3 Probability Theory

R. Heller · Date: 2026-05-11 · Last changed: 2026-11-10

The probability machinery the rest of the book leans on: Bayes' rule, common discrete and continuous distributions, the law of total probability, Q-Q diagnostics, and density estimation. The chapter is short on theory and long on R-side intuition.

This chapter contains **40 method pages** and **3 labs**. If you are not sure which method to read, return to Chapter 0 and follow the decision tree to the right node.

3.1 Method pages

Method	Source slug
Bayes' Theorem	`bayes-theorem`
Conditional Distributions	`conditional-distributions`
Conditional Probability	`conditional-probability`
Convolutions of Distributions	`convolutions`
Copulas: An Introduction	`copulas-introduction`
Discrete vs Continuous Variables	`discrete-vs-continuous`
Expectation of a Random Variable	`expectation`
Independence	`independence`
Joint Distributions	`joint-distributions`
Kolmogorov's Axioms of Probability	`kolmogorov-axioms`
Law of Total Probability	`law-of-total-probability`
Marginal Distributions	`marginal-distributions`
Random Variables	`random-variables`
Sample Space and Events	`sample-space-events`
Student's t-Distribution	`t-distribution`
The Bernoulli Distribution	`bernoulli-distribution`
The Beta Distribution	`beta-distribution`
The Binomial Distribution	`binomial-distribution`
The Chi-Squared Distribution	`chi-squared-distribution`
The Correlation Coefficient	`correlation-coefficient`
The Cumulative Distribution Function	`cumulative-distribution-function`
The Exponential Distribution	`exponential-distribution`
The F Distribution	`f-distribution`
The Gamma Distribution	`gamma-distribution`
The Geometric Distribution	`geometric-distribution`

Method	Source slug
The Hazard Function	`hazard-function`
The Hypergeometric Distribution	`hypergeometric-distribution`
The Log-Normal Distribution	`lognormal-distribution`
The Multinomial Distribution	`multinomial-distribution`
The Multivariate Normal Distribution	`multivariate-normal`
The Negative Binomial Distribution	`negative-binomial-distribution`
The Normal Distribution	`normal-distribution`
The Poisson Distribution	`poisson-distribution`
The Probability Density Function	`probability-density-function`
The Probability Mass Function	`probability-mass-function`
The Survival Function	`survival-function`
The Uniform Distribution	`uniform-distribution`
The Weibull Distribution	`weibull-distribution`
Transformations of Random Variables	`transformations-of-rv`
Variance and Covariance	`variance-covariance-of-rv`

3.2 Labs

Lab
Probability, conditional probability, Bayes' theorem
Discrete distributions: Bernoulli, binomial, Poisson
Continuous distributions and Q-Q plots

3.3 Bayes' Theorem

3.3.1 Introduction

Bayes' theorem is the rule for inverting conditional probabilities: given $P(B \mid A)$, it tells us $P(A \mid B)$. It is the mathematical bridge between diagnostic sensitivity and positive predictive value, between forward simulation and inverse inference, between a likelihood function and a posterior distribution. Every Bayesian analysis, every clinical prediction, and every probabilistic classifier is an application of Bayes' theorem in disguise.

3.3.2 Prerequisites

Conditional probability and the law of total probability.

3.3.3 Theory

For events A and B with $P(B) > 0$,

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}.$$

When $\{A_i\}$ forms a partition of the sample space, the denominator expands by the law of total probability:

$$P(A_i \mid B) = \frac{P(B \mid A_i)\, P(A_i)}{\sum_j P(B \mid A_j)\, P(A_j)}.$$

In the Bayesian vocabulary:

- $P(A)$ is the **prior**: what was believed before observing B.
- $P(B \mid A)$ is the **likelihood**: how likely B is under each hypothesis.
- $P(A \mid B)$ is the **posterior**: updated belief about A after observing B.
- $P(B)$ is the **marginal likelihood** or "evidence".

Posterior $\propto$ likelihood $\times$ prior is the single most important slogan in Bayesian statistics.

3.3.4 Assumptions

- $P(B) > 0$; otherwise conditioning is undefined.
- The prior $P(A)$ must be specified; choosing it is where domain knowledge enters.

3.3.5 R Implementation

```r
library(ggplot2)

# Diagnostic test: prior = prevalence, likelihood = sensitivity and specificity
sensitivity <- 0.92
specificity <- 0.97
prevalences <- seq(0.001, 0.5, length.out = 100)

ppv <- sensitivity * prevalences /
       (sensitivity * prevalences + (1 - specificity) * (1 - prevalences))

df <- data.frame(prevalence = prevalences, ppv = ppv)

ggplot(df, aes(prevalence, ppv)) +
  geom_line(colour = "#2A9D8F", linewidth = 1) +
  geom_vline(xintercept = 0.01, linetype = "dashed", colour = "#F4A261") +
  geom_vline(xintercept = 0.30, linetype = "dashed", colour = "#F4A261") +
  labs(x = "Disease prevalence",
       y = "Positive predictive value",
       title = "PPV depends strongly on prevalence (Bayes' theorem)") +
  theme_minimal()

c(prevalence_001 = ppv[which.min(abs(prevalences - 0.01))],
  prevalence_030 = ppv[which.min(abs(prevalences - 0.30))])
```

3.3.6 Output & Results

```
prevalence_001 prevalence_030
        0.2368         0.9290
```

At 1% prevalence, a positive test on a 92%-sensitive, 97%-specific assay gives only 24% probability of disease. At 30% prevalence, the same test yields a 93% PPV. The test characteristics are unchanged; the posterior depends crucially on the prior.

3.3.7 Interpretation

The prevalence dependence of PPV is the canonical real-world application of Bayes' theorem. Reporting "our test has 92% sensitivity and 97% specificity" is meaningless without the prevalence of the intended use population; the clinically relevant quantity is PPV.

3.3.8 Practical Tips

- Never cite sensitivity / specificity in isolation when discussing diagnostic utility; always give the PPV and NPV at the target prevalence.
- Bayes' theorem scales trivially to many hypotheses via the partition form; for continuous parameter spaces, the sums become integrals.
- Prior choice is contentious in some Bayesian applications; weakly informative priors (Cauchy, normal with wide scale) are standard defaults.
- Updating sequentially: the posterior after observation 1 becomes the prior for observation 2. Chains of updates are equivalent to a single update with all observations.
- The **likelihood ratio** $P(B \mid A)/P(B \mid A^c)$ converts the prior odds of A to the posterior odds multiplicatively – often a cleaner formulation than raw probabilities.

3.3.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.3.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.4 The Bernoulli Distribution

3.4.1 Introduction

The Bernoulli distribution models a single binary trial: success vs. failure, disease vs. healthy, click vs. no click. It is the simplest non-trivial distribution and the building block of the binomial, geometric, and negative binomial distributions.

3.4.2 Prerequisites

Random variables and basic probability.

3.4.3 Theory

A random variable X is **Bernoulli(p)** if it takes values in $\{0, 1\}$ with

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p.$$

PMF: $p_X(x) = p^x (1-p)^{1-x}$ for $x \in \{0, 1\}$.

Moments:

- $E[X] = p$.
- $\mathrm{Var}(X) = p(1-p)$, maximised at $p = 0.5$ (value 0.25).
- $E[X^k] = p$ for every $k \geq 1$ (since $X^k = X$).

The sum of n iid Bernoulli(p) variables is **Binomial**(n, p).

MLE of p from an iid sample $X_1, \ldots, X_n$: $\hat{p} = \bar{X}$, the sample proportion.

3.4.4 Assumptions

Independence across trials and a common probability p. Violations of either produce overdispersion (if p varies) or correlation (if trials are dependent).

3.4.5 R Implementation

```r
set.seed(2026)

p ← 0.3
n ← 10000

# Sample from Bernoulli(p) via rbinom with size = 1
X ← rbinom(n, size = 1, prob = p)

# Sample mean is the MLE of p
c(p_hat = mean(X), p_true = p)

# Sample variance approximates p(1 - p)
c(sample_var = var(X), theoretical = p * (1 - p))

# 95% Wald confidence interval for p
se_p ← sqrt(mean(X) * (1 - mean(X)) / n)
c(lower = mean(X) - 1.96 * se_p, upper = mean(X) + 1.96 * se_p)

# Wilson (better for small p or small n)
binom::binom.confint(sum(X), n, conf.level = 0.95, methods = "wilson")
```

3.4.6 Output & Results

```
    p_hat    p_true
   0.3002     0.3000

sample_var theoretical
   0.2101      0.2100

   lower    upper
  0.2912   0.3092

method    x     n    mean    lower    upper
Wilson 3002 10000  0.3002   0.2912   0.3094
```

Empirical mean is $0.3002 \approx p$; empirical variance is $0.2101 \approx p(1-p) = 0.21$; Wald and Wilson CIs agree in this range.

3.4.7 Interpretation

In applied work, the Bernoulli is the underlying distribution of every binary outcome: adverse event, responder status, screen-positive. Reporting usually aggregates to proportions across n independent Bernoullis – i.e., to the binomial framework.

3.4.8 Practical Tips

- For small samples or extreme p (near 0 or 1), use Wilson or Clopper-Pearson CIs instead of Wald.
- Overdispersion (variance exceeding $p(1-p)$) in grouped data signals unobserved heterogeneity; consider beta-binomial or mixed-effects logistic regression.
- `rbinom(n, size = 1, prob = p)` is the direct way to simulate Bernoullis; `sample(0:1, n, replace = TRUE, prob = c(1 - p, p))` is equivalent.

- The Bernoulli likelihood is the foundation of logistic regression; MLE of p from regression coefficients requires the logit link.
- Two dependent Bernoullis require a specification of their joint distribution – usually via a copula or a 2×2 joint PMF.

3.4.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.4.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.5 The Beta Distribution

3.5.1 Introduction

The beta distribution is the flexible family of probabilities on the unit interval $[0, 1]$. It is the natural distribution for proportions, success probabilities, and measurements bounded between two fixed limits. In Bayesian inference, it is the conjugate prior for the binomial and Bernoulli likelihoods.

3.5.2 Prerequisites

Continuous distributions, Bayesian inference basics.

3.5.3 Theory

A beta random variable with shape parameters $a > 0$ and $b > 0$ has PDF

$$f(x) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}, \qquad 0 \le x \le 1,$$

where $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$ is the beta function.

Moments:

- $E[X] = a/(a+b)$.
- $\mathrm{Var}(X) = ab/[(a+b)^2(a+b+1)]$.

Special cases:

- $a = b = 1$: Uniform$(0, 1)$.
- $a = b = 0.5$: Jeffreys prior for the binomial probability.
- $a, b > 1$: unimodal density with mode $(a - 1)/(a + b - 2)$.
- $a < 1$ and $b < 1$: bimodal (poles at 0 and 1).

Conjugacy: given a Bernoulli/Binomial likelihood with success probability θ and a Beta (a, b) prior on θ, the posterior after observing k successes in n trials is Beta $(a + k, b + n - k)$.

3.5.4 Assumptions

The variable must be bounded between 0 and 1. For probabilities near 0 or 1, the distribution can have a pole; use care when simulating.

3.5.5 R Implementation

```r
curve(dbeta(x, 2, 5),  from = 0, to = 1, col = "steelblue", lwd = 2,
      main = "Beta densities", ylab = "f(x)")
curve(dbeta(x, 5, 2),  add = TRUE, col = "#F4A261", lwd = 2)
curve(dbeta(x, 0.5, 0.5), add = TRUE, col = "#6A4C93", lwd = 2)
curve(dbeta(x, 2, 2),  add = TRUE, col = "#2A9D8F", lwd = 2)
legend("top", legend = c("Beta(2,5)", "Beta(5,2)", "Beta(0.5,0.5)", "Beta(2,2)"),
       col = c("steelblue", "#F4A261", "#6A4C93", "#2A9D8F"), lwd = 2)

# Moments
a <- 2; b <- 5
c(mean = a / (a + b), var = a * b / ((a + b)^2 * (a + b + 1)))

X <- rbeta(1e5, a, b)
c(emp_mean = mean(X), emp_var = var(X))

# Bayesian update example
prior_a <- 2; prior_b <- 2
k <- 15; n <- 20
post_a <- prior_a + k; post_b <- prior_b + n - k
c(posterior_mean = post_a / (post_a + post_b),
  posterior_95_CI_lower = qbeta(0.025, post_a, post_b),
  posterior_95_CI_upper = qbeta(0.975, post_a, post_b))
```

3.5.6 Output & Results

```
    mean      var
  0.2857   0.0255

  emp_mean   emp_var
    0.286     0.025

 posterior_mean posterior_95_CI_lower posterior_95_CI_upper
          0.708                 0.504                 0.871
```

Empirical moments match theoretical. The Bayesian update with a Beta(2,2) prior and 15 successes in 20 trials gives a posterior mean of 0.71 with a 95% credible interval of (0.50, 0.87).

3.5.7 Interpretation

Beta distributions appear everywhere in Bayesian analysis of proportions. Reporting "posterior probability of success 0.71 (95% credible interval 0.50 to 0.87)" is a beta-based statement.

3.5.8 Practical Tips

- For weakly informative priors, Beta(1, 1) (uniform) or Beta(0.5, 0.5) (Jeffreys) are common defaults.
- Very large a and b correspond to strongly informative priors; check the prior predictive distribution to verify plausibility.
- Beta regression (`betareg`) models proportions in $(0, 1)$ as a function of covariates.
- Values exactly at 0 or 1 are problematic; transform via $(x(n-1)+0.5)/n$ or use zero-one-inflated beta.
- For a quick-look sanity check, a beta with $a = b$ is symmetric and centred at 0.5, scaling by $(a + b)$ as a concentration parameter.

3.5.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.5.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.6 The Binomial Distribution

3.6.1 Introduction

The binomial distribution counts the number of successes in n independent trials, each with the same probability p. It is the most common discrete distribution in applied statistics: responder counts in a trial, positive test results in a screen, correct classifications in an evaluation.

3.6.2 Prerequisites

Bernoulli trials, basic combinatorics.

3.6.3 Theory

If $X = \sum_{i=1}^{n} B_i$ where B_i are iid Bernoulli(p), then $X \sim \text{Binomial}(n, p)$.

PMF: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ for $k = 0, 1, \ldots, n$.

Moments:

- $E[X] = np$.
- $\text{Var}(X) = np(1-p)$.

Normal approximation: $\text{Binomial}(n, p) \approx \mathcal{N}(np, np(1-p))$ when $np \geq 10$ and $n(1-p) \geq 10$. Continuity-corrected: $P(X \leq k) \approx \Phi\big((k + 0.5 - np)/\sqrt{np(1-p)}\big)$.

Poisson approximation: for large n and small p with $np = \lambda$ moderate, $\text{Binomial}(n, p) \approx \text{Poisson}(\lambda)$.

Relationship to other distributions:

- Sum of independent binomials with same p: binomial with sum of n's.
- Conditional on total successes, two independent binomials have a hypergeometric distribution of one.

3.6.4 Assumptions

- Fixed number of trials n.
- Independent trials.
- Constant success probability p.

Violations produce overdispersion (for varying p) or correlation (for dependent trials).

3.6.5 R Implementation

```r
library(ggplot2)

n ← 20; p ← 0.35

k ← 0:n
pmf ← dbinom(k, n, p)

data.frame(k, pmf = round(pmf, 4))

# Moments
c(mean = n * p, var = n * p * (1 - p))
mean(rbinom(1e5, n, p)); var(rbinom(1e5, n, p))

# Normal approximation
x ← seq(0, n, length.out = 400)
norm_approx ← dnorm(x, mean = n * p, sd = sqrt(n * p * (1 - p)))

ggplot() +
  geom_col(data = data.frame(k, pmf), aes(k, pmf), fill = "#2A9D8F") +
  geom_line(data = data.frame(x, norm_approx), aes(x, norm_approx),
            colour = "#F4A261", linewidth = 1) +
  labs(x = "k", y = "probability",
       title = "Binomial(20, 0.35) with normal approximation") +
  theme_minimal()

# Interval probability and quantile
pbinom(10, n, p)        # P(X ⩽ 10)
qbinom(0.025, n, p)     # 2.5th percentile
```

3.6.6 Output & Results

```
    k    pmf
1   0 0.0020
2   1 0.0219
...
20 20 0.0000

    mean     var
     7.0     4.55

[1] 0.9468     # P(X ⩽ 10)
[1]      3     # 2.5th percentile
```

The normal approximation tracks the PMF closely at $n = 20$, $p = 0.35$ where $np = 7$ and $n(1-p) = 13$.

3.6.7 Interpretation

Binomial reasoning pervades applied statistics: one-proportion tests, two-proportion tests, logistic regression, and every chi-squared-style comparison. Reporting a proportion with a Wilson 95% CI is reporting a binomial inference.

3.6.8 Practical Tips

- For small n or extreme p, use the exact binomial test / Clopper-Pearson interval instead of normal-based inference.
- Overdispersion relative to binomial signals either varying p across trials or correlation among trials; the beta-binomial is a common remedy.
- `dbinom(k, n, p, log = TRUE)` gives log-probabilities, useful for likelihoods with many observations.
- Compound inferences (e.g., comparing two binomial proportions) have their own CIs; prefer score intervals (`prop.test`) over Wald.
- In R the order of arguments is `dbinom(x, size, prob)` where `size` is n and `prob` is p; mixing them up is a common bug.

3.6.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.6.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.7 The Chi-Squared Distribution

3.7.1 Introduction

The chi-squared distribution arises as the sum of squared independent standard normal random variables. It is the sampling distribution of the sample variance under normality, the reference distribution for goodness-of-fit and contingency tests, and a key ingredient in the F distribution used throughout ANOVA.

3.7.2 Prerequisites

Normal distribution, basic calculus.

3.7.3 Theory

If $Z_1, \ldots, Z_k$ are iid standard normal, then

$$Q = \sum_{i=1}^{k} Z_i^2 \sim \chi^2_k.$$

The parameter k is the **degrees of freedom**. The PDF is

$$f(x) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}, \qquad x > 0.$$

Moments: $E[Q] = k$, $\mathrm{Var}(Q) = 2k$.

Sum property: if $Q_1 \sim \chi^2_{k_1}$ and $Q_2 \sim \chi^2_{k_2}$ are independent, then $Q_1 + Q_2 \sim \chi^2_{k_1+k_2}$.

Sample variance: for iid normal $X_1, \ldots, X_n$ with variance σ^2,

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

This is the basis of confidence intervals for a normal variance.

Large-df behaviour: χ^2_k is approximately normal for large k (by the CLT applied to the sum of squared normals).

3.7.4 Assumptions

For the sample-variance result: iid normal observations. For the goodness-of-fit approximation in chi-squared tests: independent observations and large enough expected cell counts.

3.7.5 R Implementation

```r
k ← 5

# PDF and CDF
x ← seq(0, 20, length.out = 400)
plot(x, dchisq(x, df = k), type = "l", col = "#2A9D8F", lwd = 2,
     main = paste("Chi-squared,", k, "df"), ylab = "f(x)")

# Moments
c(theoretical_mean = k, theoretical_var = 2 * k,
  empirical_mean   = mean(rchisq(1e5, k)),
  empirical_var    = var(rchisq(1e5, k)))

# Verify: sum of k squared standard normals
n ← 1e5
Q_direct ← rchisq(n, df = k)
Q_from_normals ← rowSums(matrix(rnorm(n * k), n, k)^2)
c(mean_direct = mean(Q_direct), mean_from_normals = mean(Q_from_normals))

# Confidence interval for a variance
set.seed(2026)
X ← rnorm(30, mean = 0, sd = 2)
s2 ← var(X)
n ← length(X)
chi_l ← qchisq(0.025, n - 1); chi_u ← qchisq(0.975, n - 1)
c((n - 1) * s2 / chi_u, (n - 1) * s2 / chi_l)
```

3.7.6 Output & Results

```
theoretical_mean  theoretical_var  empirical_mean  empirical_var
             5.0             10.0           5.003           9.98

mean_direct mean_from_normals
      5.005             5.004

[1] 2.542 7.206
```

A 95% CI for the true variance (which is 4) is (2.54, 7.21), which captures the true value. The sum-of-squared-normals construction produces the expected mean of 5.

3.7.7 Interpretation

Chi-squared tests of independence, goodness-of-fit, and variance homogeneity all reference the chi-squared distribution. Reporting a test statistic value like "$\chi^2_3 = 9.5$, $p = 0.023$" is reporting a chi-squared-distributed statistic with its p-value.

3.7.8 Practical Tips

- Chi-squared confidence intervals for variance are highly sensitive to the normality assumption.
- For large df, use the normal approximation: $\chi^2_k \approx \mathcal{N}(k, 2k)$, or more accurately $\sqrt{2\chi^2_k} \approx \mathcal{N}(\sqrt{2k-1}, 1)$.
- Chi-squared tests with small expected counts are poorly approximated; use exact tests or Monte Carlo.
- Non-central chi-squared (with non-centrality parameter λ) governs power calculations for chi-squared tests.
- The square of a t_ν is $F_{1,\nu}$, not chi-squared; do not conflate.

