no unknown coefficient to estimate.

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 0.7 (The offset for a clinic's event count)**.** In Example 0.6, 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 0.6 (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:

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

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

Example 0.8 (Aggregating independent region counts)**.** Suppose Region A records $X \sim \mathrm{Pois}(\mu_X = 12)$ cases and Region B records $Y \sim \mathrm{Pois}(\mu_Y = 18)$ cases independently. By Theorem 0.6, the combined total count $Z = X + Y$ follows a Poisson distribution:

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

The Negative-Binomial distribution

Definition 0.5 (Negative binomial distribution)**.** A random variable Y has the **negative binomial distribution** with mean $\mu > 0$ and overdispersion parameter $\rho > 0$, written $Y \sim \text{NegBin}(\mu, \rho)$, if, for $y \in \{0, 1, 2, \dots\}$:

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

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

Theorem 0.7 (The negative binomial converges to the Poisson)**.** *Fix $\mu > 0$, 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 1):*

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

Proof

Proof. Fix y. In Definition 0.5, 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 $x \stackrel{\text{def}}{=} \mu/\rho$, so $x \to 0$ as $\rho \to \infty$, and $\rho = \mu/x$:

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

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:

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

which is the Pois(μ) PMF. □

Theorem 0.8 (Mean and variance of the negative binomial distribution)**.** *If* $Y \sim \mathrm{NegBin}(\mu, \rho)$*, then:*

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

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 (Casella and Berger 2002), and $Y \mid \Lambda = \lambda \sim \mathrm{Pois}(\lambda)$, then $Y \sim \mathrm{NegBin}(\mu, \rho)$. To check this, integrate the joint density over λ:

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

Then, by the law of iterated expectations and the law of total variance, using $\mathrm{E}[Y \mid \Lambda] = \mathrm{Var}(Y \mid \Lambda) = \Lambda$ (Theorem 0.2):

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

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

Example 0.9 (Overdispersion relative to the Poisson)**.** With $\mu = 4$ and $\rho = 2$, $\mathrm{Var}(Y) = 4 + 16/2 = 12$, three times the variance of a Pois(4) count with the same mean.

Weibull distribution

Definition 0.6 (Weibull distribution)**.** A non-negative random variable T has the **Weibull distribution** with shape $\alpha > 0$ and rate $\lambda > 0$ if its survival function is:

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

Theorem 0.9 (Weibull density, hazard, and mean)**.** *If T has the Weibull distribution with shape α and rate λ, then for $t > 0$:*

$$\begin{aligned} f(t) &= \alpha\lambda t^{\alpha-1} e^{-\lambda t^\alpha} \\ \mathrm{h}(t) &= \alpha\lambda t^{\alpha-1} \\ \mathrm{E}[T] &= \Gamma(1+1/\alpha) \cdot \lambda^{-1/\alpha} \end{aligned}$$

Proof

Proof. The CDF of T is $F(t) = 1 - \mathrm{S}(t)$ (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 $f = F' = -\mathrm{S}'$ 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 $f(t)/\mathrm{S}(t)$ (hazard equals density over survival):

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

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 $dt = \frac{1}{\alpha}\lambda^{-1/\alpha} u^{1/\alpha - 1}\, du$:

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

□

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

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 0.9 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 $\mathrm{h}(t)$ is:*

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

Proof

Proof. By Theorem 0.9, $\mathrm{h}(t) = \alpha\lambda t^{\alpha-1}$ for $t > 0$, so:

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

For $t > 0$, the factors α, λ, and $t^{\alpha-2}$ are all positive, so $\frac{d}{dt}\,\mathrm{h}(t)$ 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. 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 0.10 (Exponential as a special case)**.** With $\alpha = 1$, Theorem 0.9 gives $\mathrm{h}(t) = \lambda$ and $\mathrm{E}[T] = \Gamma(2)\lambda^{-1} = 1/\lambda$, matching the exponential distribution's constant hazard and

mean. With $\alpha = 2$ and $\lambda = 1$, $\mathrm{h}(t) = 2t$ increases with t, and $\mathrm{E}[T] = \Gamma(3/2) = \sqrt{\pi}/2 \approx 0.886$.