3.7.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.7.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.8 Conditional Distributions

3.8.1 Introduction

A conditional distribution describes how one random variable behaves given a specific value (or values) of another. Conditional expectations are what regression models estimate; conditional variances quantify residual uncertainty. Together they are the language of predictive statistics.

3.8.2 Prerequisites

Joint distributions, marginal distributions, conditional probability.

3.8.3 Theory

For continuous joint density $f_{X,Y}$ with marginal $f_Y > 0$, the **conditional density of X given $Y = y$** is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Analogous expression for discrete variables using PMFs.

Conditional expectation of X given $Y = y$:

$$E[X \mid Y = y] = \int x f_{X|Y}(x \mid y)\, dx.$$

As a function of y, $E[X \mid Y]$ is itself a random variable.

Tower property (law of total expectation):

$$E[X] = E[E[X \mid Y]].$$

Conditional variance: $\operatorname{Var}(X \mid Y = y) = E[X^2 \mid Y = y] - (E[X \mid Y = y])^2$.

Law of total variance: $\operatorname{Var}(X) = E[\operatorname{Var}(X \mid Y)] + \operatorname{Var}(E[X \mid Y])$. Within-group variance plus between-group variance.

For a bivariate normal with correlation ρ:
$X \mid Y = y \sim \mathcal{N}\big(\mu_X + \rho \sigma_X \sigma_Y^{-1}(y - \mu_Y), \sigma_X^2(1 - \rho^2)\big)$.

3.8.4 Assumptions

$f_Y(y) > 0$ (or the conditional probability is defined via an alternate limiting procedure); $E[|X|] < \infty$ for the conditional expectation to exist.

3.8.5 R Implementation

```
library(mvtnorm)
set.seed(2026)

# Bivariate normal
mu ← c(0, 0); Sigma ← matrix(c(1, 0.7, 0.7, 1), 2, 2)
rho ← Sigma[1, 2] / sqrt(prod(diag(Sigma)))
XY ← rmvnorm(5e4, mean = mu, sigma = Sigma)

# Conditional on Y near y0, distribution of X
y0 ← 1.0; window ← 0.05
X_given_Y ← XY[abs(XY[, 2] - y0) < window, 1]

c(n_window    = length(X_given_Y),
  emp_mean    = mean(X_given_Y),
  theor_mean  = mu[1] + rho * (y0 - mu[2]),
  emp_sd      = sd(X_given_Y),
  theor_sd    = sqrt(Sigma[1, 1] * (1 - rho^2)))

# Tower property: E[X] = E[E[X | Y]]
c(E_X = mean(XY[, 1]),
  tower = mean(mu[1] + rho * (XY[, 2] - mu[2])))

# Law of total variance
var_total ← var(XY[, 1])
var_within ← Sigma[1, 1] * (1 - rho^2)
var_between ← rho^2 * Sigma[1, 1]
c(total = var_total,
  sum_of_parts = var_within + var_between)
```

3.8.6 Output & Results

```
 n_window    emp_mean theor_mean      emp_sd    theor_sd
   4094         0.693       0.700       0.708       0.714

        E_X       tower
      0.003      -0.001

        total sum_of_parts
        1.000        1.000
```

Empirical conditional moments match theoretical; tower property and law of total variance hold exactly (up to Monte Carlo error).

3.8.7 Interpretation

Regression is the enterprise of estimating $E[Y \mid X = x]$ as a function of x. When a manuscript reports "for every 1-unit increase in x, y increases by β", it is interpreting a slope of the conditional expectation.

3.8.8 Practical Tips

- Conditional distributions require the conditioning event to have positive probability (or a limiting argument for continuous variables).

- $E[Y \mid X]$ is a function of X and therefore a random variable; $E[Y \mid X = x]$ is a specific number.
- The conditional expectation is the best predictor of Y from X under squared-error loss (Hilbert projection).
- Jensen's inequality extends to conditional expectations: $E[g(X) \mid Y] \geq g(E[X \mid Y])$ for convex g.
- In high-dimensional conditioning, direct computation is intractable; regression models (linear, generalised, spline-based) estimate $E[Y \mid X]$ directly.

3.8.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.8.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.9 Conditional Probability

3.9.1 Introduction

Conditional probability quantifies how information about one event updates the probability of another. It is the lens through which every diagnostic test, every Bayesian update, and every hazard calculation is viewed. The rules are simple; the interpretations are subtle and recurrently misunderstood in applied research.

3.9.2 Prerequisites

Basic probability, sample spaces, and set operations.

3.9.3 Theory

For events A and B with $P(B) > 0$, the **conditional probability of A given B** is

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

Conceptually, conditioning restricts attention to the subset of the sample space where B occurred, and then asks: among those outcomes, what fraction are in A?

Rearranging gives the **product (multiplication) rule**:

$$P(A \cap B) = P(A \mid B)\,P(B) = P(B \mid A)\,P(A).$$

The product rule extends to multiple events via the chain rule:

$$P(A_1 \cap \ldots \cap A_n) = P(A_1)\,P(A_2 \mid A_1)\,P(A_3 \mid A_1 \cap A_2) \cdots$$

Conditioning preserves the axioms: fixing B, the function $P(\cdot \mid B)$ is itself a valid probability measure on the conditional sample space. All standard properties (complement rule, inclusion-exclusion) hold with P replaced by $P(\cdot \mid B)$.

3.9.4 Assumptions

- $P(B) > 0$. Conditioning on a zero-probability event is not defined in the elementary setting; measure-theoretic extensions exist but are beyond scope.
- The original probability space is fixed; conditioning does not change Ω, only the measure restricted to B.

3.9.5 R Implementation

We use a 2×2 population table to demonstrate conditioning mechanically.

```
set.seed(2026)

# Simulated screening-test data: 10 000 patients
# True disease status and test result
n ← 10000
disease_prev ← 0.05
sensitivity  ← 0.90
specificity  ← 0.95

patients ← data.frame(
  disease = rbinom(n, 1, disease_prev)
) ▷
  within({
    test ← ifelse(disease == 1,
                  rbinom(length(disease), 1, sensitivity),
                  rbinom(length(disease), 1, 1 - specificity))
  })

tab ← table(patients$test, patients$disease,
            dnn = c("test", "disease"))
tab

p_disease_given_test_pos ← tab["1", "1"] / sum(tab["1", ])
p_disease_given_test_pos

p_test_pos_given_disease ← tab["1", "1"] / sum(tab[, "1"])
p_test_pos_given_disease

p_joint ← tab["1", "1"] / n
c(joint = p_joint,
  formula = p_test_pos_given_disease * mean(patients$disease == 1))
```

3.9.6 Output & Results

```
       disease
test         0    1
     0    9030   48
     1     472  450

[1] 0.4881    # P(disease | test+)
[1] 0.9036    # P(test+ | disease)

joint   formula
0.0450    0.0450
```

The product rule is verified: $P(\text{test}+ \cap \text{disease}) = P(\text{test}+ \mid \text{disease}) \cdot P(\text{disease})$. And $P(\text{disease} \mid \text{test}+) = 0.49$ – far below the sensitivity of 0.90, because disease prevalence is low.

3.9.7 Interpretation

The final ratio – 49% positive predictive value despite 90% sensitivity – is the classic base-rate lesson in clinical testing. In a manuscript: "positive predictive value was 49%, reflecting the low prevalence of disease in the screened population despite a sensitive test."

3.9.8 Practical Tips

- $P(A \mid B) \neq P(B \mid A)$ in general; this is the prosecutor's fallacy when confused in legal and clinical settings.
- Always check that the conditioning event has positive probability; in small samples, conditional probabilities based on empty strata are undefined.
- When interpreting a conditional probability, state the conditioning event explicitly: "the probability of disease *given that the test is positive*" is clearer than "the positive predictive value" for non-specialist readers.
- Chain the product rule when computing joint probabilities of many events: $P(A \cap B \cap C) = P(A)P(B \mid A)P(C \mid A \cap B)$.
- Conditioning changes the denominator: moving from marginal to conditional probability requires dividing by $P(B)$.

3.9.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.9.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem

- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.10 Convolutions of Distributions

3.10.1 Introduction

The distribution of a sum of independent random variables is their **convolution**. Many standard distribution families are closed under convolution: sums of independent normals are normal, sums of independent Poissons are Poisson, sums of independent gammas (with the same rate) are gamma. This closure is what makes convolution a core operation in probability.

3.10.2 Prerequisites

Joint distributions, independence, integration.

3.10.3 Theory

For independent X and Y with densities f_X and f_Y, the density of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x)\, dx.$$

The integral is the **convolution** $f_X * f_Y$. For discrete PMFs, the sum replaces the integral.

Characteristic functions convert convolutions to products: $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$. This makes convolutions of many distributions analytically tractable via Fourier-analytic methods.

Closure examples:

- $\mathcal{N}(\mu_1, \sigma_1^2) + \mathcal{N}(\mu_2, \sigma_2^2) \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ (independent).
- $\text{Poisson}(\lambda_1) + \text{Poisson}(\lambda_2) \sim \text{Poisson}(\lambda_1 + \lambda_2)$ (independent).
- $\text{Gamma}(\alpha_1, \beta) + \text{Gamma}(\alpha_2, \beta) \sim \text{Gamma}(\alpha_1 + \alpha_2, \beta)$ (same rate; independent).
- $\chi^2_{n_1} + \chi^2_{n_2} \sim \chi^2_{n_1+n_2}$ (independent).
- $\text{Binomial}(n_1, p) + \text{Binomial}(n_2, p) \sim \text{Binomial}(n_1 + n_2, p)$ (same p; independent).

Note that the Cauchy family is also closed under sums: sums of independent Cauchys are Cauchy – a consequence of the stable distribution family to which Cauchy belongs.

3.10.4 Assumptions

Independence is essential; the convolution formula fails for dependent X and Y.

3.10.5 R Implementation

```r
set.seed(2026)

# Normal + normal is normal
X ← rnorm(1e5, mean = 0, sd = 1)
Y ← rnorm(1e5, mean = 0, sd = 1)
Z ← X + Y
c(emp_mean = mean(Z), emp_sd = sd(Z),
  theoretical_mean = 0, theoretical_sd = sqrt(2))

# Poisson + Poisson is Poisson
X_pois ← rpois(1e5, lambda = 3)
Y_pois ← rpois(1e5, lambda = 2)
Z_pois ← X_pois + Y_pois
c(emp_mean = mean(Z_pois),
  emp_var  = var(Z_pois),
  theoretical_mean = 5,
  theoretical_var  = 5)

# Numerical convolution of two PDFs via integrate()
# f_X: uniform[0,1], f_Y: exponential(rate=1)
# Compute density of Z = X + Y at z = 0.5
f_convolve ← function(z) {
  integrate(function(x) dunif(x) * dexp(z - x, rate = 1),
            lower = 0, upper = min(1, z))$value
}
f_convolve(0.5)
f_convolve(1.5)
```

3.10.6 Output & Results

```
  emp_mean    emp_sd  theoretical_mean  theoretical_sd
     0.006    1.414             0.000          1.414

   emp_mean   emp_var  theoretical_mean  theoretical_var
      5.002     5.014             5.000            5.000

[1] 0.3935    # density at z = 0.5
[1] 0.3894    # density at z = 1.5
```

Empirical sums match theoretical distributions for both the continuous (normal) and discrete (Poisson) closure cases; numerical convolution computes the sum density for a non-standard case.

3.10.7 Interpretation

Convolutions underlie every sum-of-variables calculation in probability. The Gaussian CLT, at its core, says that sums of iid variables with finite variance converge (after standardisation) to the normal regardless of the component distributions.

3.10.8 Practical Tips

- For explicit convolutions of standard distributions, use the characteristic-function approach: sums of independent stable distributions are again stable.

- Numerical convolutions can be computed via FFT (`fft` in R) for efficiency on large grids.
- In discrete cases, direct summation often suffices: `outer()` followed by appropriate tabulation.
- Convolutions preserve location, scale, shape of stable families but generally destroy non-stable structure (skewness, kurtosis drift across convolutions).
- For CLT-like arguments, the speed of convergence to normal depends on the finite third moment (Berry-Esseen inequality).

3.10.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.10.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.11 Copulas: An Introduction

3.11.1 Introduction

A copula is a multivariate distribution with uniform marginals. It captures the dependence structure of a joint distribution independently of the marginals. Sklar's theorem says every joint distribution decomposes into its marginals plus a unique copula; this decomposition is the cleanest way to model dependence when the marginals are non-normal.

3.11.2 Prerequisites

Joint and marginal distributions; cumulative distribution function; basic multivariate probability.

3.11.3 Theory

Sklar's theorem (1959). For any joint CDF F_{XY} with marginals F_X, F_Y, there exists a copula $C : [0,1]^2 \to [0,1]$ such that

$$F_{XY}(x, y) = C(F_X(x), F_Y(y)).$$

If F_X, F_Y are continuous, C is unique. In the other direction, for any copula C and marginals F_X, F_Y, the composition defines a valid joint CDF.

Common bivariate copulas:

- **Gaussian copula**: induced by a bivariate normal with correlation ρ. Symmetric, no tail dependence.
- **t-copula**: from a bivariate t. Same symmetric structure as Gaussian but with heavier tail dependence.
- **Clayton copula**: asymmetric lower-tail dependence.
- **Gumbel copula**: asymmetric upper-tail dependence.
- **Frank copula**: symmetric, no tail dependence, but wider range of dependence than Gaussian.

Tail dependence: the lower and upper tail-dependence coefficients $\lambda_L = \lim_{u \to 0^+} C(u,u)/u$ and $\lambda_U = \lim_{u \to 1^-} (1 - 2u + C(u,u))/(1-u)$ quantify how strongly extreme values of one variable co-occur with extremes of the other.

The **probability integral transform** turns any continuous marginal into a uniform: $U = F_X(X) \sim \text{Uniform}(0,1)$. Copulas operate on the uniform scale and transform back to the original scale via F_X^{-1}.

3.11.4 Assumptions

Continuous marginals (for uniqueness); otherwise the copula is unique only up to the joint probabilities of atoms.

3.11.5 R Implementation

```r
library(copula)

# Fit a Gaussian copula to bivariate data
set.seed(2026)
n <- 1000

# Simulate data with non-normal marginals but Gaussian copula
cop <- normalCopula(param = 0.7, dim = 2)
mvd <- mvdc(copula = cop,
            margins = c("gamma", "exp"),
            paramMargins = list(list(shape = 3, rate = 0.8),
                                list(rate = 0.3)))
X <- rMvdc(n, mvd)

# Visualise the dependence through the copula (uniform-scaled)
U <- pobs(X)   # empirical pseudo-observations on [0, 1]^2
plot(U, pch = 16, cex = 0.3, col = "#2A9D8F",
     main = "Empirical copula (pseudo-observations)",
     xlab = "U1", ylab = "U2")

# Fit a Gaussian copula
gc_fit <- fitCopula(normalCopula(dim = 2), data = U, method = "ml")
coef(gc_fit)

# Compare Clayton (lower tail dep) vs Gumbel (upper tail dep)
cc_fit <- fitCopula(claytonCopula(dim = 2), data = U, method = "ml")
gu_fit <- fitCopula(gumbelCopula(dim = 2), data = U, method = "ml")
c(gaussian = AIC(gc_fit), clayton = AIC(cc_fit), gumbel = AIC(gu_fit))
```

3.11.6 Output & Results

```
[1] 0.7036       # MLE of copula correlation recovers the true 0.7

gaussian  clayton  gumbel
  -822.1   -551.0  -683.8
```

Gaussian copula has the best AIC, matching the data-generating model. The Clayton and Gumbel copulas mis-specify the tail structure and fit worse.

3.11.7 Interpretation

Copula models are standard in finance (risk aggregation) and increasingly common in epidemiology (joint survival) and genetics (multivariate phenotypes). They let analysts specify marginals appropriate to each variable (gamma for concentrations, Beta for proportions) without forcing joint normality.

3.11.8 Practical Tips

- Do not fit copulas without first checking that the marginals are well-specified; a bad marginal confounds copula diagnosis.
- The Gaussian copula underestimates joint extremes; for risk applications use a t-copula or Archimedean copulas (Clayton, Gumbel).
- Empirical copulas are good exploratory tools; fit parametric copulas only after inspecting the pseudo-observation scatter.
- Multivariate copulas beyond dimension 3 are often built via vine copulas (`VineCopula` package); direct multivariate Archimedean copulas scale poorly.
- Copulas capture dependence, not marginal distributions; do not confuse copula correlation with Pearson correlation of the raw variables.

3.11.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.11.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.12 The Correlation Coefficient

3.12.1 Introduction

The correlation coefficient rescales covariance into a unitless number between -1 and $+1$, making comparisons across different measurement scales meaningful. A correlation of $+1$ means perfect positive linear dependence, -1 perfect negative, and 0 no linear association. It is the single most reported bivariate summary in applied research.

3.12.2 Prerequisites

Variance and covariance.

3.12.3 Theory

For random variables X, Y with finite non-zero variances,

$$\rho(X, Y) = \frac{\operatorname{Cov}(X, Y)}{\sigma_X \sigma_Y}.$$

Cauchy-Schwarz guarantees $\rho \in [-1, 1]$. Equality at ± 1 occurs if and only if $Y = aX + b$ almost surely with $\operatorname{sign}(a) = \operatorname{sign}(\rho)$.

The sample correlation coefficient (Pearson's r) is the plug-in estimator using sample covariance and SDs:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}.$$

Key caveats:

- $\rho = 0$ does **not** imply independence unless (X, Y) is jointly normal.
- ρ measures only the *linear* component of association; non-linear dependence is invisible to it.
- Outliers affect r disproportionately; a robust alternative is the Spearman rank correlation.

3.12.4 Assumptions

- Both variances finite and non-zero.
- Linearity is the interpretive assumption: ρ captures only linear dependence.

3.12.5 R Implementation

```r
library(mvtnorm)
set.seed(2026)

# Simulate bivariate normals at several correlation levels
rho_grid <- c(-0.8, -0.3, 0, 0.3, 0.8)
n <- 500
results <- sapply(rho_grid, function(rho) {
  Sigma <- matrix(c(1, rho, rho, 1), 2, 2)
  XY <- rmvnorm(n, sigma = Sigma)
  cor(XY[, 1], XY[, 2])
})
data.frame(true = rho_grid, empirical = round(results, 3))

# Zero correlation does not imply independence
X <- rnorm(1e4)
Y <- X^2 - 1      # quadratic, mean zero
cor(X, Y)

# Outlier sensitivity
X <- rnorm(100); Y <- 0.3 * X + rnorm(100)
c(no_outlier = cor(X, Y),
  with_outlier = cor(c(X, 20), c(Y, -20)))
```

3.12.6 Output & Results

```
  true  empirical
1 -0.8     -0.790
2 -0.3     -0.307
3  0.0     -0.023
4  0.3      0.309
5  0.8      0.815

[1] 0.005          # zero Pearson despite clear dependence

no_outlier with_outlier
     0.309       -0.059
```

Sample correlations recover the population values within Monte Carlo error. The $X, X^2 - 1$ example shows zero correlation with obvious dependence. A single outlying point flips the sign of a small-sample correlation.

3.12.7 Interpretation

Reported in a paper as "r = .68 (95% CI .56-.77), $p < .001$", the correlation conveys both direction and magnitude of linear association. Always accompany r with a scatter plot to verify linearity; never report r alone.

3.12.8 Practical Tips

- Use Pearson only for roughly linear relationships between continuous variables.
- For monotone non-linear association, use Spearman's rho or Kendall's tau.
- A 95% CI for ρ uses the Fisher z-transformation: atanh(r) has an approximately normal sampling distribution.

- Correlation does not imply causation; confounders can create strong associations between unrelated variables.
- In multivariate data, a single correlation matrix can be visualised with `corrplot::corrplot()` or `ggcorrplot::ggcorrplot()`; always plot the matrix and the pairwise scatters.

3.12.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.12.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.13 The Cumulative Distribution Function

3.13.1 Introduction

The cumulative distribution function (CDF) is the one representation that fits every random variable, whether discrete, continuous, or mixed. It encodes the distribution completely and supports every computation that a PMF or PDF does.

3.13.2 Prerequisites

Random variables, basic probability.

3.13.3 Theory

The CDF of a random variable X is

$$F_X(x) = P(X \le x), \qquad x \in \mathbb{R}.$$

Every CDF has the following properties:

1. **Monotonic non-decreasing**: $x_1 \le x_2 \Rightarrow F_X(x_1) \le F_X(x_2)$.
2. **Right-continuous**: $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$.
3. **Limits**: $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to +\infty} F_X(x) = 1$.

Conversely, any function with these three properties defines a unique probability distribution on $\mathbb{R}$.

Key computations:

- Interval probability: $P(a < X \leq b) = F_X(b) - F_X(a)$.
- PDF (continuous X): $f_X(x) = F_X'(x)$.
- PMF (discrete X): $p_X(x) = F_X(x) - F_X(x^-)$, the jump size.
- Quantile function: $Q_X(u) = \inf\{x : F_X(x) \geq u\}$, the (generalised) inverse of F_X.

In R, each distribution family has a `p*()` function returning the CDF; the quantile function is `q*()`. The empirical CDF is obtained with `ecdf()`.

3.13.4 Assumptions

None beyond a well-defined probability measure.

3.13.5 R Implementation

```r
library(ggplot2)

set.seed(2026)
x ← rnorm(200, mean = 75, sd = 10)

# Empirical CDF compared to theoretical CDF
Fn ← ecdf(x)
grid ← seq(50, 100, length.out = 401)
df ← data.frame(
  x       = grid,
  F_empir  = Fn(grid),
  F_theor  = pnorm(grid, mean = 75, sd = 10)
)

ggplot(df, aes(x = x)) +
  geom_step(aes(y = F_empir), colour = "#2A9D8F") +
  geom_line(aes(y = F_theor), colour = "#F4A261", linewidth = 1) +
  labs(y = "CDF", title = "Empirical (step) vs theoretical (line) CDF") +
  theme_minimal()

# Interval probability via the CDF
P_70_to_80 ← pnorm(80, 75, 10) - pnorm(70, 75, 10)
P_70_to_80

# Quantile: the value below which a given probability lies
qnorm(0.975, 75, 10)
```

3.13.6 Output & Results

```
[1] 0.3829        # P(70 < X ⩽ 80)
[1] 94.6          # 97.5th percentile
```

The empirical step-function is close to the theoretical normal CDF, and the differences narrow as n grows (Glivenko-Cantelli).

3.13.7 Interpretation

In a paper, the CDF appears most often as a figure showing the empirical CDF against a theoretical reference (Q-Q plot is a rotation of this idea). Interval probabilities computed from the CDF are the standard form of reporting "percentage within the normal range".

3.13.8 Practical Tips

- The CDF is the unique object that characterises the distribution; two variables with the same CDF have the same distribution.
- Right-continuity vs left-continuity matters only at discrete atoms; the convention is right-continuous in most modern texts.
- The quantile function inverts the CDF; for continuous X, it is the ordinary functional inverse.
- The inverse-CDF method generates random draws: $X = F^{-1}(U)$ where $U \sim \text{Uniform}(0, 1)$.
- For mixed distributions, the CDF is the most transparent representation; PDFs and PMFs require generalised definitions.

3.13.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.13.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.14 Discrete vs Continuous Variables

3.14.1 Introduction

Random variables come in two canonical kinds – **discrete** (countable support) and **continuous** (uncountable support with a density). Most textbook distributions belong to one or the other. A subtler third category, **mixed**, combines discrete atoms with continuous parts and arises naturally in censored, truncated, or zero-inflated data.

3.14.2 Prerequisites

Random variables, PMF, and PDF.

3.14.3 Theory

A random variable X is:

- **Discrete** if it takes values in a countable set $\{x_1, x_2, \ldots\}$. Its distribution is described by the **probability mass function** $p(x_i) = P(X = x_i)$, with $\sum_i p(x_i) = 1$.
- **Continuous** (absolutely continuous) if there exists a **probability density function** f such that $P(X \in A) = \int_A f(x)\,dx$ for every measurable set A. For continuous X, $P(X = x) = 0$ for every single point x.
- **Mixed** if its distribution has both discrete atoms (points with positive probability mass) and a continuous part. Examples: censored survival times (mass at the censoring time, density elsewhere); zero-inflated counts (mass at 0, positive distribution above).

The CDF $F_X(x) = P(X \leq x)$ is defined for all three. For discrete X, F is a step function; for continuous X, F is smooth; for mixed X, F is piecewise smooth with jumps at the atoms.

The mean and variance are defined the same way (as $E[X]$ and $E[(X - E[X])^2]$) across the three cases, computed via sums for discrete and integrals for continuous.

3.14.4 Assumptions

The classification is mathematical rather than empirical; *any* random variable falls into one of the three categories.

3.14.5 R Implementation

```
library(ggplot2)
set.seed(2026)

# Discrete: Poisson counts
dis ← rpois(5000, lambda = 3)
table(dis)[1:7]

# Continuous: exponential waiting times
con ← rexp(5000, rate = 0.5)
quantile(con, c(0.25, 0.5, 0.75))

# Mixed: censored normal (values below 0 set to exactly 0)
mix ← pmax(rnorm(5000, mean = 1.0, sd = 1.5), 0)
c(mass_at_zero = mean(mix == 0),
  mean_overall = mean(mix))

# Plot all three
df ← rbind(
  data.frame(x = dis,    kind = "discrete (Poisson)"),
  data.frame(x = con,    kind = "continuous (exponential)"),
  data.frame(x = mix,    kind = "mixed (censored normal)")
)

ggplot(df, aes(x = x)) +
  geom_histogram(bins = 40, fill = "#2A9D8F") +
  facet_wrap(~ kind, scales = "free", ncol = 1) +
  theme_minimal()
```

3.14.6 Output & Results

```
dis       0      1      2      3      4      5      6
count   243    746   1124   1128    835    505    275

      25%       50%       75%
 0.5675    1.3735    2.7711

mass_at_zero  mean_overall
      0.2526        1.1342
```

The Poisson histogram has bars at integer values (discrete), the exponential has a smooth decaying curve (continuous), and the censored-normal has a spike at zero plus a continuous upper tail (mixed).

3.14.7 Interpretation

In reporting, the kind of random variable determines appropriate summaries:

- Discrete: report counts, proportions, mode, or a PMF table.
- Continuous: report mean, SD, quantiles, or a histogram/density.
- Mixed: report the atom probability separately from the continuous-part summary.

A paper analysing a zero-inflated outcome should acknowledge the mixed distribution explicitly rather than forcing it into a continuous or discrete mould.

3.14.8 Practical Tips

- Count data that "cap out" at a large value are often mixed; report the probability of the cap and analyse the remainder conditionally.
- Censored survival times are mixed; classical survival methods (Kaplan-Meier, Cox) handle the mixture natively.
- Continuous variables that are rounded to, say, integer values are still "almost continuous" in practice; treat as continuous for inference.
- Zero-inflated Poisson (`pscl::zeroinfl`) is the canonical regression model for mixed count data.
- When in doubt, plot a histogram (or ECDF) and inspect for atoms; the presence of a spike at a specific value signals a mixed distribution.

3.14.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.14.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.15 Expectation of a Random Variable

3.15.1 Introduction

The **expectation** $E[X]$ is the probability-weighted average of the values a random variable can take. It is the centre of gravity of the distribution, the long-run average of repeated independent draws, and the first moment in the moment hierarchy. Along with variance, it is the single most commonly reported summary of a distribution.

3.15.2 Prerequisites

Random variables, PMF or PDF, basic summation and integration.

3.15.3 Theory

For a discrete random variable with PMF p_X:

$$E[X] = \sum_x x\, p_X(x).$$

For a continuous random variable with PDF f_X:

$$E[X] = \int x\, f_X(x)\, dx.$$

Law of the Unconscious Statistician (LOTUS): for a measurable function g,

$$E[g(X)] = \sum_x g(x) p_X(x) \quad \text{or} \quad \int g(x) f_X(x)\, dx,$$

without needing to find the distribution of $g(X)$ first.

Linearity. Expectation is linear regardless of dependence:

$$E[aX + bY] = aE[X] + bE[Y], \qquad E\left[\sum_i X_i\right] = \sum_i E[X_i].$$

Product rule for independent variables: $E[XY] = E[X]E[Y]$ when $X \perp Y$.

Monotonicity: $X \leq Y$ (pointwise) $\Rightarrow E[X] \leq E[Y]$.

Expectation need not exist. The Cauchy distribution has no finite mean because the integral diverges. Existence requires $E[|X|] < \infty$.

3.15.4 Assumptions

- $E[|X|] < \infty$ for the expectation to be well-defined.
- Fubini/Tonelli conditions for interchanging integration and summation when needed (automatic for non-negative variables and for integrable ones).

3.15.5 R Implementation

```r
set.seed(2026)

# Exact expectation for discrete RV: binomial
n ← 10; p ← 0.3
k ← 0:n
E_X ← sum(k * dbinom(k, n, p))
c(formula = E_X, closed_form = n * p)

# Exact expectation for continuous RV: gamma
integrate(function(x) x * dgamma(x, shape = 3, rate = 0.5), 0, Inf)
3 / 0.5   # closed form alpha/beta

# LOTUS: E[g(X)] without the distribution of g(X)
# For X ~ Gamma(3, 0.5), compute E[log(X)]
integrate(function(x) log(x) * dgamma(x, shape = 3, rate = 0.5), 0, Inf)

# Monte Carlo estimate (LLN justifies this)
X ← rgamma(1e5, shape = 3, rate = 0.5)
mean(log(X))

# Linearity check: E[2X + 3Y] = 2E[X] + 3E[Y] regardless of dependence
X ← rnorm(1e5, 5, 2)
Y ← 0.7 * X + rnorm(1e5, 0, 1)        # dependent
c(direct = mean(2 * X + 3 * Y),
  sum_of_parts = 2 * mean(X) + 3 * mean(Y))
```

3.15.6 Output & Results

```
   formula  closed_form
       3.0          3.0

6 with absolute error < 8.5e-05

[1] 1.506         # E[log X] via integration
[1] 1.505         # same via Monte Carlo

     direct  sum_of_parts
     27.43         27.43
```

Closed-form, numerical integration, and Monte Carlo all agree, and linearity of expectation holds despite dependence between X and Y.

3.15.7 Interpretation

When a paper reports a mean, it is (implicitly) reporting an estimate of the expectation of a random variable. The sample mean converges to the expectation as n grows (law of large numbers), so inferences about population-level expectations derive their validity from this convergence.

3.15.8 Practical Tips

- Linearity is unconditional on dependence; this is the single most useful fact about expectation. It underlies the unbiasedness of the sample mean regardless of the

correlation structure.
- The product rule requires independence: $E[XY] = E[X]E[Y]$ fails when X, Y are dependent.
- Jensen's inequality: $E[g(X)] \geq g(E[X])$ for convex g (reversed for concave). Taking a log of an expectation is not the expectation of the log.
- For heavy-tailed distributions (Cauchy, stable with index < 1), the mean may not exist; rely on medians instead.
- Monte Carlo estimation of expectations works for any integrable X; convergence rate is $O(n^{-1/2})$ (CLT).

3.15.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.15.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.16 The Exponential Distribution

3.16.1 Introduction

The exponential distribution models waiting times between events in a Poisson process with constant rate. It is the only continuous distribution with the memoryless property, and it is the simplest parametric model in survival and reliability analysis.

3.16.2 Prerequisites

Continuous random variables and the Poisson process.

3.16.3 Theory

A non-negative random variable X is **exponential with rate** $\lambda > 0$ if

$$f(x) = \lambda e^{-\lambda x}, \qquad x \geq 0.$$

CDF: $F(x) = 1 - e^{-\lambda x}$. Survival function: $S(x) = e^{-\lambda x}$.

Moments:

- $E[X] = 1/\lambda$ (the mean waiting time).

- $\mathrm{Var}(X) = 1/\lambda^2$.

Memoryless property: $P(X > s + t \mid X > s) = P(X > t)$. The probability of waiting another t units, given we have already waited s, is the same as the unconditional probability of waiting t.

Constant hazard: $h(x) = f(x)/S(x) = \lambda$. The instantaneous event rate does not depend on how long the object has already survived.

Relationship to Poisson: in a Poisson process with rate λ, inter-arrival times are iid exponential with rate λ.

Scale vs rate: some R functions use rate (λ), others use scale ($1/\lambda$). Check the documentation.

3.16.4 Assumptions

Constant hazard (no ageing / no improvement). Violations motivate the Weibull or log-normal distributions.

3.16.5 R Implementation

```r
lambda ← 0.5

X ← rexp(1e5, rate = lambda)
c(mean = mean(X), theoretical = 1 / lambda,
  var  = var(X),  theoretical_var = 1 / lambda^2)

# Memoryless property
P_past_2 ← mean(X > 2)
P_past_5_given_2 ← mean(X[X > 2] > 5)        # P(X > 5 | X > 2)
P_past_3 ← mean(X > 3)                        # should equal above by memorylessness
c(conditional = P_past_5_given_2, unconditional = P_past_3)

# Constant hazard
breaks ← seq(0, 10, by = 1)
hazards ← sapply(seq_len(length(breaks) - 1), function(i) {
  in_bin ← X ⩾ breaks[i] & X < breaks[i + 1]
  sum(in_bin) / sum(X ⩾ breaks[i])
})
round(hazards, 3)
```

3.16.6 Output & Results

```
    mean theoretical      var  theoretical_var
    1.996        2.0    3.985              4.0

   conditional  unconditional
        0.2232         0.2229

  [1] 0.394 0.392 0.392 0.400 0.391 0.395 0.394 0.389 0.399 0.394
```

Empirical mean and variance match theoretical; memoryless property holds; empirical hazard is constant at $\approx 1 - e^{-\lambda}$ per unit interval.

3.16.7 Interpretation

Exponential survival is the simplest hypothesis one can test in reliability and survival. When observed hazard is non-constant, report the Weibull or a flexible parametric model instead.

3.16.8 Practical Tips

- `rexp(n, rate)` in R uses the rate parameterisation; `scale = 1/rate`.
- Log-transform to compress the heavy right tail for visualisation.
- Constant hazard is a strong and testable assumption; plot $-\log S(t)$ against t; linearity indicates exponentiality.
- For right-censored data, use survival-analysis tools (`survreg(..., dist = "exponential")`) rather than naive `rexp`.
- Exponential is the discrete geometric's continuous analogue; both share memorylessness.

3.16.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.16.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.17 The F Distribution

3.17.1 Introduction

The F distribution arises as the ratio of two independent scaled chi-squared random variables. It is the reference distribution for the F-statistic in ANOVA, for tests of variance equality between two normal samples, and for omnibus tests in linear regression. Its shape is controlled by two degree-of-freedom parameters, one for the numerator and one for the denominator.

3.17.2 Prerequisites

Chi-squared distribution.

3.17.3 Theory

If $U \sim \chi^2_{d_1}$ and $V \sim \chi^2_{d_2}$ are independent, then

$$F = \frac{U/d_1}{V/d_2} \sim F_{d_1, d_2}.$$

The numerator df is d_1; the denominator df is d_2.

Moments (when they exist):

- $E[F] = d_2/(d_2 - 2)$ for $d_2 > 2$.
- $\operatorname{Var}(F) = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2-2)^2(d_2-4)}$ for $d_2 > 4$.

Relationships to other distributions:

- $1/F_{d_1,d_2} \sim F_{d_2,d_1}$.
- $t_\nu^2 \sim F_{1,\nu}$.
- As $d_2 \to \infty$, $d_1 \cdot F_{d_1,d_2} \to \chi^2_{d_1}$.

One-way ANOVA: with k groups and n total observations, the F-statistic

$$F = \frac{\mathrm{MS}_{\text{between}}}{\mathrm{MS}_{\text{within}}}$$

follows $F_{k-1, n-k}$ under the null of equal means.

Variance ratio test: for two independent normal samples with variances $\sigma_1^2 = \sigma_2^2$,

$$F = s_1^2 / s_2^2 \sim F_{n_1 - 1, n_2 - 1}.$$

3.17.4 Assumptions

Independence between U and V; normality of the underlying data for the derived sample statistics.

3.17.5 R Implementation

```r
# PDF visualisation
x ← seq(0, 5, length.out = 400)
plot(x, df(x, 5, 20), type = "l", col = "#2A9D8F", lwd = 2,
     main = "F(5, 20) density", ylab = "f(x)")
lines(x, df(x, 5, 5),   col = "#F4A261", lwd = 2)
lines(x, df(x, 20, 20), col = "#6A4C93", lwd = 2)
legend("topright", c("F(5,20)", "F(5,5)", "F(20,20)"),
       col = c("#2A9D8F", "#F4A261", "#6A4C93"), lwd = 2)

# Verify construction
n ← 1e5; d1 ← 4; d2 ← 10
U ← rchisq(n, df = d1); V ← rchisq(n, df = d2)
F_sim ← (U/d1) / (V/d2)
c(emp_mean = mean(F_sim), theoretical = d2 / (d2 - 2))

# F-statistic from ANOVA
set.seed(2026)
df ← data.frame(y = c(rnorm(30, 50, 8), rnorm(30, 55, 8), rnorm(30, 53, 8)),
                g = factor(rep(c("A", "B", "C"), each = 30)))
anova(lm(y ~ g, data = df))
```

3.17.6 Output & Results

```
emp_mean theoretical
    1.252      1.250

Analysis of Variance Table

Response: y
           Df Sum Sq Mean Sq F value  Pr(>F)
g           2  394.7   197.3    3.57 0.03218 *
Residuals 87 4811.2    55.3
```

The empirical F ratio recovers the theoretical mean, and the ANOVA F-statistic of 3.57 on (2, 87) df has p-value 0.032.

3.17.7 Interpretation

Almost every ANOVA and linear-model-comparison result in applied statistics is reported as an F-statistic with numerator and denominator df. Reporting “F(2, 87) = 3.57, p = 0.032” is a standard form.

3.17.8 Practical Tips

- Order the df correctly: F(df1, df2) is not the same as F(df2, df1) in general.
- For two-sided variance tests, some software uses max/min of s_1^2/s_2^2 to avoid ambiguity.
- Non-central F (with non-centrality parameter λ) governs power calculations for ANOVA.
- The F-distribution is highly sensitive to non-normality; for skewed data prefer Welch’s ANOVA or non-parametric alternatives.

- `qf(0.95, d1, d2)` gives the critical value for a one-sided 5% test.

3.17.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.17.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.18 The Gamma Distribution

3.18.1 Introduction

The gamma distribution is a flexible right-skewed distribution on positive real numbers. It is parameterised by a shape and a rate (or scale) and includes the exponential, Erlang, and chi-squared as special cases. In applied work it models waiting times, concentrations, and response amplitudes – any positive continuous variable with light-to-moderate right skew.

3.18.2 Prerequisites

Exponential distribution, basic calculus.

3.18.3 Theory

A gamma random variable with shape $\alpha > 0$ and rate $\beta > 0$ has PDF

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x > 0.$$

Scale parameterisation: $\theta = 1/\beta$, so $f(x) = \frac{1}{\theta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\theta}$.

Moments:

- $E[X] = \alpha/\beta$ (rate form) = $\alpha\theta$ (scale form).
- $\mathrm{Var}(X) = \alpha/\beta^2 = \alpha\theta^2$.

Special cases:

- Shape $\alpha = 1$: exponential(β).

- Integer shape $\alpha = k$: Erlang distribution, which equals the sum of k iid exponentials with rate β.
- Shape $\alpha = \nu/2$, rate $\beta = 1/2$: chi-squared with ν df.

Sum property: independent Gamma(α_1, β) and Gamma(α_2, β) sum to Gamma$(\alpha_1 + \alpha_2, \beta)$.

Conjugacy: gamma is conjugate to the Poisson likelihood (prior on λ); gamma-Poisson mixture yields negative binomial.

3.18.4 Assumptions

Values must be positive. The flexibility comes from the two parameters; monotone hazard (increasing for $\alpha > 1$, decreasing for $\alpha < 1$, constant for $\alpha = 1$).

3.18.5 R Implementation

```r
library(MASS)

shape ← 2.5; rate ← 0.8

X ← rgamma(1e5, shape = shape, rate = rate)
c(mean = mean(X), theoretical_mean = shape / rate,
  var  = var(X),  theoretical_var = shape / rate^2)

# Visualise the PDF
curve(dgamma(x, shape, rate), from = 0, to = 15,
      main = paste("Gamma(shape =", shape, ", rate =", rate, ")"),
      xlab = "x", ylab = "f(x)", col = "#2A9D8F", lwd = 2)

# MLE fit
fit ← fitdistr(X, "gamma")
fit$estimate

# Sum of independent exponentials equals Erlang (gamma with integer shape)
k ← 5; lambda ← 1
S ← replicate(5000, sum(rexp(k, rate = lambda)))
c(mean_empirical = mean(S), mean_theoretical = k / lambda,
  var_empirical  = var(S),  var_theoretical  = k / lambda^2)
```

3.18.6 Output & Results

```
  mean theoretical_mean      var theoretical_var
  3.13             3.125     3.92            3.906

    shape    rate
  2.507   0.802       # MLE recovers true parameters

mean_empirical mean_theoretical  var_empirical  var_theoretical
         4.989            5.000          5.014            5.000
```

Empirical moments match theoretical; MLE recovers the true parameters; the sum of 5 iid exponentials has the predicted Erlang moments.

3.18.7 Interpretation

Gamma is a default parametric model for positive continuous variables that do not fit well to normal or log-normal. In generalised linear models, a gamma family (`glm(..., family = Gamma)`) is appropriate for positive continuous outcomes with variance proportional to the squared mean.

3.18.8 Practical Tips

- Watch R's `rgamma(n, shape, rate)` vs `rgamma(n, shape, scale = ...)`; only one of `rate` and `scale` should be given.
- Fitting a gamma to data: `MASS::fitdistr(x, "gamma")` or `fitdistrplus::fitdist`.
- Small shape ($\alpha < 1$) gives a distribution with a pole at zero; reasonable for heavily skewed data but breaks with zero values in a sample.
- For modelling variance in mixed models, the gamma distribution appears as the distribution of random-effect variances in certain hierarchical setups.
- Log-gamma (log of a gamma) is often approximately normal and easier to work with for some computations.