The multivariate normal distribution

Definition 0.7 (Multivariate normal distribution)**.** A $p \times 1$ random vector $\tilde{X} = (X_1, \ldots, X_p)^\top$ has the **multivariate normal distribution** (or **multivariate Gaussian distribution**) with mean parameter $\tilde{\mu} \in \mathbb{R}^p$ and variance parameter , a $p \times p$ positive definite matrix, written $\tilde{X} \sim \mathrm{N}_p(\tilde{\mu},)$, if $\tilde{X}$ has joint density:

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

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

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 $\Pr(\tilde{X} \in A)$. The density is well defined because is positive definite:

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

Theorem 0.10 (The multivariate normal density integrates to 1, with mean $\tilde{\mu}$ and variance)**.** *If $\tilde{X} \sim \mathrm{N}_p(\tilde{\mu},)$ (Definition 0.7), then the density in Definition 0.7 integrates to 1 over $\mathbb{R}^p$, $\mathrm{E}\,\tilde{X} = \tilde{\mu}$ (the expectation of a random vector), and $\mathrm{Var}\big(\tilde{X}\big) =$ (the variance of a random vector).*

Remark. The proof changes variables in a p-dimensional integral, which these notes do not develop; see https://en.wikipedia.org/wiki/Multivariate_normal_distribution. As with the univariate normal distribution, these facts are where the parameters' names come from. They also show why 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 0.11 (The univariate normal is the case $p = 1$)**.** With $p = 1$, $\tilde{\mu} = (\mu)$, and $= (\sigma^2)$ for $\sigma^2 > 0$, $\det() = \sigma^2$ (determinant, $p = 1$) and $^{-1} = (1/\sigma^2)$, so:

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

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 is a $p \times p$ positive definite matrix and $\tilde{x}, \tilde{\mu} \in \mathbb{R}^p$, then*

$$(\tilde{x} - \tilde{\mu})^\top {}^{-1}(\tilde{x} - \tilde{\mu}) \geq 0,$$

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

Proof

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

Definition 0.8 (Mahalanobis distance)**.** For a $p \times p$ positive definite matrix and $\tilde{x}, \tilde{\mu} \in \mathbb{R}^p$, the **Mahalanobis distance** from $\tilde{x}$ to $\tilde{\mu}$ with respect to is:

$$\Delta(\tilde{x}) \stackrel{\text{def}}{=} \sqrt{(\tilde{x} - \tilde{\mu})^\top {}^{-1}(\tilde{x} - \tilde{\mu})}$$

Remark. The square root is defined by Lemma 0.1. See also https://en.wikipedia.org/wiki/Mahalanobis_distance.

Corollary 0.3 (The multivariate normal density decreases with Mahalanobis distance)**.** *If $\tilde{X} \sim \mathrm{N}_p(\tilde{\mu},)$, then:*

$$\mathrm{p}(\tilde{X} = \tilde{x}) = \frac{1}{(2\pi)^{p/2} \det()^{1/2}} e^{-\frac{\Delta(\tilde{x})^2}{2}}$$

where Δ is the Mahalanobis distance to $\tilde{\mu}$ with respect to (Definition 0.8). So the density is highest at $\tilde{x} = \tilde{\mu}$, decreases as $\Delta(\tilde{x})$ increases, and is the same at every $\tilde{x}$ with the same $\Delta(\tilde{x})$.