3.18.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.18.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.19 The Geometric Distribution

3.19.1 Introduction

The geometric distribution counts the number of Bernoulli trials required to obtain the first success. It is the discrete analogue of the exponential waiting-time distribution and shares the memoryless property: the probability of further trials to success does not depend on how many have already been tried.

3.19.2 Prerequisites

Bernoulli trials.

3.19.3 Theory

Two conventions exist. Let X be the number of Bernoulli(p) trials *up to and including* the first success:

$$P(X = k) = (1-p)^{k-1}p, \qquad k = 1, 2, 3, \ldots$$

Alternatively, Y is the number of failures *before* the first success:

$$P(Y = k) = (1-p)^{k}p, \qquad k = 0, 1, 2, \ldots$$

R uses the second convention (`rgeom`, `dgeom` return Y, i.e., failures before first success).

Moments for the first convention (X = trials until first success):

- $E[X] = 1/p$, $\mathrm{Var}(X) = (1-p)/p^2$.

For Y (R's convention): $E[Y] = (1-p)/p$, $\mathrm{Var}(Y) = (1-p)/p^2$.

Memoryless property: $P(X > m+n \mid X > m) = P(X > n)$. Having waited m trials without success does not change the distribution of further waiting time.

3.19.4 Assumptions

Independent Bernoulli trials with constant probability p.

3.19.5 R Implementation

```r
p ← 0.3

# R's convention: Y = failures before first success
Y ← rgeom(1e5, p)
c(mean_Y = mean(Y), theoretical = (1 - p) / p)

# Number of trials to first success X = Y + 1
X ← Y + 1
c(mean_X = mean(X), theoretical = 1 / p)

# PMF comparison
k ← 0:15
data.frame(k, pmf_Y = dgeom(k, p))

# Memoryless: P(Y > 5 + 3 | Y > 5) = P(Y > 3)
Y_past_5 ← Y[Y > 5]
c(conditional = mean(Y_past_5 > 5 + 3),
  unconditional = mean(Y > 3))
```

3.19.6 Output & Results

```
   mean_Y theoretical
    2.334       2.333

   mean_X theoretical
    3.334       3.333

    k      pmf_Y
1   0     0.3000
2   1     0.2100
3   2     0.1470
...

conditional  unconditional
     0.3441         0.3430
```

The memoryless property holds empirically: conditional probability of further 3 failures, given already 5 failures, equals the unconditional probability of at least 3 failures.

3.19.7 Interpretation

The geometric distribution is the baseline model for “time to event” in the discrete case: cycles to pregnancy in IVF, attempts before a successful procedure, trials before a rare positive test. Memorylessness is a strong assumption that often fails in real data where fatigue or learning changes p over time.

3.19.8 Practical Tips

- Check which convention (trials vs. failures) is expected when comparing to textbook formulas.
- Memorylessness is testable empirically: compare the distribution of residual trials given survival to each time.
- For non-memoryless discrete waiting times, the negative binomial (shape > 1) captures over-dispersion relative to geometric.
- The geometric converges to the exponential when discretising time with fine intervals.
- In clinical or biological settings, censoring is common; naive `rgeom` ignores right-censoring – use survival methods for realistic data.

3.19.9 For Reviewers

Use the shared reviewer checklist, together with this method’s assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript’s application of the method.

3.19.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.20 The Hazard Function

3.20.1 Introduction

The hazard function $h(t)$ gives the instantaneous rate at which events occur at time t, given that they have not occurred before t. It is the primary object in survival analysis and reliability engineering: proportional-hazards models, accelerated-failure-time models, and competing-risks methods all reason about hazards rather than distributions directly.

3.20.2 Prerequisites

Continuous distributions; survival function $S(t) = P(T > t)$.

3.20.3 Theory

For a non-negative continuous random variable T with density f and survival function $S(t) = 1 - F(t)$,

$$h(t) = \frac{f(t)}{S(t)}.$$

Intuitively, $h(t)\,dt$ is approximately $P(t \le T < t + dt \mid T \ge t)$ for small dt.

The **cumulative hazard** is

$$H(t) = \int_0^t h(u)\,du = -\log S(t).$$

Conversely, $S(t) = e^{-H(t)}$. Once either of f, F, S, h, H is known, the others follow.

Hazard shapes for common distributions:

- Exponential: $h(t) = \lambda$ (constant).
- Weibull with shape k: $h(t) = (k/\lambda)(t/\lambda)^{k-1}$ – increasing if $k > 1$, decreasing if $k < 1$.
- Gompertz: $h(t) = \alpha e^{\beta t}$ – exponentially increasing; common in human mortality.
- Log-normal: $h(t)$ rises then falls.

3.20.4 Assumptions

T is a non-negative continuous random variable with a density. Proportional-hazards models further assume $h(t \mid x) = h_0(t)\exp(\beta^\top x)$.

3.20.5 R Implementation

```r
library(survival); library(ggplot2)

# Constant hazard (exponential)
x <- seq(0.01, 5, length.out = 400)
h_exp <- rep(0.5, length(x))
h_weibull_inc <- (2.0 / 1.5) * (x / 1.5)^(2.0 - 1)
h_weibull_dec <- (0.5 / 1.5) * (x / 1.5)^(0.5 - 1)

df <- data.frame(
  t   = rep(x, 3),
  h   = c(h_exp, h_weibull_inc, h_weibull_dec),
  kind = factor(rep(c("exp lambda=0.5",
                      "weibull k=2 (incr)",
                      "weibull k=0.5 (decr)"), each = length(x)))
)

ggplot(df, aes(t, h, colour = kind)) +
  geom_line(linewidth = 1) +
  scale_colour_manual(values = c("#6A4C93", "#2A9D8F", "#F4A261")) +
  labs(x = "t", y = "h(t)", colour = "",
       title = "Hazard functions for three canonical models") +
  theme_minimal()

# Empirical cumulative hazard from censored data
set.seed(2026)
T <- rexp(100, rate = 0.3)
status <- rep(1, 100)
fit <- survfit(Surv(T, status) ~ 1)
plot(fit, fun = "cumhaz", col = "#2A9D8F", lwd = 2,
     xlab = "t", ylab = "H(t)",
     main = "Empirical cumulative hazard (Nelson-Aalen)")
abline(0, 0.3, col = "#F4A261", lty = 2)
```

3.20.6 Output & Results

The plot shows three hazards: exponential (flat), increasing Weibull, decreasing Weibull. The empirical cumulative hazard from 100 exponential event times is approximately linear with slope $\lambda = 0.3$, matching the theoretical $H(t) = 0.3t$.

3.20.7 Interpretation

In survival analysis, hazards are the natural target of regression: “the hazard of death was 40 % higher in the treated group than in controls (HR = 1.40, 95 % CI 1.10-1.78)”. The proportional-hazards model is ubiquitous because it avoids specifying the baseline hazard shape.

3.20.8 Practical Tips

- A constant hazard implies exponentiality; a log-log plot of $-\log S(t)$ against $\log t$ that is linear indicates Weibull.
- Empirical hazards are noisy at the tail where few subjects remain; report cumulative hazard (smoother) or kernel-smoothed hazard estimates.

- The **Nelson-Aalen** estimator is the standard non-parametric estimator of the cumulative hazard under right-censoring.
- In competing risks, cause-specific hazards differ from the sub-distribution hazard of Fine-Gray; the two answer different questions.
- Time-varying hazard ratios invalidate the proportional-hazards assumption; use stratification or time-interactions.

3.20.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.20.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.21 The Hypergeometric Distribution

3.21.1 Introduction

The hypergeometric distribution models the number of successes in a sample drawn without replacement from a finite population. It differs from the binomial by dependence across draws: removing a success reduces the probability of a subsequent success. It underlies Fisher's exact test and audit sampling.

3.21.2 Prerequisites

Combinatorics, binomial distribution.

3.21.3 Theory

A population of N items contains K "successes" and $N-K$ "failures". A sample of size n is drawn without replacement. Let X be the number of successes in the sample:

$$P(X=k)=\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}},\qquad \max(0,n-(N-K))\le k\le \min(n,K).$$

Moments:

- $E[X]=nK/N$ (same as binomial with $p=K/N$).
- $\operatorname{Var}(X)=n\cdot\frac{K(N-K)}{N^2}\cdot\frac{N-n}{N-1}$.

The factor $(N-n)/(N-1)$ is the **finite-population correction**; it shrinks the variance relative to binomial sampling, reflecting the reduced uncertainty when a substantial fraction of the population is sampled.

As $N \to \infty$ with $K/N = p$ fixed, hypergeometric converges to binomial: when the population is large relative to the sample, with/without replacement makes little difference.

3.21.4 Assumptions

- Finite population of known size N.
- Known number of successes K in the population.
- Simple random sampling without replacement.

3.21.5 R Implementation

```r
N ← 200; K ← 60; n ← 30

# PMF and moments
k ← 0:n
pmf ← dhyper(k, K, N - K, n)
c(mean = n * K / N,
  var  = n * K / N * (1 - K / N) * (N - n) / (N - 1),
  emp_mean = sum(k * pmf),
  emp_var  = sum((k - sum(k * pmf))^2 * pmf))

# Compare with binomial: same mean, larger variance
c(binom_var = n * (K / N) * (1 - K / N))

# Fisher's exact test uses the hypergeometric
tab ← matrix(c(12, 18, 48, 122), 2, 2)
fisher.test(tab)
```

3.21.6 Output & Results

```
     mean       var   emp_mean    emp_var
      9.0     5.487       9.00      5.487

binom_var
     6.30

        Fisher's Exact Test for Count Data

data:  tab
p-value = 0.1412
alternative hypothesis: true odds ratio is not equal to 1
```

The hypergeometric variance is smaller than the binomial variance (5.49 vs 6.30) because the finite-population correction accounts for dependence. Fisher's exact test in the 2×2 example is a hypergeometric computation under the null of independence.

3.21.7 Interpretation

Hypergeometric inference appears whenever sampling without replacement matters. In quality audits, the exact probability of a given number of defective items being observed in a sample is hypergeometric. In contingency-table analysis of small cells, Fisher's exact test is hypergeometric.

3.21.8 Practical Tips

- For large populations ($N > 20n$), binomial approximation is typically adequate and simpler.
- For 2×2 contingency tables, Fisher's exact test is hypergeometric-based and preferable to chi-squared with small expected counts.
- `dhyper(x, m, n, k)` uses R's quirky naming: `m` = successes in population, `n` = failures in population, `k` = sample size.
- In survey sampling with stratification, the within-stratum variance uses a finite-population correction factor.
- Generalisations to multiple categories yield the multivariate hypergeometric; useful in categorical audit and card-game probabilities.

3.21.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.21.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.22 Independence

3.22.1 Introduction

Two events are *independent* when knowing that one occurred tells you nothing about whether the other will. Independence is the property that makes simulation tractable, factorises joint distributions, and underlies the iid assumption of almost every classical statistical method. It is also one of the most commonly mis-stated concepts: independent is not the same as disjoint, and pairwise independence does not imply joint independence.

3.22.2 Prerequisites

Basic probability, conditional probability, and the product rule.

3.22.3 Theory

Two events A and B are **independent** if

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

Equivalently, when $P(B) > 0$: $P(A \mid B) = P(A)$. Knowing B does not update the probability of A.

For n events $A_1, \ldots, A_n$, two levels of independence exist:

- **Pairwise independence**: $P(A_i \cap A_j) = P(A_i)P(A_j)$ for every $i \neq j$.
- **Mutual independence**: every subset factorises. $P(A_{i_1} \cap \ldots \cap A_{i_k}) = \prod_{j=1}^{k} P(A_{i_j})$ for every subset.

Mutual is strictly stronger: it is possible for three events to be pairwise independent yet not mutually independent (Bernstein's example).

For random variables X and Y:

$$X \perp Y \iff F_{XY}(x, y) = F_X(x)F_Y(y) \quad \forall x, y,$$

or equivalently (when densities exist) $f_{XY}(x, y) = f_X(x)f_Y(y)$.

Independence vs. uncorrelatedness. Independence implies zero correlation, but not vice versa: X and X^2 for $X \sim \mathcal{N}(0, 1)$ are uncorrelated yet functionally dependent.

3.22.4 Assumptions

None beyond the probability space being well-defined.

3.22.5 R Implementation

```r
set.seed(2026)

# Two independent Bernoulli trials
n ← 1e5
A ← rbinom(n, 1, 0.3)
B ← rbinom(n, 1, 0.6)

# Check: P(A and B) ≈ P(A) * P(B)
c(joint = mean(A == 1 & B == 1),
  product = mean(A == 1) * mean(B == 1))

# Dependence: construct two correlated Bernoullis
C ← ifelse(A == 1, rbinom(n, 1, 0.9), rbinom(n, 1, 0.2))
c(joint = mean(A == 1 & C == 1),
  product = mean(A == 1) * mean(C == 1))

# Uncorrelated but dependent example
X ← rnorm(n)
Y ← X^2
c(correlation = cor(X, Y),
  joint_factorises = abs(mean(X < 0 & Y > 2) -
                         mean(X < 0) * mean(Y > 2)))
```

3.22.6 Output & Results

```
  joint  product
 0.1801   0.1800

  joint  product
 0.2714   0.1798

correlation joint_factorises
     0.0028           0.1589
```

First pair is independent (joint equals product). Second pair is dependent (joint much larger than product). Third shows zero correlation with clearly non-zero deviation from factorisation.

3.22.7 Interpretation

In study design, independence is usually a *design* property: separate random assignments ensure treatment is independent of unmeasured confounders; random selection ensures the sample is independent of other observable sources. When the design does not guarantee independence (clustered data, repeated measures), the statistical model must account for it.

3.22.8 Practical Tips

- Independence is a property of a particular probability measure P. Two events can be independent under one distribution and dependent under another.
- Pairwise-but-not-mutual independence is rare in applied settings but matters in cryptography and combinatorial probability.
- "iid" (independent and identically distributed) is the default assumption of most statistical methods; violations require mixed-effects models, GEE, or time-series methods.
- Use simulation to check independence when analytical factorisation is cumbersome; the deviation from $P(A)P(B)$ quantifies the dependence.
- For continuous variables, a formal independence test is available via distance correlation (`energy::dcor`).

3.22.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.22.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem

- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.23 Joint Distributions

3.23.1 Introduction

A joint distribution describes how two or more random variables behave together. It contains strictly more information than the marginals alone: the marginals tell you about each variable in isolation, while the joint tells you how they interact. Dependence, correlation, and conditional behaviour are all properties of the joint distribution.

3.23.2 Prerequisites

Random variables, PMF, PDF, independence.

3.23.3 Theory

For two random variables X and Y:

- **Joint CDF**: $F_{X,Y}(x, y) = P(X \le x, Y \le y)$.
- **Joint PMF** (discrete): $p_{X,Y}(x, y) = P(X = x, Y = y)$.
- **Joint PDF** (continuous): $f_{X,Y}(x, y)$ with $\iint f = 1$.

The joint CDF satisfies $\partial^2 F / \partial x \partial y = f$ (in the continuous case where derivatives exist).

Independence: $X \perp Y$ if and only if $f_{X,Y}(x, y) = f_X(x) f_Y(y)$ for all x, y (or the PMF equivalent). Equivalently, the joint CDF factorises.

The **bivariate normal** is the canonical joint distribution in applied work:

$$
\begin{aligned}
f_{X,Y}(x, y) = {} & \frac{1}{2\pi \sigma_X \sigma_Y \sqrt{1 - \rho^2}} \\
& \times \exp\Bigg\{ -\frac{1}{2(1 - \rho^2)} \Bigg[\frac{(x - \mu_X)^2}{\sigma_X^2} \\
& \quad - \frac{2\rho (x - \mu_X)(y - \mu_Y)}{\sigma_X \sigma_Y} + \frac{(y - \mu_Y)^2}{\sigma_Y^2} \Bigg] \Bigg\}.
\end{aligned}
$$

Its marginals, conditionals, and any linear combination are all normal.

3.23.4 Assumptions

A joint distribution is well-defined whenever the random variables live on a common probability space.

3.23.5 R Implementation

```r
library(mvtnorm); library(ggplot2)
set.seed(2026)

# Bivariate normal sample
Sigma ← matrix(c(1, 0.6,
                 0.6, 1), 2, 2)
XY ← rmvnorm(2000, mean = c(0, 0), sigma = Sigma)
df ← data.frame(X = XY[, 1], Y = XY[, 2])

ggplot(df, aes(X, Y)) +
  geom_point(alpha = 0.3, colour = "#2A9D8F") +
  geom_density_2d(colour = "#F4A261") +
  labs(title = "Bivariate normal with correlation 0.6") +
  theme_minimal()

# Joint density at a point
dmvnorm(c(0, 0), mean = c(0, 0), sigma = Sigma)

# Probability of a rectangle via the CDF
p_rect ← pmvnorm(lower = c(-1, -1), upper = c(1, 1),
                 mean = c(0, 0), sigma = Sigma)
p_rect
```

3.23.6 Output & Results

```
[1] 0.1989       # density at (0,0)
[1] 0.6372       # P(-1 ⩽ X ⩽ 1, -1 ⩽ Y ⩽ 1)
```

The rectangle has probability 0.64 under the bivariate normal with correlation 0.6 – higher than the product of the marginal rectangle probabilities (0.68 × 0.68 = 0.47) because the variables are positively correlated.

3.23.7 Interpretation

Most multivariate methods assume some form of joint distribution. Multivariate normal is the default assumption in linear discriminant analysis, MANOVA, and Gaussian graphical models. When the joint distribution is clearly non-normal, consider copula models or non-parametric density estimation.

3.23.8 Practical Tips

- Visualise joint distributions with 2D density plots, hexbin plots, or pairs plots; scatter plots alone can miss cluster structure.
- The marginals never determine the joint distribution; two datasets with identical marginals can have very different bivariate structures (Anscombe's quartet generalisation).
- For continuous data, `MASS::kde2d` and `ggplot2::stat_density_2d` give kernel joint densities.
- Independence tests: for continuous data use Hoeffding's D (`Hmisc::hoeffd`) or distance correlation (`energy::dcor`); both detect non-linear dependence Pearson

misses.
- Joint discrete distributions are usually tabulated as contingency tables; chi-squared tests of independence are the canonical hypothesis tests.

3.23.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.23.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.24 Kolmogorov's Axioms of Probability

3.24.1 Introduction

Modern probability theory rests on three axioms proposed by Andrey Kolmogorov in 1933. They turn probability from a collection of intuitions into a rigorous mathematical theory consistent with measure theory. Every theorem about probability – the law of large numbers, the central limit theorem, Bayes' rule – is ultimately a consequence of these three rules and the definitions that accompany them.

3.24.2 Prerequisites

The reader should know what a set is and should be comfortable with the basic set operations (union, intersection, complement).

3.24.3 Theory

A **probability space** is the triple $(\Omega, \mathcal{F}, P)$, where:

- Ω is the **sample space**, the set of all possible outcomes.
- $\mathcal{F}$ is a **sigma-algebra** of subsets of Ω – the collection of events to which we assign probabilities. It is closed under complements and countable unions.
- P is the **probability measure**, a function $P : \mathcal{F} \to [0, 1]$ satisfying Kolmogorov's three axioms:

1. **Non-negativity**: $P(A) \geq 0$ for every $A \in \mathcal{F}$.
2. **Normalisation**: $P(\Omega) = 1$.
3. **Countable additivity**: for any sequence of pairwise disjoint events $A_1, A_2, \ldots$ in $\mathcal{F}$, $P(\bigcup_i A_i) = \sum_i P(A_i)$.

From these axioms every familiar property follows:

- $P(\emptyset) = 0$.
- $P(A^c) = 1 - P(A)$ (complement rule).
- If $A \subseteq B$ then $P(A) \leq P(B)$ (monotonicity).
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ (inclusion-exclusion).

The axioms are the minimal, sufficient rules. Violating any of them produces a theory that is either inconsistent or fails to capture behaviour real data exhibit.

3.24.4 Assumptions

The axioms themselves are assumptions; they cannot be derived from more primitive principles. Their justification is empirical: they match the long-run frequencies observed in real random experiments.

3.24.5 R Implementation

We can verify the axioms empirically by simulation for a simple probability space.

```
set.seed(2026)

# Fair six-sided die: Omega = {1, ..., 6}
rolls ← sample(1:6, size = 1e5, replace = TRUE)

# Axiom 1: every event probability in [0, 1]
empirical_p ← prop.table(table(rolls))
all(empirical_p ⩾ 0 & empirical_p ⩽ 1)

# Axiom 2: probability of the whole sample space is 1
mean(rolls %in% 1:6)

# Axiom 3: additivity for disjoint events
A ← rolls %in% c(1, 2)   # "low" outcomes
B ← rolls %in% c(3, 4)   # "mid" outcomes
# A and B are disjoint; P(A or B) should equal P(A) + P(B)
c(p_union = mean(A | B),
  p_A_plus_p_B = mean(A) + mean(B))

# Derived property: complement rule
c(p_not_A = mean(!A),
  one_minus_p_A = 1 - mean(A))

# Inclusion-exclusion for non-disjoint events
C ← rolls %in% c(2, 3, 4)   # overlaps with A
c(p_union = mean(A | C),
  formula = mean(A) + mean(C) - mean(A & C))
```

3.24.6 Output & Results

```
[1] TRUE          # all empirical probabilities in [0, 1]

[1] 1             # P(Omega) = 1

  p_union p_A_plus_p_B
   0.6661       0.6661

  p_not_A one_minus_p_A
   0.6669        0.6669

  p_union  formula
   0.6669   0.6669
```

Empirical frequencies converge to axiomatic probabilities as n grows; the axioms are satisfied by the simulated frequencies.

3.24.7 Interpretation

The axioms are not something one "reports" in a paper; they are the framework within which every probabilistic claim is made. When a manuscript states $P(\text{event}) = 0.23$, it assumes a probability space has been defined implicitly and that P is a valid measure on it.

3.24.8 Practical Tips

- In applied work the sample space is usually obvious (outcomes of a trial, categories of a classification), and the sigma-algebra is the full power set. Measure-theoretic niceties matter only in continuous or infinite settings.
- Countable additivity is strictly stronger than finite additivity; the two agree for finite sample spaces.
- Empirically estimated probabilities (sample proportions) satisfy the axioms by construction, but converge to the true probability only in the limit (law of large numbers).
- When combining events, always check disjointness before adding probabilities; use inclusion-exclusion when events overlap.
- The axioms say nothing about *interpretation* of probability (frequentist, Bayesian, logical); they are compatible with all three.

3.24.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.24.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.25 Law of Total Probability

3.25.1 Introduction

The law of total probability (LTP) expresses the unconditional probability of an event as a weighted sum of conditional probabilities. It is the bookkeeping rule for "conditioning on every possibility and adding them up", and it appears as the denominator in Bayes' theorem, as the marginalisation step in mixture models, and whenever a probability computation is easier after splitting on a latent variable.

3.25.2 Prerequisites

Conditional probability and set partitions.

3.25.3 Theory

Let $\{B_1, B_2, \ldots, B_n\}$ be a **partition** of the sample space: the events are pairwise disjoint and their union is Ω. Then for any event A,

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)\, P(B_i).$$

Proof sketch: $A = \bigcup_i (A \cap B_i)$ where the intersections are disjoint. By additivity, $P(A) = \sum_i P(A \cap B_i)$, and each $P(A \cap B_i) = P(A \mid B_i) P(B_i)$.

In continuous form, if X is a continuous random variable and A is an event,

$$P(A) = \int P(A \mid X = x) f_X(x)\, dx,$$

which is the marginalisation used to compute the probability of an event under a mixture or hierarchical model.

3.25.4 Assumptions

- The B_i form a partition: disjoint, exhaustive, positive probability each.
- Otherwise standard probability-space assumptions.

3.25.5 R Implementation

```r
# Population of patients divided by three risk strata with different event rates
strata_prop  ← c(low = 0.60, medium = 0.30, high = 0.10)
event_rates  ← c(low = 0.02, medium = 0.08, high = 0.25)

p_event ← sum(event_rates * strata_prop)
p_event

# Simulate to verify
set.seed(2026)
n ← 1e5
stratum ← sample(names(strata_prop), n, replace = TRUE, prob = strata_prop)
event   ← rbinom(n, 1, event_rates[stratum])
mean(event)

# Bayes using LTP for the denominator
p_low_given_event ← event_rates["low"] * strata_prop["low"] / p_event
p_low_given_event
```

3.25.6 Output & Results

```
[1] 0.061         # P(event) = 0.02*0.6 + 0.08*0.3 + 0.25*0.1 = 0.061
[1] 0.0608        # empirical verification

      low
   0.1967         # P(low-risk | event)
```

The marginal event rate of 6.1% masks substantial heterogeneity across strata. Given that an event occurred, the probability that the patient is in the low-risk stratum is 19.7% – higher than some intuitions suggest, because low-risk patients outnumber high-risk by a factor of six.

3.25.7 Interpretation

Most population-level quantities in epidemiology and public health are weighted sums of stratum-specific rates. Reporting the weighted sum without the underlying decomposition hides heterogeneity; reporting the decomposition clarifies what drives the pooled rate.

3.25.8 Practical Tips

- Verify the partition: probabilities of the B_i must sum to 1; the events must not overlap.
- Conditioning on the "right" variable simplifies otherwise intractable probabilities – choose a latent with clean conditional structure.
- The LTP generalises to nested partitions and to countably infinite ones.
- In Bayesian computation, the LTP is the denominator of Bayes' theorem ($P(B)$ as the marginal likelihood) and is often the hard quantity to compute.
- For continuous marginalisations, numerical integration (`integrate()`) or Monte Carlo (`mean(f(rsample(...)))`) replace the sum.

3.25.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.25.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.26 The Log-Normal Distribution

3.26.1 Introduction

If $\log X$ is normally distributed, then X is log-normal. The distribution is strictly positive and right-skewed, and it models variables that result from multiplicative rather than additive processes – concentrations, income, survival times, and many biomedical lab values.

3.26.2 Prerequisites

Normal distribution, basic properties of logarithms.

3.26.3 Theory

A random variable $X > 0$ is log-normal with parameters μ and σ (on the log scale) if $\log X \sim \mathcal{N}(\mu, \sigma^2)$.

PDF:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{ -\frac{(\log x - \mu)^2}{2\sigma^2} \right\}, \qquad x > 0.$$

Moments:

- $E[X] = \exp(\mu + \sigma^2/2)$.
- $\mathrm{Var}(X) = \left[\exp(\sigma^2) - 1\right] \exp(2\mu + \sigma^2)$.
- Median $= e^{\mu}$; mode $= e^{\mu - \sigma^2}$.

The **geometric mean** of a log-normal sample equals e^{μ}, the median; it is the natural point estimate when the data combine multiplicatively.

Multiplicative CLT: the product of a large number of independent positive random variables tends to log-normal (by applying the CLT to the sum of logs).

3.26.4 Assumptions

Strictly positive continuous data whose logarithm is approximately normal.

3.26.5 R Implementation

```r
mu    ← 1.0
sigma ← 0.5

X ← rlnorm(1e5, meanlog = mu, sdlog = sigma)
c(mean_X          = mean(X),
  theoretical     = exp(mu + sigma^2 / 2),
  median_X        = median(X),
  theoretical_med = exp(mu))

# Log-transform makes it normal
hist(log(X), breaks = 50, freq = FALSE, col = "#2A9D8F")
curve(dnorm(x, mu, sigma), add = TRUE, col = "#F4A261", lwd = 2)

# Geometric mean equals median (for lognormal)
gm ← exp(mean(log(X)))
c(geometric_mean = gm, median = median(X), theoretical = exp(mu))

# MLE
est_mu    ← mean(log(X))
est_sigma ← sd(log(X))
c(mu_hat = est_mu, sigma_hat = est_sigma)
```

3.26.6 Output & Results

```
       mean_X   theoretical      median_X  theoretical_med
        3.082         3.080         2.716            2.718

 geometric_mean    median   theoretical
          2.714     2.716         2.718

   mu_hat    sigma_hat
    0.9995       0.500
```

Empirical arithmetic mean is larger than the median (right skew), the geometric mean equals the median (a property of log-normals), and the log-scale MLEs recover the true μ and σ.

3.26.7 Interpretation

For log-normal variables (concentrations, titres), report the geometric mean and a geometric-mean-based confidence interval rather than the arithmetic mean and SD. In a manuscript: “geometric mean antibody titre 588 (95 % CI 441-782)”.

3.26.8 Practical Tips

- Plot on a log scale when visualising log-normal data; the skew is removed.
- For regression on log-normal outcomes, log-transform the response and fit a linear model on the log scale.
- Back-transformation of the mean of the log-transformed variable gives the geometric mean, not the arithmetic mean.
- For inference on the arithmetic mean, use $\hat{\mu} + \hat{\sigma}^2/2$ on the log scale and back-transform, but this requires a bias correction in small samples.
- Zero or negative values cannot be log-transformed; substitute or switch to a zero-inflated or Tobit model.

3.26.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.26.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.27 Marginal Distributions

3.27.1 Introduction

Given a joint distribution of several random variables, the marginal distribution of any one of them is obtained by integrating (or summing) out the others. Marginalisation is the probability-theoretic version of "don't care": collapsing multi-dimensional information into a single dimension.

3.27.2 Prerequisites

Joint distributions, integration/summation over probability densities.

3.27.3 Theory

For a joint density $f_{X,Y}(x, y)$:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy.$$

Analogously, for a joint PMF: $p_X(x) = \sum_y p_{X,Y}(x, y)$.

The joint CDF gives marginals by pushing the other argument to infinity: $F_X(x) = F_{X,Y}(x, \infty)$.

Key property: the marginals do not determine the joint. Many joint distributions share the same marginals but differ in dependence structure (verified by Sklar's theorem: marginals + copula = joint).

For a multivariate normal, the marginal of each component is univariate normal with the corresponding mean and variance; more generally, any subset of a multivariate normal is multivariate normal.

3.27.4 Assumptions

Integrability of the joint density; otherwise standard probability-space assumptions.

3.27.5 R Implementation

```r
library(mvtnorm)
set.seed(2026)

# Bivariate normal with correlation 0.6
Sigma ← matrix(c(1, 0.6, 0.6, 1), 2, 2)
XY ← rmvnorm(5000, sigma = Sigma)

# Marginal distributions of X and Y
c(mean_X = mean(XY[, 1]), sd_X = sd(XY[, 1]),
  mean_Y = mean(XY[, 2]), sd_Y = sd(XY[, 2]))

# Each marginal is standard normal regardless of correlation
ks.test(XY[, 1], "pnorm")
ks.test(XY[, 2], "pnorm")

# Demonstrate: same marginals, different joints
# Joint 1: correlation 0.6 (bivariate normal)
# Joint 2: same marginals but comonotonic (perfect rank correlation)
U ← pnorm(XY[, 1])   # uniform marginal via probability integral transform
X1 ← qnorm(U); Y1 ← qnorm(U)   # perfectly dependent copy
cov(XY)
cov(cbind(X1, Y1))
```

3.27.6 Output & Results

```
   mean_X      sd_X    mean_Y      sd_Y
   0.0099    1.0050   -0.0027    1.0018

        One-sample Kolmogorov-Smirnov test
D = 0.012, p-value = 0.4672          # X marginal matches N(0,1)

        One-sample Kolmogorov-Smirnov test
D = 0.015, p-value = 0.2132          # Y marginal matches N(0,1)

          [,1]      [,2]
[1,]     1.00     0.61
[2,]     0.61     1.00

          X1       Y1
X1       1.00     1.00       # comonotonic: correlation 1 with same marginals
Y1       1.00     1.00
```

Same marginals (both standard normal), very different joint distributions: correlation 0.6 versus perfect dependence.

3.27.7 Interpretation

In reporting, marginal summaries (marginal mean, SD) are usually what Table 1 presents. When dependence matters – pair-of-eyes measurements, twin pairs, repeated measures – a marginal analysis is incomplete and must be complemented with joint or conditional reporting.

3.27.8 Practical Tips

- Always report marginals alongside any joint analysis; they provide basic sanity checks.
- Marginalisation is trivial for multivariate normals: take the relevant subvector of the mean and the submatrix of the covariance.
- For continuous joint densities without closed form, marginalise numerically via `integrate()` or Monte Carlo.
- Empirical marginals are obtained by projection: for a sample $\{(X_i, Y_i)\}$, the marginal sample of X is $\{X_i\}$.
- In latent-variable models, marginalising over the latent gives the observed-data likelihood – the central quantity in EM and MCMC.

3.27.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.27.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.28 The Multinomial Distribution

3.28.1 Introduction

The multinomial distribution generalises the binomial from two categories to k: the counts of outcomes falling into each of several categories across n independent trials. It underlies the analysis of categorical variables with more than two levels – disease stages, treatment responses, species counts – and sits at the heart of chi-squared goodness-of-fit tests.

3.28.2 Prerequisites

Binomial distribution.

3.28.3 Theory

Let $(X_1, \ldots, X_k)$ count outcomes across k categories in n independent trials, where the probability of outcome j on any trial is p_j, and $\sum p_j = 1$.

PMF:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} \prod_{j=1}^{k} p_j^{x_j},$$

for $x_j \geq 0$ and $\sum x_j = n$.

Moments:

- $E[X_j] = np_j$.
- $\mathrm{Var}(X_j) = np_j(1 - p_j)$.
- $\mathrm{Cov}(X_j, X_\ell) = -np_jp_\ell$ (negative because one count gain forces another's loss).

Marginals: $X_j \sim \mathrm{Binomial}(n, p_j)$. Pairwise joint for (X_j, X_ℓ): a trinomial with $(p_j, p_\ell, 1 - p_j - p_\ell)$.

MLE of $(p_1, \ldots, p_k)$ is the vector of observed proportions $\hat{p}_j = x_j/n$.

3.28.4 Assumptions

- Fixed total n trials.
- Independent trials.
- Constant category probabilities across trials.

3.28.5 R Implementation

```r
set.seed(2026)

p ← c(0.40, 0.30, 0.20, 0.10)
n ← 200

# Single realisation
one_draw ← rmultinom(1, size = n, prob = p)
data.frame(category = 1:length(p), count = as.integer(one_draw),
           expected = n * p)

# Simulate many replicates, check moments
many ← rmultinom(1e4, n, p)
rowMeans(many); apply(many, 1, var)

# Expected variance
n * p * (1 - p)

# Chi-squared goodness-of-fit
obs ← c(82, 66, 34, 18)
chisq.test(obs, p = p)

# Multinomial likelihood of observed counts
dmultinom(obs, size = sum(obs), prob = p)
```

3.28.6 Output & Results

```
  category count expected
1        1    77       80
2        2    66       60
3        3    38       40
4        4    19       20

rowMeans:  80.07  60.04  39.95  19.94
apply var: 48.2   42.0   32.1   17.8

Expected var:
[1] 48.00 42.00 32.00 18.00

        Chi-squared test for given probabilities

data:  obs
X-squared = 3.2, df = 3, p-value = 0.363
```

The empirical means and variances match the theoretical values. The chi-squared test does not reject the specified category probabilities.

3.28.7 Interpretation

Reporting multinomial fits usually focuses on category proportions with uncertainty. For large samples, the chi-squared goodness-of-fit test is the standard tool for comparing observed multinomial counts to hypothesised probabilities.

3.28.8 Practical Tips

- For small expected counts (<5), use Fisher's exact test or Monte Carlo chi-squared (`chisq.test(..., simulate.p.value = TRUE)`).
- Multinomial categories are mutually exclusive by definition; overlapping classifications require a different framework (multilabel models).
- Conditional on the total, the multinomial is a sufficient statistic for the unordered category counts.
- Bayesian inference for multinomial probabilities uses the Dirichlet prior (conjugate); posterior is Dirichlet with shape parameters updated by observed counts.
- For regression with multinomial outcomes, use multinomial logistic regression (`nnet::multinom`).

3.28.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.28.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.29 The Multivariate Normal Distribution

3.29.1 Introduction

The multivariate normal (MVN) is the joint distribution that plays the same role for random vectors that the normal plays for random scalars. It is parameterised by a mean vector and a covariance matrix, and it underlies linear discriminant analysis, principal component analysis, Gaussian graphical models, and most of multivariate inference.

3.29.2 Prerequisites

Normal distribution, linear algebra (vectors, matrices, positive-definite matrices).

3.29.3 Theory

A random vector $\mathbf{X} \in \mathbb{R}^p$ is MVN with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and positive definite covariance $\Sigma \in \mathbb{R}^{p \times p}$ if its density is

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

The exponent is the squared **Mahalanobis distance** from $\mathbf{x}$ to $\boldsymbol{\mu}$.

Key properties:

- **Marginals are MVN**: every subset of components is MVN.
- **Conditionals are MVN**: given a subset of components, the rest is MVN with mean and covariance depending on the conditioning values.
- **Linear transformations**: if $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$ and A is any matrix, then $A\mathbf{X} \sim \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^\top)$.
- **Zero correlation implies independence** in the MVN family (unlike general distributions).
- **Mahalanobis distance squared** $\sim \chi^2_p$ for $\mathbf{X} \sim \mathcal{N}_p(\boldsymbol{\mu}, \Sigma)$.

3.29.4 Assumptions

Σ must be symmetric positive-definite. Degenerate cases (singular Σ) give a distribution supported on a lower-dimensional affine subspace.

3.29.5 R Implementation

```r
library(mvtnorm); library(ggplot2)

set.seed(2026)
mu    <- c(0, 1)
Sigma <- matrix(c(1, 0.6, 0.6, 2), 2, 2)

X <- rmvnorm(2000, mean = mu, sigma = Sigma)
colMeans(X)
cov(X)

df <- data.frame(X1 = X[, 1], X2 = X[, 2])
ggplot(df, aes(X1, X2)) +
  geom_point(alpha = 0.3, colour = "#2A9D8F") +
  stat_ellipse(level = 0.95, colour = "#F4A261", linewidth = 1) +
  coord_equal() +
  theme_minimal()

# Conditional distribution: X1 | X2 = x2
# mu_1 + Sigma_12 Sigma_22^{-1} (x2 - mu_2)
x2_val <- 2.0
cond_mean <- mu[1] + Sigma[1, 2] / Sigma[2, 2] * (x2_val - mu[2])
cond_var  <- Sigma[1, 1] - Sigma[1, 2]^2 / Sigma[2, 2]
c(mean = cond_mean, var = cond_var)

# Mahalanobis distance
md2 <- mahalanobis(X, center = mu, cov = Sigma)
c(mean_md2 = mean(md2), expected = 2)   # chi^2_2 mean = 2
```

3.29.6 Output & Results

```
 [1]  0.003 1.012

         [,1]      [,2]
[1,]    1.002    0.605
[2,]    0.605    1.989

    mean      var
  0.3000   0.8200

mean_md2 expected
    2.00     2.00
```

Empirical mean and covariance closely match the theoretical values, and the Mahalanobis distance squared has the expected chi-squared mean.

3.29.7 Interpretation

MVN assumptions underlie multiple regression (normal errors), linear discriminant analysis (normal class densities), and most factor-analytic and SEM approaches. Violations of joint normality invalidate ellipsoidal confidence regions and some likelihood-based tests.

3.29.8 Practical Tips

- Check joint normality with chi-squared Q-Q plots of Mahalanobis distance squared, not just marginal Q-Q plots.
- Near-singular Σ (high collinearity) causes numerical instability; regularise with `Sigma + k I` or use generalised inverses.
- For high-dimensional MVN (p large), shrinkage estimators (Ledoit-Wolf) stabilise the covariance.
- Sampling from MVN: `mvtnorm::rmvnorm`, or directly via Cholesky: $\mathbf{X} = \boldsymbol{\mu} + L\mathbf{z}$ with $LL^\top = \Sigma$ and $\mathbf{z} \sim \mathcal{N}_p(0, I)$.
- The Kullback-Leibler divergence between two MVNs has a closed form; useful in information-geometric methods.

3.29.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.29.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson

- Continuous distributions and Q-Q plots

3.30 The Negative Binomial Distribution

3.30.1 Introduction

The negative binomial distribution is the workhorse model for overdispersed counts – counts whose variance exceeds their mean. It arises naturally as the number of failures before the r-th success in a sequence of Bernoulli trials, or as a gamma mixture of Poisson distributions. It dominates modern count-regression practice (RNA-seq, public-health incidence, insurance claims).

3.30.2 Prerequisites

Geometric and Poisson distributions; gamma distribution is useful for the mixture interpretation.

3.30.3 Theory

Two standard parameterisations:

Counts-of-failures form: Y is the number of failures before the r-th success in iid Bernoulli(p) trials:

$$P(Y = k) = \binom{k + r - 1}{k} p^r (1 - p)^k, \qquad k = 0, 1, 2, \ldots$$

Mean-dispersion form: parameterise by mean μ and dispersion (size) parameter θ with

$$P(Y = k) = \frac{\Gamma(k + \theta)}{k!\,\Gamma(\theta)} \left(\frac{\theta}{\theta + \mu}\right)^{\theta} \left(\frac{\mu}{\theta + \mu}\right)^{k}.$$

Moments in the mean-dispersion form:

- $E[Y] = \mu$.
- $\mathrm{Var}(Y) = \mu + \mu^2/\theta$.

Gamma-Poisson interpretation: if $\Lambda \sim \mathrm{Gamma}(\theta, \theta/\mu)$ and $Y \mid \Lambda \sim \mathrm{Poisson}(\Lambda)$, then marginally $Y \sim \mathrm{NegBin}(\mu, \theta)$. As $\theta \to \infty$, the NB collapses to Poisson(μ).

Geometric as a special case: $r = 1$ (or $\theta = 1$) recovers the geometric distribution.

3.30.4 Assumptions

No common-rate assumption (unlike Poisson); the heterogeneity in underlying rates is parameterised by θ.