Proof

Proof. Substitute Definition 0.8 into Definition 0.7. The function $\delta \mapsto \mathrm{e}^{-\delta^2/2}$ decreases on $\delta \geq 0$, and by Lemma 0.1, $\Delta(\tilde{x}) = 0$ only at $\tilde{x} = \tilde{\mu}$. □

Example 0.12 (Mahalanobis distance for diagonal variance matrices)**.** If is diagonal with diagonal elements $\sigma_1^2, \ldots, \sigma_p^2$, all positive, then $^{-1}$ is diagonal with diagonal elements $1/\sigma_1^2, \ldots, 1/\sigma_p^2$ (multiplying the two gives $\mathbf{I}_p$; matrix inverse), so:

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

Two special cases:

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

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

Theorem 0.11 (A multivariate normal vector with diagonal variance has independent normal components)**.** *If $\tilde{X} \sim \mathrm{N}_p(\tilde{\mu},)$ and is diagonal with diagonal elements $\sigma_1^2, \ldots, \sigma_p^2$, then:*

- *the joint density of $\tilde{X}$ 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, $\det() = \prod_{i=1}^{p} \sigma_i^2$, so $\det()^{1/2} = \prod_{i=1}^{p} \sigma_i$. Using Example 0.12 for the quadratic form:

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

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

Proof

Proof. By Theorem 0.10, $\mathrm{Var}\left(\tilde{X}\right) =$, whose (i, j)-th element is $\mathrm{Cov}(X_i, X_j)$ (elements of the variance matrix). If is diagonal, the components are independent by Theorem 0.11. 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
- $\mathrm{E}[X_i^2] < \infty$, because $\mathrm{Var}(X_i)$, the (i, i)-th element of , is defined, and a variance is defined only when $\mathrm{E}[X_i^2] < \infty$.

So is diagonal by independent components give a diagonal variance matrix. □

Remark. In general, uncorrelated random variables need not be independent (an example). Corollary 0.4 needs the whole vector to be multivariate normal: two random variables that are each normal and uncorrelated can still be dependent (https://en.wikipedia.org/wiki/Normally_distributed_and_uncorrelated_does_not_imply_independent).

Theorem 0.12 (Mahalanobis distance in eigenvector coordinates)**.** *Let* *be positive definite, with eigendecomposition* $= \mathbf{Q}\mathbf{Q}^\top$*, where* $\mathbf{Q}$ *has columns* $\tilde{q}_1, \ldots, \tilde{q}_p$ *and* *has diagonal elements* $\lambda_1, \ldots, \lambda_p$*. Then:*

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

Proof

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

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

□

Remark. y_i is the coordinate of $\tilde{x}-\tilde{\mu}$ along the eigenvector $\tilde{q}_i$. So the set of points at Mahalanobis distance δ from $\tilde{\mu}$, where the multivariate normal density is constant (Corollary 0.3), is an ellipse ($p=2$) or ellipsoid centered at $\tilde{\mu}$, with axes along the eigenvectors of and half-length $\delta\sqrt{\lambda_i}$ along $\tilde{q}_i$.

Example 0.13 (Elliptical contours of a bivariate normal density)**.** Let $\tilde{\mu} = \tilde{0}$ and $= \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$, which has eigenvalues 3 and 1, with eigenvectors $\tilde{q}_1 = \frac{1}{\sqrt{2}}(1,1)^\top$ and $\tilde{q}_2 = \frac{1}{\sqrt{2}}(1,-1)^\top$ (an eigendecomposition example). By Theorem 0.12:

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

The point $\tilde{x} = \sqrt{3}\,\tilde{q}_1 = \sqrt{3/2}\,(1,1)^\top$ has $x_1+x_2 = \sqrt{6}$ and $x_1-x_2 = 0$, so $\Delta(\tilde{x})^2 = 6/6 = 1$; the point $\tilde{x} = \tilde{q}_2$ has $x_1+x_2 = 0$ and $x_1-x_2 = \sqrt{2}$, so $\Delta(\tilde{x})^2 = 2/2 = 1$. So the contour $\Delta(\tilde{x}) = 1$ of the $\mathrm{N}_2\big(\tilde{0},\big)$ 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.

Mixture distributions

Definition 0.9 (Mixture density)**.** Let $\mathrm{p}_1, \ldots, \mathrm{p}_K$ be densities on $\mathbb{R}$, and let $w_1, \ldots, w_K$ be numbers with:

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

The **mixture** of $\mathrm{p}_1, \ldots, \mathrm{p}_K$ with **mixing weights** $w_1, \ldots, w_K$ is the function:

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

The densities $\mathrm{p}_1, \ldots, \mathrm{p}_K$ are the mixture's **components**.