3.30.5 R Implementation

```r
library(MASS)

# R's rnbinom: parameterise by size (= theta) and mu
mu <- 5; theta <- 2
Y <- rnbinom(1e5, size = theta, mu = mu)

c(mean     = mean(Y),
  variance = var(Y),
  expected_var = mu + mu^2 / theta)

# PMF comparison to Poisson
k <- 0:20
pmf_nb   <- dnbinom(k, size = theta, mu = mu)
pmf_pois <- dpois(k, lambda = mu)

data.frame(k, nb = round(pmf_nb, 4), poisson = round(pmf_pois, 4))

# Fit by MLE using MASS::glm.nb
df <- data.frame(y = Y, x = rnorm(1e5))
fit <- glm.nb(y ~ x, data = df)
summary(fit)$theta
```

3.30.6 Output & Results

```
     mean  variance expected_var
    5.020    17.48        17.50

    k      nb  poisson
1   0  0.2041   0.0067
2   1  0.2041   0.0337
3   2  0.1701   0.0842
4   3  0.1296   0.1404
5   4  0.0925   0.1755
6   5  0.0634   0.1755
...

[1] 1.998        # estimated theta
```

Empirical variance (17.5) matches the theoretical $\mu + \mu^2/\theta = 5 + 25/2 = 17.5$. The MLE of θ is close to the true value 2.

3.30.7 Interpretation

Negative binomial is the default count-data model in modern applied work. In a manuscript: "incidence rates were compared using negative binomial regression to account for overdispersion (dispersion estimated as $\theta = 2.0$)". Always report the dispersion parameter alongside the rate ratios.

3.30.8 Practical Tips

- Detect overdispersion by fitting a Poisson first and checking residual deviance $\approx$ df; ratios > 1.5 suggest overdispersion.

- `glm.nb` jointly estimates regression coefficients and dispersion; faster-but-less-flexible alternatives include `glmmTMB` and `brms`.
- Zero-inflated variants (`pscl::zeroinfl` with `dist = "negbin"`) handle excess zeros on top of overdispersion.
- Reparameterisations: some textbooks use (n, p), some (μ, θ); R's `rnbinom(n, size, prob)` or `rnbinom(n, size, mu)` accepts either.
- In RNA-seq, DESeq2 and edgeR both model counts as negative binomial with shrinkage on the dispersion parameter.

3.30.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.30.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.31 The Normal Distribution

3.31.1 Introduction

The normal distribution – also called the Gaussian distribution – is the most important continuous distribution in applied statistics. It arises naturally as the limit of averages (via the central limit theorem), as the distribution of measurement errors, and as the default assumption in classical regression and ANOVA. Even when a raw variable is not itself normal, its transformations, its averages, or its residuals often are.

3.31.2 Prerequisites

Familiarity with probability density and cumulative distribution functions, and with the idea of expectation and variance, is assumed. The reader should know how to call R functions and generate a histogram.

3.31.3 Theory

The normal distribution with mean μ and variance σ^2 has the probability density function

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

defined for all real x. The mean, median, and mode all coincide at μ; the standard deviation σ controls the spread. The distribution is symmetric about μ, with thin tails – about 68% of the mass falls within one standard deviation of the mean, 95% within two, and 99.7% within three. These "68-95-99.7" rule-of-thumb numbers are used constantly in applied work.

Several properties make the normal distribution uniquely convenient:

- **Closure under linear transformations**: if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
- **Closure under sums**: the sum of independent normals is normal.
- **Maximum entropy**: among all distributions with a given mean and variance, the normal has the highest entropy – it is the least informative choice given those moments.
- **Conjugacy**: in Bayesian analysis with known variance, the normal prior is conjugate to the normal likelihood.

The standard normal is $\mathcal{N}(0, 1)$ and its CDF is denoted $\Phi(z)$. Any normal can be reduced to the standard normal by the linear transformation $Z = (X - \mu)/\sigma$.

3.31.4 Assumptions

When a procedure *assumes normality* it usually means one of two things: (i) the individual observations are normally distributed, or (ii) the residuals from the fitted model are normally distributed. The latter is much more common in practice, and is what matters for linear regression and ANOVA. For inference on the mean, the CLT often makes this distinction academic: the sampling distribution of the mean can be approximately normal even when the observations are not.

3.31.5 R Implementation

R provides the standard `d`, `p`, `q`, `r` interface for every distribution:

```
library(ggplot2)

dnorm(0)
pnorm(1.96)
qnorm(0.975)
rnorm(5, mean = 10, sd = 2)

grid <- data.frame(x = seq(-4, 4, length.out = 401))
grid$density <- dnorm(grid$x)

ggplot(grid, aes(x = x, y = density)) +
  geom_line(colour = "steelblue", linewidth = 1) +
  geom_area(data = subset(grid, x >= -1.96 & x <= 1.96),
            aes(y = density), fill = "steelblue", alpha = 0.25) +
  labs(x = "z", y = "Density",
       title = "Standard normal with the central 95% shaded") +
  theme_minimal(base_size = 12)
```

`dnorm(0)` returns the density at the mean, $1/\sqrt{2\pi} \approx 0.3989$. `pnorm(1.96)` returns the CDF at 1.96, which is approximately 0.975. `qnorm(0.975)` returns the 97.5% quantile of the standard normal, which is approximately 1.96.

3.31.6 Output & Results

The plot shows the familiar bell curve with the central 95% shaded between $z = \pm 1.96$. The `rnorm()` call produces a sample of five values, for example `c(9.02, 11.14, 7.88, 12.47, 10.06)`, from a normal with mean 10 and SD 2.

3.31.7 Interpretation

In a manuscript, reporting a variable as "approximately normal, mean 75.2 (SD 8.4)" communicates that the distribution is roughly symmetric and bell-shaped, and that 95% of values lie in the range $75.2 \pm 1.96 \times 8.4 = (58.7, 91.7)$. If the distribution is skewed or heavy-tailed, reporting a mean and SD is misleading, and the summary should be a median with IQR.

3.31.8 Practical Tips

- Check normality with a Q-Q plot, not with a formal test. In small samples, formal tests lack power; in large samples, they reject for trivial deviations that do not affect inference.
- Data that are strictly positive and right-skewed (concentrations, incubation times, expression levels) are often approximately *log-normal*: their logarithm is normal. Analyses on the log scale may be more appropriate.
- In high dimensions (many variables per observation), joint normality is a much stronger assumption than marginal normality. Check both.
- "Normal" does not mean "common". Many natural phenomena are not normally distributed; always verify rather than assume.

3.31.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.31.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.32 The Poisson Distribution

3.32.1 Introduction

The Poisson distribution models counts of rare, independent events in a fixed window of time or space. Calls per hour, mutations per kilobase, photons per pixel, hospital admissions per day – all naturally Poisson when the underlying rate is approximately constant and events do not cluster.

3.32.2 Prerequisites

Discrete distributions and the notion of a rate.

3.32.3 Theory

A random variable X is **Poisson**(λ) if

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

Moments:

- $E[X] = \lambda$.
- $\mathrm{Var}(X) = \lambda$ (equi-dispersion).
- Higher central moments: skewness $1/\sqrt{\lambda}$, excess kurtosis $1/\lambda$; the distribution becomes approximately normal for large λ.

Poisson limit theorem: $\mathrm{Binomial}(n, \lambda/n) \xrightarrow{d} \mathrm{Poisson}(\lambda)$ as $n \to \infty$. The Poisson is the "law of small numbers" limit of rare-event binomial trials.

Closure under sums: independent $X \sim \mathrm{Poisson}(\lambda_1)$ and $Y \sim \mathrm{Poisson}(\lambda_2)$ yield $X + Y \sim \mathrm{Poisson}(\lambda_1 + \lambda_2)$.

Connection to exponential: inter-arrival times in a Poisson process are iid exponential with rate λ.

3.32.4 Assumptions

The underlying events must be (approximately) independent with a constant rate. Violations of equi-dispersion (variance exceeding mean) signal **overdispersion**, commonly addressed by the negative binomial.

3.32.5 R Implementation

```r
library(ggplot2)

lambda ← 3.5

k ← 0:15
pmf ← dpois(k, lambda)

data.frame(k, pmf = round(pmf, 4))

# Moments
X ← rpois(1e5, lambda)
c(mean = mean(X), var = var(X), skew = mean((X - mean(X))^3) / sd(X)^3)

# Plot PMF
ggplot(data.frame(k, pmf), aes(k, pmf)) +
  geom_col(fill = "#2A9D8F") +
  scale_x_continuous(breaks = k) +
  labs(x = "k", y = "P(X = k)",
       title = paste("Poisson(", lambda, ")")) +
  theme_minimal()

# Verify Poisson limit from binomial
n_vals ← c(10, 100, 1000, 10000)
results ← sapply(n_vals, function(n) {
  p ← lambda / n
  X ← rbinom(1e5, n, p)
  c(n = n, mean = mean(X), var = var(X))
})
t(results)
```

3.32.6 Output & Results

```
    k    pmf
1   0 0.0302
2   1 0.1057
3   2 0.1850
4   3 0.2158
5   4 0.1888
6   5 0.1322
...

    mean     var    skew
    3.503   3.502   0.536

          n    mean     var
[1,]    10    3.49    3.13
[2,]   100    3.50    3.40
[3,]  1000    3.50    3.49
[4,] 10000    3.50    3.50
```

Mean and variance are both ≈ 3.5 in the empirical sample; the Poisson limit from binomial tightens as $n \to \infty$.

3.32.7 Interpretation

Poisson regression (`glm(..., family = poisson)`) estimates log-rates from count data; reports of incidence-rate ratios and event-per-person-time summaries are typically Poisson-based. Always check for overdispersion before accepting a Poisson fit at face value.

3.32.8 Practical Tips

- Equi-dispersion is the signature assumption; real count data often violate it.
- Use quasi-Poisson (`family = quasipoisson`) or negative binomial (`MASS::glm.nb`) when variance exceeds mean.
- Zero-inflated Poisson (`pscl::zeroinfl`) handles excess zeros.
- For rates per unit exposure (e.g., per person-year), include an offset in the regression: `glm(y ~ x, offset = log(exposure), family = poisson)`.
- The normal approximation works for $\lambda \geq 10$; for smaller λ, use the exact PMF.

3.32.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.32.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.33 The Probability Density Function

3.33.1 Introduction

For a continuous random variable, probability is distributed smoothly across the support rather than concentrated on countable points. The **probability density function (PDF)** encodes this distribution: integrating the PDF over a set gives the probability that the variable lands in that set.

3.33.2 Prerequisites

Calculus (integration) and the concept of a continuous random variable.

3.33.3 Theory

A continuous random variable X has a PDF f_X if, for every measurable set A,

$$P(X \in A) = \int_A f_X(x)\,dx.$$

A valid PDF satisfies:

1. $f_X(x) \geq 0$ for all x.
2. $\int_{-\infty}^{\infty} f_X(x)\,dx = 1$.

Crucially, $f_X(x)$ is **not** a probability. For continuous X, $P(X = x) = 0$ for every x. The density can exceed 1 at a given point, provided the integral is bounded. Density has units of inverse x: for x in seconds, $f_X(x)$ has units of $1/\text{second}$.

Useful relationships:

- CDF: $F_X(x) = \int_{-\infty}^{x} f_X(u)\,du$.
- PDF from CDF: $f_X(x) = F_X'(x)$ where F_X is differentiable.
- Expected value: $E[X] = \int x f_X(x)\,dx$.
- Variance: $\text{Var}(X) = \int (x - E[X])^2 f_X(x)\,dx$.

In R, each standard continuous distribution has a `d*()` function returning its PDF.

3.33.4 Assumptions

The variable is continuous (absolutely continuous with respect to Lebesgue measure). Mixed or discrete variables do not have ordinary PDFs (though the mixed case has a generalised density with respect to a mixed measure).

3.33.5 R Implementation

```r
library(ggplot2)

x_grid ← seq(-4, 4, length.out = 801)
fx ← dnorm(x_grid, mean = 0, sd = 1)

ggplot(data.frame(x = x_grid, f = fx), aes(x, f)) +
  geom_line(colour = "#2A9D8F", linewidth = 1) +
  geom_area(data = subset(data.frame(x = x_grid, f = fx),
                          x ⩾ -1.96 & x ⩽ 1.96),
            aes(y = f), fill = "#2A9D8F", alpha = 0.25) +
  labs(x = "x", y = "f(x)",
       title = "Standard normal PDF with central 95% shaded") +
  theme_minimal()

# Verify normalisation by numerical integration
integrate(dnorm, -Inf, Inf)

# Probability of an interval via integration
integrate(dnorm, -1.96, 1.96)        # about 0.95

# Equivalently via the CDF
pnorm(1.96) - pnorm(-1.96)

# Moments via integration
integrate(function(x) x * dnorm(x), -Inf, Inf)        # E[X]
integrate(function(x) x^2 * dnorm(x), -Inf, Inf)      # E[X^2]
```

3.33.6 Output & Results

```
1 with absolute error < 9.4e-05

0.9500042 with absolute error < 1e-11

[1] 0.9500042

-4.4e-11 with absolute error < ...   # E[X] ≈ 0
1 with absolute error < ...          # E[X^2] = Var(X) for standard normal
```

Integrating the standard normal PDF over $[-1.96, 1.96]$ gives 0.95 – the well-known central probability – and the `pnorm` difference recovers the same value via the CDF.

3.33.7 Interpretation

For applied purposes, a PDF is visualised by its curve; probabilities are computed via `p*()` functions in R rather than by explicit integration. When reporting a continuous distribution in a paper, the density plot, mean, SD, and median are the standard summaries.

3.33.8 Practical Tips

- Density at a point is not probability; do not compare $f(x_1)$ to $f(x_2)$ as probabilities of outcomes 1 and 2. Compare $P(X \in (x_1 - h, x_1 + h))$ for small h instead.
- The density can be unbounded (e.g., Gamma with shape < 1 near 0); integrability, not boundedness, is what matters.

- Kernel density estimates (`density()`) approximate the PDF from a sample.
- Log-densities (`dnorm(..., log = TRUE)`) avoid underflow when evaluating likelihoods with many observations.
- The relationship $f = F'$ is the starting point for the inverse-CDF method: to generate X, apply F^{-1} to a uniform random variable.

3.33.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.33.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.34 The Probability Mass Function

3.34.1 Introduction

For a discrete random variable, the probability mass function (PMF) gives the probability that the variable takes each specific value. It is the discrete analogue of the density function, and the probability of any event reduces to summing the PMF over the relevant values.

3.34.2 Prerequisites

Discrete random variables and basic probability.

3.34.3 Theory

For a discrete random variable X with countable support $\{x_1, x_2, \ldots\}$, the PMF is

$$p_X(x) = P(X = x), \qquad x \in \mathbb{R}.$$

A valid PMF satisfies:

1. $p_X(x) \geq 0$ for all x.
2. $\sum_x p_X(x) = 1$ (total probability is 1).
3. $p_X(x) = 0$ for x outside the support.

The probability of an event is a sum over the PMF:

$$P(X \in A) = \sum_{x \in A} p_X(x).$$

The CDF is the running cumulative sum:

$$F_X(x) = P(X \le x) = \sum_{x_i \le x} p_X(x_i).$$

Expected value and variance follow the usual formulas:

$$E[X] = \sum_x x\, p_X(x), \qquad \mathrm{Var}(X) = \sum_x (x - E[X])^2\, p_X(x).$$

Common discrete PMFs in R use the `d<name>()` family: `dbinom`, `dpois`, `dgeom`, `dhyper`, `dnbinom`, `dmultinom`.

3.34.4 Assumptions

The variable must be discrete. Assigning a PMF to a continuous variable produces a degenerate object (zero at every point, failing normalisation).

3.34.5 R Implementation

```r
library(ggplot2)

# Binomial PMF for n = 10 trials, p = 0.3
n ← 10; p ← 0.3
k ← 0:n
pmf ← dbinom(k, size = n, prob = p)

data.frame(k, pmf)

# Verify normalisation
sum(pmf)

# Compute probabilities of events via sums
sum(pmf[k ⩾ 3 & k ⩽ 5])        # P(3 ⩽ X ⩽ 5)
sum(pmf[k %in% c(0, n)])        # P(X = 0 or X = n)

# Expected value and variance
sum(k * pmf)
sum((k - sum(k * pmf))^2 * pmf)

ggplot(data.frame(k, pmf), aes(k, pmf)) +
  geom_col(fill = "#2A9D8F") +
  scale_x_continuous(breaks = k) +
  labs(x = "k", y = "P(X = k)",
       title = "Binomial(10, 0.3) PMF") +
  theme_minimal()
```

3.34.6 Output & Results

```
   k         pmf
1  0 0.028247525
2  1 0.121060821
3  2 0.233474440
4  3 0.266827932
5  4 0.200120949
6  5 0.102919345
7  6 0.036756909
8  7 0.009001692
9  8 0.001446773
10 9 0.000137781
11 10 0.0000059049

[1] 1            # normalisation
[1] 0.5695       # P(3 ⩽ X ⩽ 5)
[1] 0.0283       # P(X = 0 or X = 10)
[1] 3            # E[X] = np
[1] 2.1          # Var(X) = np(1-p)
```

3.34.7 Interpretation

A PMF table or bar chart is the natural display of a discrete distribution. For a manuscript, reporting "the number of responders out of 10 was binomially distributed with $p = 0.3$" conveys the full PMF implicitly.

3.34.8 Practical Tips

- Make sure the PMF sums to exactly 1 when you define a custom distribution; normalise explicitly.
- For mixtures of Poissons, Bernoullis, and similar, sums of PMFs computed componentwise give the mixture PMF.
- Use `choose(n, k) * p^k * (1 - p)^(n - k)` for symbolic verification of the binomial PMF.
- Numerical underflow can corrupt PMFs at extreme values; work on the log scale (`dbinom(k, ..., log = TRUE)`) when n is large.
- Histograms of samples approximate the PMF in finite samples; for a true PMF-bar-chart, use `geom_col()` at the exact support values.

3.34.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.34.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem

- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.35 Random Variables

3.35.1 Introduction

A **random variable** is a function that assigns a real number to each outcome of a random experiment. The concept packages probabilistic thinking into the language of numbers – means, variances, correlations – and is the abstraction on which every statistical method rests.

3.35.2 Prerequisites

Sample spaces, events, and the probability measure.

3.35.3 Theory

Given a probability space $(\Omega, \mathcal{F}, P)$, a **random variable** is a measurable function

$$X : \Omega \to \mathbb{R}$$

such that $\{X \leq x\} = \{\omega : X(\omega) \leq x\} \in \mathcal{F}$ for every $x \in \mathbb{R}$. Measurability ensures that we can assign probabilities to events defined in terms of X.

The **distribution** of X is the probability measure it induces on $\mathbb{R}$:

$$P_X(A) = P(X \in A) = P(\{\omega : X(\omega) \in A\}).$$

In practice, we rarely work with Ω directly; we work with the distribution P_X, summarised by its CDF $F_X(x) = P(X \leq x)$ and (when it exists) its PMF p_X or density f_X.

Random variables are classified as:

- **Discrete**: takes values in a countable set. Described by a PMF.
- **Continuous** (absolutely continuous): has a density. Described by a PDF.
- **Mixed**: combines discrete atoms and a continuous part. Common for censored / thresholded data.

Transformations of random variables give new random variables: if X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is measurable, then $Y = g(X)$ is a random variable with distribution determined by X and g.

3.35.4 Assumptions

A random variable is well-defined provided the underlying probability space exists and the function is measurable. Measurability is automatic for every function that arises in practice (continuous or piecewise-continuous).

3.35.5 R Implementation

In R, we usually work with random variables via their distributions.

```r
set.seed(2026)

# Discrete random variable: binomial
X_discrete <- rbinom(10000, size = 10, prob = 0.3)
head(sort(unique(X_discrete)))
prop.table(table(X_discrete))[1:5]     # empirical PMF

# Continuous random variable: normal
X_continuous <- rnorm(10000, mean = 0, sd = 1)
quantile(X_continuous, c(0.025, 0.5, 0.975))

# Transformation: Y = g(X) = X^2 of a standard normal
Y <- X_continuous^2
summary(Y)
var(Y)     # theoretical variance is 2 (Y ~ chi-squared with 1 df, variance 2)

# Mixed RV: censored at 0 (negative values truncated)
X_mixed <- pmax(rnorm(10000), 0)
mean(X_mixed == 0)
mean(X_mixed)
```

3.35.6 Output & Results

```
X_discrete       0       1       2       3       4
empirical  0.0306  0.1195  0.2314  0.2643  0.2059

         2.5%          50%         97.5%
   -1.935       -0.0089       1.9689

       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
      0.000     0.128     0.455     1.008     1.319    17.59
[1] 2.078                                              # var(X^2)

[1] 0.499                                              # P(X = 0)
[1] 0.399                                              # mean of censored
```

Each quantity (PMF, quantile, variance of a transformation, mixed distribution's atom) is a statistic of the corresponding random variable.

3.35.7 Interpretation

Random variables let us move from "outcome space" thinking to "number space" thinking. A methods section typically declares the random variable implicitly: "Let X be the systolic blood pressure of a randomly selected patient" specifies the random variable without bothering with the underlying Ω.

3.35.8 Practical Tips

- For most applied work, think of random variables as "generators of data" with a specified distribution; the measurability formalism is background.

- A random variable is distinct from its realisation. "X" is the variable; "x_i" is the observed value for unit i.
- Vector-valued random variables (**random vectors**) generalise the scalar case; their distribution is the joint distribution of the components.
- Transformations preserve measurability; common operations (sums, products, functions) are always measurable in practice.
- In R, `d*`, `p*`, `q*`, `r*` families of functions parameterise distributions; think of them as computing quantities of random variables with the specified distribution.

3.35.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.35.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.36 Sample Space and Events

3.36.1 Introduction

Before any probability can be assigned, we need to know what *could* happen. The **sample space** is the set of all possible outcomes of a random experiment; **events** are subsets of the sample space to which we assign probabilities. Clean reasoning about probability begins with clear identification of both.

3.36.2 Prerequisites

Familiarity with basic set theory – unions, intersections, complements – is sufficient.

3.36.3 Theory

The **sample space** Ω is the set of all possible outcomes. Examples:

- Coin toss: $\Omega = \{H, T\}$.
- Two dice: $\Omega = \{(i, j) : i, j \in \{1, \ldots, 6\}\}$, 36 outcomes.
- Waiting time: $\Omega = [0, \infty)$.
- DNA base at position i: $\Omega = \{A, C, G, T\}$.

An **event** is a subset $A \subseteq \Omega$. We say event A occurs if the realised outcome ω is in A. Key operations:

- **Union** $A \cup B$: the event that A or B (or both) occurs.
- **Intersection** $A \cap B$: the event that both A and B occur.
- **Complement** A^c: the event that A does not occur.
- **Difference** $A \setminus B = A \cap B^c$: A but not B.

Two key identities – **De Morgan's laws** – relate complements and set operations:

$$(A \cup B)^c = A^c \cap B^c, \qquad (A \cap B)^c = A^c \cup B^c.$$

Events A and B are **disjoint** (mutually exclusive) if $A \cap B = \emptyset$. They are **independent** (under probability P) if $P(A \cap B) = P(A)P(B)$. Disjointness is about sets; independence is about probabilities – the two concepts should not be confused.

3.36.4 Assumptions

No assumptions beyond the underlying probability space being well-defined.

3.36.5 R Implementation

We can enumerate finite sample spaces explicitly and manipulate events as logical vectors.

```
# Two dice: enumerate the 36-element sample space
omega ← expand.grid(die1 = 1:6, die2 = 1:6)
head(omega, 3)

# Define events as logical vectors indexing rows of omega
A ← with(omega, die1 == 1)            # first die is 1
B ← with(omega, die1 + die2 == 7)     # sum equals 7

# Set operations
A_union_B     ← A | B
A_inter_B     ← A & B
A_complement  ← !A
A_minus_B     ← A & !B

# Under a uniform probability on 36 outcomes:
p ← function(E) mean(E)   # proportion of outcomes in E
c(P_A         = p(A),
  P_B         = p(B),
  P_A_union_B = p(A_union_B),
  P_A_and_B   = p(A_inter_B))

# Verify inclusion-exclusion: P(A U B) = P(A) + P(B) - P(A n B)
c(lhs = p(A_union_B),
  rhs = p(A) + p(B) - p(A_inter_B))

# Verify De Morgan: (A U B)^c = A^c n B^c
all((!(A | B)) == (!A & !B))
```

3.36.6 Output & Results

```
  die1 die2
1    1    1
2    2    1
3    3    1

       P_A        P_B P_A_union_B  P_A_and_B
    0.1667     0.1667      0.3056     0.02778

     lhs      rhs
  0.3056   0.3056

[1] TRUE
```

All identities are satisfied. $P(A \cup B) = 11/36$ because 11 outcomes satisfy "first die is 1 OR sum is 7", and the formula recovers the same value.

3.36.7 Interpretation

When writing a methods section, the sample space is usually defined implicitly by the design ("patients were followed for 30 days and classified as event / no event"). Making this explicit – "the sample space is $\{\text{event}, \text{no event}\}$" – is rarely necessary except in pedagogical contexts.

3.36.8 Practical Tips

- For continuous sample spaces ($\Omega = \mathbb{R}$ or higher-dimensional), not every subset is an event; measurability restrictions apply. In finite or countable sample spaces, every subset is an event.
- Do not conflate *disjoint* with *independent*. Two disjoint events with positive probability are maximally dependent: if one occurs, the other cannot.
- De Morgan's laws generalise to countable unions and intersections; they are indispensable when reasoning about complements of complex events.
- Encoding finite sample spaces in R (via `expand.grid` or combinations) makes exploratory probability calculations explicit and checkable.
- Inclusion-exclusion extends to many events: $P(\bigcup A_i) = \sum P(A_i) - \sum P(A_i \cap A_j) + \ldots$.

3.36.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.36.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.37 The Survival Function

3.37.1 Introduction

For a time-to-event random variable $T \geq 0$, the **survival function** $S(t) = P(T > t)$ is the probability of not having experienced the event by time t. It is the complement of the CDF and the single most important descriptive quantity in survival analysis.

3.37.2 Prerequisites

Random variables, CDF, basic survival-analysis concepts.

3.37.3 Theory

For a non-negative random variable T,

$$S(t) = P(T > t) = 1 - F(t).$$

Properties:

- $S(0) = 1$, $\lim_{t\to\infty} S(t) = 0$.
- Non-increasing.
- Right-continuous (by convention).

Related quantities:

- PDF: $f(t) = -S'(t)$.
- Hazard: $h(t) = f(t)/S(t)$.
- Cumulative hazard: $H(t) = -\log S(t)$.

Classical forms:

- Exponential: $S(t) = e^{-\lambda t}$.
- Weibull: $S(t) = \exp\big[-(t/\lambda)^k\big]$.
- Log-normal: $S(t) = 1 - \Phi\big((\log t - \mu)/\sigma\big)$.

Kaplan-Meier is the non-parametric MLE of S from right-censored data: a step function that drops at each observed event time by a factor reflecting the fraction of the at-risk set that experiences the event.

3.37.4 Assumptions

Non-negative random variable representing a time to an event. For Kaplan-Meier: independent censoring.

3.37.5 R Implementation

```
library(survival)
set.seed(2026)

# Theoretical survival functions
t_grid ← seq(0, 10, length.out = 400)
S_exp     ← exp(-0.3 * t_grid)
S_weibull ← exp(-(t_grid / 3)^2)
S_lnorm   ← 1 - plnorm(t_grid, meanlog = log(3), sdlog = 0.5)

plot(t_grid, S_exp, type = "l", col = "#6A4C93", lwd = 2,
     main = "Theoretical survival functions", xlab = "t", ylab = "S(t)")
lines(t_grid, S_weibull, col = "#2A9D8F", lwd = 2)
lines(t_grid, S_lnorm,   col = "#F4A261", lwd = 2)
legend("topright", c("exponential", "Weibull k=2", "log-normal"),
       col = c("#6A4C93", "#2A9D8F", "#F4A261"), lwd = 2)

# Empirical Kaplan-Meier from censored data
n ← 80
T_true ← rweibull(n, shape = 2, scale = 3)
C      ← runif(n, 0, 10)
T_obs  ← pmin(T_true, C)
delta  ← as.integer(T_true ⩽ C)

km ← survfit(Surv(T_obs, delta) ~ 1)
plot(km, conf.int = TRUE, main = "Kaplan-Meier estimate",
     xlab = "t", ylab = "S(t)", col = "#2A9D8F")
lines(t_grid, exp(-(t_grid / 3)^2), col = "#F4A261", lwd = 2)

# Median survival
summary(km)$table["median"]
qweibull(0.5, shape = 2, scale = 3)
```

3.37.6 Output & Results

```
   median
    2.499

[1] 2.498
```

Empirical and theoretical median survival times agree. The Kaplan-Meier step function closely tracks the true Weibull survival curve.

3.37.7 Interpretation

Reporting a median survival with its 95 % CI from a Kaplan-Meier estimate is the standard headline statistic of survival studies. Survival functions are typically compared across groups via log-rank tests and visualised together on the same axes.

3.37.8 Practical Tips

- Report medians with CIs when survival drops below 0.5; report specific-time survival (1-year, 5-year) when the study ends before the median is reached.
- Kaplan-Meier is distribution-free; parametric survival models (Weibull, log-normal) give smooth curves but commit to a shape.
- Include a risk table below Kaplan-Meier plots (`survminer::ggsurvplot(..., risk.table = TRUE)`).
- Under heavy censoring, the tail of the survival curve is poorly estimated; avoid over-interpreting late-follow-up differences.
- Survival functions never take values above 1 or below 0; if your empirical estimator returns such values, there is a bug.

3.37.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.37.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.38 Student's t-Distribution

3.38.1 Introduction

The t-distribution (also called Student's t, after W. S. Gosset's pseudonym) arises as the ratio of a standard normal to the scaled square root of an independent chi-squared. It is the reference distribution for t-tests and for small-sample confidence intervals for a normal mean with unknown variance.

3.38.2 Prerequisites

Normal distribution, chi-squared distribution.

3.38.3 Theory

If $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_\nu$ are independent, then

$$T = \frac{Z}{\sqrt{V/\nu}} \sim t_\nu,$$

the t-distribution with ν **degrees of freedom**.

PDF:

$$f(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)}\left(1+\frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

Properties:

- Symmetric about 0.
- $E[T] = 0$ for $\nu > 1$; $\mathrm{Var}(T) = \nu/(\nu-2)$ for $\nu > 2$.
- Heavier tails than normal; excess kurtosis $6/(\nu-4)$ for $\nu > 4$.
- As $\nu \to \infty$, $t_\nu \to \mathcal{N}(0,1)$.

Pivotal quantity: for iid normal $X_1, \ldots, X_n$,

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

This is the basis of the one-sample t-test and its confidence interval.

3.38.4 Assumptions

Normal underlying data (or large enough n for CLT to make the approximation reasonable).

3.38.5 R Implementation

```r
# PDF comparison
x ← seq(-5, 5, length.out = 400)
plot(x, dnorm(x), type = "l", col = "black", lwd = 2,
     main = "t with 3, 10, 30 df vs N(0,1)", ylab = "f(x)")
lines(x, dt(x, df = 3),  col = "#F4A261", lwd = 2)
lines(x, dt(x, df = 10), col = "#6A4C93", lwd = 2)
lines(x, dt(x, df = 30), col = "#2A9D8F", lwd = 2)
legend("topright", c("normal", "t(3)", "t(10)", "t(30)"),
       col = c("black", "#F4A261", "#6A4C93", "#2A9D8F"), lwd = 2)

# Verify construction
n ← 1e5; nu ← 5
Z ← rnorm(n); V ← rchisq(n, df = nu)
T_sim ← Z / sqrt(V / nu)
c(emp_mean = mean(T_sim), theoretical = 0,
  emp_var  = var(T_sim),  theoretical_var = nu / (nu - 2))

# Small-sample CI for a mean
set.seed(2026)
X ← rnorm(15, mean = 10, sd = 3)
t_crit ← qt(0.975, df = length(X) - 1)
mean(X) + c(-1, 1) * t_crit * sd(X) / sqrt(length(X))
```

3.38.6 Output & Results

```
emp_mean theoretical  emp_var theoretical_var
  -0.003           0    1.665           1.667

[1]  8.76  12.06     # 95% CI using t_14
```

Empirical moments match theoretical at $\nu = 5$, and the small-sample CI properly uses t_{14} critical values.

3.38.7 Interpretation

Reporting a t-interval or t-test with "$t(n-1) = \ldots$" in a manuscript is reporting a t-distribution-based inference. The df equals $n - 1$ for the one-sample t-test; Welch's two-sample t-test uses Satterthwaite df that may be fractional.

3.38.8 Practical Tips

- For $\nu \geq 30$ the t is nearly indistinguishable from the normal; for $\nu \leq 10$ the tail differences are substantial.
- The t-distribution is the one-parameter generalisation of the normal that allows heavier tails.
- Robust regression models sometimes use a t-likelihood to achieve outlier resistance.
- Non-central t (with non-centrality μ) governs power calculations for t-tests.
- For very heavy tails ($\nu \leq 2$), variance is infinite; mean exists only for $\nu > 1$.

3.38.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.38.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.39 Transformations of Random Variables

3.39.1 Introduction

Given a random variable X and a function g, the distribution of $Y = g(X)$ is determined by the distribution of X and the function. The change-of-variables formula computes the new distribution; for strictly monotone differentiable g this involves a Jacobian factor.

3.39.2 Prerequisites

Random variables, PDF, basic calculus.

3.39.3 Theory

Univariate, monotone case. Let X have density f_X on support $\mathcal{X}$, and let g be strictly monotone and differentiable on $\mathcal{X}$. Then $Y = g(X)$ has density

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}(y)}{dy} \right|.$$

The absolute value of the derivative of the inverse is the **Jacobian**.

Non-monotone case. If g is piecewise monotone, partition the domain into monotone segments and sum the contributions:

$$f_Y(y) = \sum_i f_X(x_i) / |g'(x_i)|$$

where $\{x_i\}$ are the solutions of $g(x) = y$.

Multivariate case. For $\mathbf{Y} = g(\mathbf{X})$ with $g : \mathbb{R}^n \to \mathbb{R}^n$ a diffeomorphism,

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(g^{-1}(\mathbf{y})) \, |\det J_{g^{-1}}(\mathbf{y})|,$$

where $J_{g^{-1}}$ is the Jacobian matrix of g^{-1}.

Common examples:

- $Y = aX + b$: $f_Y(y) = f_X((y-b)/a)/|a|$.
- $Y = X^2$, X standard normal: $Y \sim \chi^2_1$.
- $Y = e^X$, X normal: Y is log-normal.
- $Y = F^{-1}(U)$, $U \sim \text{Uniform}(0,1)$: the inverse-CDF method for generating arbitrary distributions.

3.39.4 Assumptions

- g must be measurable.
- For the change-of-variables formula to apply, g must be strictly monotone and differentiable (or piecewise so).

3.39.5 R Implementation

```r
set.seed(2026)

# Case 1: Y = X^2 for X ~ N(0, 1); Y should be chi-squared with 1 df
X <- rnorm(1e5)
Y <- X^2

par(mfrow = c(1, 2))
hist(Y, breaks = 100, freq = FALSE, main = "Empirical", col = "#2A9D8F")
curve(dchisq(x, df = 1), from = 0, to = 10, add = TRUE, col = "#F4A261", lwd = 2)

# Case 2: Y = exp(X) for X ~ N(mu, sigma) is log-normal
mu <- 1; sigma <- 0.5
X2 <- rnorm(1e5, mu, sigma)
Y2 <- exp(X2)
curve(dlnorm(x, mu, sigma), from = 0, to = 10,
      main = "exp(N(1, 0.25)) vs lognormal PDF",
      col = "#2A9D8F", lwd = 2)
hist(Y2, breaks = 100, freq = FALSE, add = TRUE, border = "#F4A261")

par(mfrow = c(1, 1))

# Case 3: Inverse-CDF method to generate exponential from uniform
U <- runif(1e5)
E <- -log(1 - U)    # F^{-1}(u) = -log(1 - u) for rate = 1
c(mean = mean(E), theoretical = 1)
```

3.39.6 Output & Results

```
    mean  theoretical
   0.999        1.000
```

Empirical histograms match the theoretical transformed densities. The inverse-CDF method recovers the exponential mean.

3.39.7 Interpretation

Many “derived” distributions in statistics (chi-squared from normals, log-normal from normals, Cauchy from normals, Beta from gammas) arise as transformations of simpler distributions. Recognising the transformation short-cuts many derivations.

3.39.8 Practical Tips

- Always check strict monotonicity before applying the simple Jacobian formula; square-mapping X^2 is a classic non-monotone case.
- For inverse-CDF simulation, you need the CDF’s inverse; use `qnorm`, `qgamma`, etc., in R.
- The Jacobian for a multivariate linear transformation $Y = AX$ is $|det A|$, independent of location.
- Delta method is the “linearised” version of the change-of-variables formula, used for asymptotic distributions.

- Non-bijective transformations (e.g., $|X|$) require summing over preimages, as in the non-monotone case.

3.39.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.39.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.40 The Uniform Distribution

3.40.1 Introduction

The uniform distribution spreads probability equally across its support. The continuous uniform on $[a, b]$ is the building block of random-number generation; the discrete uniform on $\{1, \ldots, k\}$ describes fair dice and random selection. It is the simplest non-trivial distribution.

3.40.2 Prerequisites

Continuous and discrete random variables.

3.40.3 Theory

Continuous uniform $\mathrm{Uniform}(a, b)$ has PDF

$$f(x) = \frac{1}{b-a}, \qquad a \le x \le b,$$

with moments $E[X] = (a+b)/2$ and $\mathrm{Var}(X) = (b-a)^2/12$.

Discrete uniform on $\{x_1, \ldots, x_k\}$: $P(X = x_i) = 1/k$ for each i. For $\{1, \ldots, k\}$: $E[X] = (k+1)/2$, $\mathrm{Var}(X) = (k^2-1)/12$.

Probability integral transform: if X has a continuous CDF F_X, then $F_X(X) \sim \mathrm{Uniform}(0, 1)$. The inverse: $X = F_X^{-1}(U)$ where $U \sim \mathrm{Uniform}(0, 1)$ generates X from uniform draws. This is the inverse-CDF method of simulation.

3.40.4 Assumptions

Continuous uniform assumes the variable genuinely has constant density; discrete uniform assumes genuinely equal probabilities across a specified set.

3.40.5 R Implementation

```
set.seed(2026)

# Continuous uniform
U ← runif(1e5, min = 0, max = 1)
c(mean = mean(U), theoretical = 0.5,
  var  = var(U),  theoretical_var = 1/12)

# Inverse-CDF method: generate exponential from uniform
lambda ← 2
E ← -log(1 - U) / lambda   # F^{-1}(u) = -log(1 - u) / lambda
c(empirical_mean = mean(E), exponential_mean = 1 / lambda)

# Discrete uniform via sample()
die ← sample(1:6, 1e5, replace = TRUE)
c(mean_die = mean(die), theoretical = 3.5,
  var_die  = var(die),  theoretical_var = 35/12)
```

3.40.6 Output & Results

```
       mean theoretical           var  theoretical_var
     0.5003      0.5000       0.0835           0.0833

 empirical_mean  exponential_mean
         0.4983             0.500

   mean_die  theoretical     var_die  theoretical_var
      3.504        3.500       2.924            2.917
```

Empirical moments match theoretical exactly up to Monte Carlo error, and the inverse-CDF method recovers the exponential mean.

3.40.7 Interpretation

Uniforms rarely appear as substantive models of real data but pervade simulation, randomisation, and Monte Carlo methods. Reporting “trial assignment was simulated via `sample()`” is implicitly invoking a discrete uniform.

3.40.8 Practical Tips

- `runif(n)` defaults to $[0, 1]$; specify min/max explicitly to avoid surprises.
- The uniform distribution has zero skewness and excess kurtosis -1.2 (light tails).
- For true randomness from hardware sources, avoid pseudo-random `runif`; use `/dev/urandom` or the `random` R package with online services.
- Uniform priors are often overly informative on transformed scales (a uniform on σ is not uniform on σ^2).

- The Jacobian of the inverse-CDF transform is $1/f_X(x)$; this is why the probability integral transform works.

3.40.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.40.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.41 Variance and Covariance

3.41.1 Introduction

Variance measures how far a single random variable ranges around its mean; covariance measures how two random variables move together around their means. Together they populate the covariance matrix of a random vector, which is the central object in multivariate statistics, regression, and portfolio theory.

3.41.2 Prerequisites

Expectation and basic properties of random variables.

3.41.3 Theory

Variance of X:

$$\operatorname{Var}(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2.$$

Covariance of X and Y:

$$\operatorname{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y].$$

Variance is a special case: $\operatorname{Cov}(X, X) = \operatorname{Var}(X)$.

Properties:

- Non-negativity: $\operatorname{Var}(X) \geq 0$, with equality iff X is constant almost surely.
- Scale: $\operatorname{Var}(aX + b) = a^2 \operatorname{Var}(X)$.
- Symmetry: $\operatorname{Cov}(X, Y) = \operatorname{Cov}(Y, X)$.
- Bilinearity: $\operatorname{Cov}(aX + b, cY + d) = ac \operatorname{Cov}(X, Y)$.

- Sum: $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$.
- Independence implies zero covariance: $X \perp Y \Rightarrow \mathrm{Cov}(X,Y) = 0$. The converse is false.

The **covariance matrix** of a random vector $\mathbf{X} = (X_1, \ldots, X_p)$ is the $p \times p$ matrix Σ with entries $\Sigma_{ij} = \mathrm{Cov}(X_i, X_j)$. Σ is symmetric positive semi-definite.

3.41.4 Assumptions

$E[X^2] < \infty$ and $E[Y^2] < \infty$ for variance and covariance to be defined.