Theorem 0.13 (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*

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

with p_c and w_c as in Definition 0.9. Then:

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

Proof

Proof. By the marginal PMF from a joint density-mass function, $\mathrm{P}(C = c) = \int_{-\infty}^{\infty} w_c \, \mathrm{p}_c(x) \, dx = w_c$, because p_c is a density and so integrates to 1. Then, by the definition of the conditional density, $\mathrm{p}(X = x \mid C = c) = w_c \, \mathrm{p}_c(x)/w_c = \mathrm{p}_c(x)$.
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:

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

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

Remark. Theorem 0.13 describes how to generate a draw from a mixture: first draw a component label C with $\mathrm{P}(C = c) = w_c$, then draw X from the component density 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 0.14 (The mean of a mixture is the weighted mean of the component means)**.** *If X has density p_{mix} (Definition 0.9), and each component p_c has a defined mean $\mu_c = \int_{-\infty}^{\infty} x \, \mathrm{p}_c(x)\, dx$, then:*

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

Proof

Proof. The integral defining $\mathrm{E}[X]$ converges absolutely, because $\int |x| \, \mathrm{p}_{\text{mix}}(x)\, dx = \sum_c w_c \int |x| \, \mathrm{p}_c(x)\, dx$, a finite sum of finite numbers. Then:

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

The first step is the definition of expectation. □

Theorem 0.15 (Which component did an observation come from?)**.** *Under the conditions of Theorem 0.13, for each x with $\mathrm{p}_{mix}(x) > 0$:*

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

Proof

Proof. By Theorem 0.13, p_{mix} is a density of X. By the definition of the conditional PMF (the case X continuous, C discrete):

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

□

Remark. Theorem 0.15 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 $\mathrm{p}_c(x)$ at the observation and divided by the overall density $\mathrm{p}_{\text{mix}}(x)$, which plays the role of the law of total probability. In machine learning, $\mathrm{P}(C = c \mid X = x)$ is called the **responsibility** of component c for x.

Example 0.14 (A two-component normal mixture)**.** Let p_1 be the $\mathrm{N}(0,1)$ density and 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 $\mathrm{p}_{\text{mix}} = 0.3\,\mathrm{p}_1 + 0.7\,\mathrm{p}_2$.

- **Mean.** By Theorem 0.14 and the normal means, $\mathrm{E}[X] = 0.3 \cdot 0 + 0.7 \cdot 6 = 4.2$.
- **Density.** p_{mix} has peaks near the two component means, $\mathrm{p}_{\text{mix}}(0) \approx 0.120$ and $\mathrm{p}_{\text{mix}}(6) \approx 0.279$, with a trough between them, $\mathrm{p}_{\text{mix}}(3) \approx 0.004$. The density at the mean, $\mathrm{p}_{\text{mix}}(4.2) \approx 0.055$, is less than half its value at either peak.
- **Responsibilities.** By Theorem 0.15, at $x = 3$, halfway between the component means, $\mathrm{p}_1(3) = \mathrm{p}_2(3)$ by the symmetry of the normal density, so:

$$\mathrm{P}(C = 2 \mid X = 3) = \frac{0.7\,\mathrm{p}_2(3)}{0.3\,\mathrm{p}_1(3) + 0.7\,\mathrm{p}_2(3)} = \frac{0.7}{0.3 + 0.7} = 0.7,$$

  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 $\mathrm{P}(C = 2 \mid X = 4.2) \approx 0.9997$.

Remark. Example 0.14 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 0.15 (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 $\mathrm{P}(Y = y) = \int_0^\infty g(\lambda)\,\frac{\lambda^y e^{-\lambda}}{y!}\,d\lambda$, where the gamma density g plays the role of the mixing weights and the integral replaces the sum in Definition 0.9 (see the proof of Theorem 0.8).

References

Casella, George, and Roger Berger. 2002. *Statistical Inference*. 2nd ed. Cengage Learning. https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/.