3.41.5 R Implementation

```r
library(mvtnorm)
set.seed(2026)

# Univariate variance
X ← rnorm(1e5, mean = 10, sd = 3)
c(theoretical = 9, empirical = var(X))

# Covariance from bivariate normal with known Sigma
Sigma ← matrix(c(4, 1.8,
                 1.8, 9), nrow = 2)
n ← 1e5
XY ← rmvnorm(n, mean = c(0, 0), sigma = Sigma)
cov(XY)
cov(XY) / (sqrt(Sigma[1, 1] * Sigma[2, 2]))   # correlation = 0.3

# Variance of a sum verifies the formula
X ← XY[, 1]; Y ← XY[, 2]
c(var_sum = var(X + Y),
  formula = var(X) + var(Y) + 2 * cov(X, Y))

# Zero covariance does not imply independence
X ← rnorm(1e5)
Y ← X^2
c(cov = cov(X, Y),
  correlation = cor(X, Y))
```

3.41.6 Output & Results

```
theoretical   empirical
      9.0         9.04

          [,1]      [,2]
[1,]   3.998     1.791
[2,]   1.791     8.984

          [,1]      [,2]
[1,]   0.999     0.299
[2,]   0.299     0.998

 var_sum formula
  16.56   16.56

       cov   correlation
     0.003        0.002
```

Empirical variance and covariance match the theoretical values within Monte Carlo error. X and X^2 have zero covariance but are clearly dependent – covariance detects only linear association.

3.41.7 Interpretation

Reporting a mean and a variance (or SD) gives a rough summary of a distribution's location and spread; reporting a covariance matrix gives the joint structure of several variables. Most multivariate methods (PCA, MANOVA, linear discriminant analysis) work with sample covariance matrices.

3.41.8 Practical Tips

- Sample variance uses divisor $n - 1$ for unbiasedness; R's `var()` uses $n - 1$ by default.
- $\mathrm{Cov}(X, Y) = 0$ does not imply independence except in the Gaussian case.
- The covariance matrix must be positive semi-definite; estimation from small samples can produce near-singular matrices.
- Linearly transforming a random vector by matrix A changes the covariance matrix to $A\Sigma A^\top$.
- For robust multivariate covariance estimation in contaminated data, `robustbase::covMcd` is the standard MCD estimator.

3.41.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.41.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.42 The Weibull Distribution

3.42.1 Introduction

The Weibull distribution generalises the exponential to allow monotone non-constant hazard rates. It is the default parametric model in reliability engineering and a common choice for parametric survival analysis. The shape parameter controls whether hazard is increasing (ageing), decreasing (early failures dominate), or constant (exponential).

3.42.2 Prerequisites

Exponential distribution, hazard and survival functions.

3.42.3 Theory

A Weibull random variable with shape $k > 0$ and scale $\lambda > 0$ has PDF

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right], \qquad x \geq 0.$$

CDF: $F(x) = 1 - \exp\big[-(x/\lambda)^{k}\big]$.

Hazard function: $h(x) = (k/\lambda)(x/\lambda)^{k-1}$. Three regimes:

- $k < 1$: decreasing hazard (early failures / "infant mortality").
- $k = 1$: constant hazard (exponential).
- $k > 1$: increasing hazard (wear-out / ageing).

Moments:

- $E[X] = \lambda\,\Gamma(1 + 1/k)$.
- $\mathrm{Var}(X) = \lambda^2\left[\Gamma(1 + 2/k) - \Gamma(1 + 1/k)^2\right]$.

Accelerated failure time form: $\log T = \mu + \sigma W$ where W is standard extreme-value; the AFT parameters relate to Weibull's $k = 1/\sigma$ and $\lambda = e^{\mu}$.

3.42.4 Assumptions

Non-negative continuous data, monotone hazard (log-hazard linear in log-time).

3.42.5 R Implementation

```r
library(survival)

# PDFs for three shape regimes
x <- seq(0, 5, length.out = 400)
plot(x, dweibull(x, shape = 0.5, scale = 1), type = "l", col = "#F4A261",
     lwd = 2, main = "Weibull PDFs", ylab = "f(x)")
lines(x, dweibull(x, shape = 1, scale = 1), col = "#6A4C93", lwd = 2)
lines(x, dweibull(x, shape = 2.5, scale = 1), col = "#2A9D8F", lwd = 2)
legend("topright", c("k = 0.5 (decreasing h)", "k = 1 (exponential)", "k = 2.5
(increasing h)"),
       col = c("#F4A261", "#6A4C93", "#2A9D8F"), lwd = 2)

# Simulate and fit
set.seed(2026)
X <- rweibull(500, shape = 2.0, scale = 1.5)
fit <- survreg(Surv(X, rep(1, length(X))) ~ 1, dist = "weibull")
summary(fit)

# Convert survreg output to Weibull parameters
c(shape = 1 / fit$scale,
  scale = exp(coef(fit)))
```

3.42.6 Output & Results

```
Call: survreg(...)

                Value Std. Error    z      p
(Intercept)  0.40578    0.01574 25.78  <2e-16
Log(scale)  -0.69312    0.03293 -21.0  <2e-16

Scale = 0.500

   shape    scale
   2.000    1.501
```

The fit recovers $k \approx 2$ and $\lambda \approx 1.5$ – the true parameters. The increasing-hazard regime is visible in the PDF.

3.42.7 Interpretation

In survival analysis, a Weibull model is often the starting parametric fit: “we fitted a Weibull regression with shape $k = 1.3$ (95 % CI 1.1-1.5), indicating increasing hazard over follow-up.” Weibull is the default choice when the proportional-hazards assumption holds on a transformed scale.

3.42.8 Practical Tips

- R’s `rweibull(n, shape, scale)` conventions differ from `survreg`, which parameterises via intercept and scale = $1/k$.
- For right-censored data, use `survreg(..., dist = "weibull")`, not `fitdistr` on complete data.

- Check the fit via a log-log plot: $\log(-\log S(t))$ vs $\log t$ should be approximately linear.
- If shape estimates are very close to 1 with narrow CI, consider the simpler exponential.
- `flexsurv::flexsurvreg(..., dist = "weibull")` is a modern alternative with richer diagnostics.

3.42.9 For Reviewers

Use the shared reviewer checklist, together with this method's assumptions and practical tips.

Method-specific reviewer notes have not yet been developed for this entry. The links above support a structured review; they do not certify the example or a manuscript's application of the method.

3.42.10 See also — labs in this chapter

- Probability, conditional probability, Bayes' theorem
- Discrete distributions: Bernoulli, binomial, Poisson
- Continuous distributions and Q-Q plots

3.43 Week 2, Session 5 — Continuous distributions and Q-Q plots

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

3.43.1 Learning objectives

- Draw densities of the normal, *t*, chi-square, *F*, and exponential distributions and describe how shape changes with parameters.
- Use Q-Q plots to assess whether data are consistent with a hypothesised distribution.
- Recognise the degrees-of-freedom relationships among the normal-derived distributions.

3.43.2 Prerequisites

Lab 2.4.

3.43.3 Background

The continuous distributions in this lab cover the majority of routine parametric inference. The normal is the distribution of means under the central limit theorem. The *t* is what happens to the normal when the variance is itself estimated from the data — with more

uncertainty (smaller n), heavier tails. The chi-square is the distribution of sums of squared standard normals. The F is a ratio of chi-squares. The exponential appears in survival and queueing.

The Q-Q plot is the workhorse for assessing distributional fit. Plotting sample quantiles against theoretical quantiles gives a straight line when the model is right. Deviations at the tails indicate skew or heavy tails; an S-shape indicates non-normal symmetric behaviour. Q-Q plots are how assumptions of the t-test and the analysis of variance get checked in practice.

Degrees of freedom are easy to mis-remember. The short version: t has df = n − 1 for a one-sample test; chi-square from a sum of k squared standard normals has df = k; F has two df — numerator and denominator — that come from the chi-squares in its ratio.

3.43.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

3.43.5 1. Hypothesis

We are not testing a hypothesis. We are characterising five continuous distributions and using Q-Q plots to assess whether a sample is consistent with a model.

3.43.6 2. Visualise

Plot densities side by side.

```
grid ← tibble(x = seq(-5, 5, length.out = 400))

densities ← bind_rows(
  grid ▷ mutate(density = dnorm(x),          dist = "Normal(0,1)"),
  grid ▷ mutate(density = dt(x, df = 3),     dist = "t(3)"),
  grid ▷ mutate(density = dt(x, df = 30),    dist = "t(30)")
)
```

```
ggplot(densities, aes(x, density, colour = dist)) +
  geom_line(linewidth = 1) +
  labs(x = NULL, y = "Density", colour = NULL)
```

The t with 3 df has visibly heavier tails; at 30 df it is nearly normal.

```
chi_grid ← tibble(x = seq(0.01, 20, length.out = 400))
chi ← bind_rows(
  chi_grid ▷ mutate(density = dchisq(x, df = 2), dist = "chi-sq(2)"),
  chi_grid ▷ mutate(density = dchisq(x, df = 5), dist = "chi-sq(5)"),
  chi_grid ▷ mutate(density = dchisq(x, df = 10), dist = "chi-sq(10)")
)
```

```
ggplot(chi, aes(x, density, colour = dist)) +
  geom_line(linewidth = 1) +
  labs(x = NULL, y = "Density", colour = NULL)
```

3.43.7 3. Assumptions

We assume samples are independent and identically distributed. For the Q-Q check, the sample must be reasonably sized (say, at least 20 observations) for the plot to be informative.

```
# Facet across df for the F distribution.
f_grid ← expand_grid(
  x   = seq(0.01, 5, length.out = 300),
  df1 = c(2, 5, 10),
  df2 = c(10, 30)
) ▷
  mutate(density = df(x, df1, df2))
```

```
f_grid ▷
  ggplot(aes(x, density, colour = factor(df1))) +
  geom_line(linewidth = 1) +
  facet_wrap(~ df2, labeller = label_both) +
  labs(x = NULL, y = "Density", colour = "df1")
```

3.43.8 4. Conduct

Q-Q plots. Simulate data, compare to a hypothesised distribution.

```
n ← 200
samples ← tibble(
  normal_sample = rnorm(n),
  heavy_tail    = rt(n, df = 3),
  exponential   = rexp(n, rate = 1)
)
```

```
samples ▷
  pivot_longer(everything(), names_to = "sample", values_to = "value") ▷
  ggplot(aes(sample = value)) +
  stat_qq(alpha = 0.6) +
  stat_qq_line(colour = "firebrick") +
  facet_wrap(~ sample, scales = "free") +
  labs(x = "Theoretical normal quantiles",
       y = "Sample quantiles")
```

The normal sample lies on the 45° reference line; the heavy-tailed sample splays at both ends; the exponential curves away on one side.

Shapiro-Wilk as a quick numerical check on the three samples.

```r
sw_results ← samples ▷
  pivot_longer(everything(), names_to = "sample", values_to = "value") ▷
  group_by(sample) ▷
  summarise(p = shapiro.test(value)$p.value, .groups = "drop")
sw_results
```

3.43.9 5. Concluding statement

> The normal sample (*n* = `n`) was consistent with normality both visually (Q-Q line) and by Shapiro-Wilk (*p* = `signif(sw_results$p[sw_results$sample=="normal_sample"], 2)`). The *t*(3) sample showed clear tail departures, Shapiro-Wilk *p* = `signif(sw_results$p[sw_results$sample=="heavy_tail"], 2)`. The exponential sample was strongly right-skewed, $p < 0.001$.

A Q-Q plot is the most information-dense way to inspect a distributional assumption. Numerical tests of normality (Shapiro-Wilk, Kolmogorov-Smirnov) should corroborate, not replace, the plot.

3.43.10 Common pitfalls

- Reading the centre of a Q-Q plot and ignoring the tails. The tails are where the interesting departures live.
- Declaring a distribution wrong on Shapiro-Wilk in a sample of 5000 when the Q-Q plot is straight.
- Forgetting that *t* approaches normal as df → ∞; do not test with df = 500 in a fit-diagnostic mindset.

3.43.11 Further reading

- Rice JA. *Mathematical Statistics and Data Analysis*.
- Kutner MH et al. *Applied Linear Statistical Models*.

3.43.12 Session info

```r
sessionInfo()
```

3.43.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

3.44 Week 2, Session 4 — Discrete distributions: Bernoulli, binomial, Poisson

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

3.44.1 Learning objectives

- Simulate Bernoulli trials and aggregate them to a binomial.
- Show the Poisson approximation to the binomial when n is large and p is small.
- Use `dbinom()`, `pbinom()`, and `rbinom()` (and their Poisson counterparts) to answer probability questions.

3.44.2 Prerequisites

Lab 2.2 (probability).

3.44.3 Background

The Bernoulli distribution is the simplest non-trivial distribution: one coin, two outcomes, a single probability p. The binomial adds structure — n independent Bernoulli trials with the same p — and gives the distribution of the count of successes. The Poisson arises in the limit where n is large, p is small, and np stays fixed; it describes counts of rare events over a fixed window of opportunity, without needing to specify n and p separately.

Most count-like quantities in biomedicine can be modelled by one of these three. Whether a tumour is responding (Bernoulli), how many of twenty patients relapse (binomial), how many emergency admissions arrive in an hour (Poisson) — the same three distributions cover a great deal of ground. The simulation-first approach makes them concrete before we write a formula.

There is a common conflation between the binomial and the Poisson in the clinical literature. When the sample size is large and the event is rare, the two distributions give nearly identical probabilities and either is defensible. When the event is not rare, the Poisson can assign meaningful probability to impossible events (more successes than trials), so the binomial is strictly better.

3.44.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

3.44.5 1. Hypothesis

We characterise three discrete distributions by simulation, then compare the empirical frequencies to the theoretical probabilities. No hypothesis is tested.

3.44.6 2. Visualise

A Bernoulli trial: one coin, success probability 0.3.

```
n_trials ← 1000
bern ← rbinom(n_trials, 1, 0.3)
mean(bern)   # empirical p
```

A binomial count: how many successes in 20 trials each at p = 0.3, repeated 10,000 times.

```
n_rep ← 10000
bin  ← rbinom(n_rep, size = 20, prob = 0.3)

bin_tibble ← tibble(k = bin) ▷
  count(k) ▷
  mutate(prop = n / sum(n),
         theo = dbinom(k, 20, 0.3))
```

```
bin_tibble ▷
  ggplot(aes(k, prop)) +
  geom_col(fill = "grey70", alpha = 0.8) +
  geom_point(aes(y = theo), colour = "firebrick", size = 2) +
  labs(x = "Successes in 20 trials",
       y = "Proportion (bars) vs. P(k) (points)")
```

3.44.7 3. Assumptions

Binomial assumes fixed n, independence, and constant p. Poisson assumes a constant rate and independent counts over disjoint intervals. When n is large and p is small with $np = \lambda$, the binomial approaches Poisson.

```
# n=1000, p=0.003, so np = 3. Compare binomial and Poisson probabilities.
probs ← tibble(
  k = 0:10,
  binom  = dbinom(k, 1000, 0.003),
  pois   = dpois(k, 3)
)
probs
```

3.44.8 4. Conduct

Answer two specific questions.

```
# Q1: In a trial of 20 patients at p=0.3 response rate, what is
# P(at least 8 responders)?
p_ge8 ← 1 - pbinom(7, size = 20, prob = 0.3)
p_ge8

# Q2: An ER admits patients at a rate of 3/hour. What is
# P(7 or more patients in a given hour)?
p_pois_ge7 ← 1 - ppois(6, lambda = 3)
p_pois_ge7
```

Simulate Q2 to confirm.

```
sim ← rpois(10000, lambda = 3)
mean(sim ⩾ 7)
```

```
tibble(k = 0:15) ▷
  mutate(pois_density = dpois(k, 3)) ▷
  ggplot(aes(k, pois_density)) +
  geom_col(fill = "steelblue") +
  labs(x = "Number of events", y = "P(X = k)")
```

3.44.9 5. Concluding statement

> A response rate of 0.3 in a 20-patient trial implies a probability of `round(p_ge8, 3)` of observing eight or more responders. An ER with a mean admission rate of 3/hour has a probability of `round(p_pois_ge7, 3)` of receiving seven or more admissions in a given hour; 10,000 simulated hours recovered an empirical proportion of `round(mean(sim ⩾ 7), 3)`. The binomial with n = 1000, p = 0.003 and the Poisson with λ = 3 gave nearly indistinguishable probabilities across k = 0–10.

Discrete distributions are under-appreciated. Many designs that are reached for as continuous (eg regressions on a rate) have a cleaner count-based model sitting next to them.

3.44.10 Common pitfalls

- Using the normal approximation to the binomial when n is small. For small n use `pbinom()` directly.
- Modelling a rate as a proportion when the denominator is person-time, not persons.
- Treating a Poisson with a known λ as if its variance had to be estimated. For Poisson, mean = variance.

3.44.11 Further reading

- Rosner B. *Fundamentals of Biostatistics*, chapter on discrete distributions.
- Agresti A. *Categorical Data Analysis*, opening chapter.

3.44.12 Session info

```
sessionInfo()
```

3.44.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

3.45 Week 2, Session 2 — Probability, conditional probability, Bayes' theorem

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

3.45.1 Learning objectives

- Distinguish marginal, joint, and conditional probability and compute each from a contingency table.
- State Bayes' theorem and apply it to a medical-testing scenario.
- Simulate a population and recover conditional probabilities by counting.

3.45.2 Prerequisites

Basic probability.

3.45.3 Background

Probability is the language for reasoning under uncertainty. In biostatistics, most of the probabilities we care about are conditional — the probability that a patient has a disease *given* a test result; the probability of survival *given* a treatment; the probability of an adverse event *given* a risk factor. The distinction between marginal and conditional probability is the distinction between "how common is X in the population" and "how common is X in a particular subgroup".

Bayes' theorem is the rule for flipping a conditional probability. Given $P(A|B)$ and the marginal probabilities $P(A)$ and $P(B)$, it returns $P(B|A)$. In medical testing this lets us move from *how often does the test fire when the disease is present* (sensitivity) to *how likely is disease given a positive test* (positive predictive value). The two are routinely confused; Bayes' theorem is the cure.

The counterintuitive part of Bayes is the role of prevalence. For a test with fixed sensitivity and specificity, PPV rises and falls with the prevalence in the tested population. Screening a very-low-risk population with an imperfect test guarantees that most positives will be false positives, regardless of the test's quality.

3.45.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

3.45.5 1. Hypothesis

We are not testing a hypothesis in the Fisherian sense; we are computing a conditional probability. The quantity of interest is the positive predictive value of a test with known sensitivity and specificity in a population of known prevalence.

3.45.6 2. Visualise

Simulate a population of 100,000 people. Prevalence of disease = 1%. Test sensitivity = 95%. Test specificity = 95%.

```
N ← 100000
prev ← 0.01
sens ← 0.95
spec ← 0.95

pop ← tibble(
  id = seq_len(N),
  disease = rbinom(N, 1, prev)
) ▷
  mutate(
    test = if_else(disease == 1,
                   rbinom(N, 1, sens),
                   rbinom(N, 1, 1 - spec))
  )
```

```
tab ← table(disease = pop$disease, test = pop$test)
tab
```

```
as_tibble(tab) ▷
  mutate(disease = recode(disease, `0` = "no disease", `1` = "disease"),
         test    = recode(test,    `0` = "neg", `1` = "pos")) ▷
  ggplot(aes(test, disease, fill = n)) +
  geom_tile() +
  geom_text(aes(label = n), colour = "white") +
  scale_fill_viridis_c() +
  labs(x = "Test result", y = "True state", fill = "Count")
```

3.45.7 3. Assumptions

We assume each person is independent, the test's operating characteristics are fixed, and no selection bias. Real screening has verification bias, spectrum bias, and imperfect gold standards; we are ignoring all of them to isolate the Bayesian calculation.

```
p_disease ← mean(pop$disease)
p_positive ← mean(pop$test)
p_pos_given_disease ← mean(pop$test[pop$disease == 1])
p_pos_given_no_disease ← mean(pop$test[pop$disease == 0])

tibble(quantity = c("P(D)", "P(T+)",
                    "P(T+|D)", "P(T+|~D)"),
       value = c(p_disease, p_positive,
                 p_pos_given_disease, p_pos_given_no_disease))
```

3.45.8 4. Conduct

Apply Bayes' theorem by hand.

```
ppv_bayes ← (sens * prev) / (sens * prev + (1 - spec) * (1 - prev))
ppv_bayes

# By direct counting from the simulation.
ppv_count ← mean(pop$disease[pop$test == 1])
ppv_count
```

Same quantity, two routes. Bayes is algebra on the marginals; simulation is counting from the table. They must agree, within sampling error.

```
# Sweep prevalence from 0.001 to 0.5 and see PPV change.
sweep ← tibble(
  prev = seq(0.001, 0.5, length.out = 200),
  ppv = (sens * prev) / (sens * prev + (1 - spec) * (1 - prev))
)
```

```
sweep ▷
  ggplot(aes(prev, ppv)) +
  geom_line(linewidth = 1) +
  geom_hline(yintercept = 0.5, linetype = 2, colour = "grey50") +
  labs(x = "Prevalence", y = "Positive predictive value")
```

3.45.9 5. Concluding statement

With sensitivity = `sens` and specificity = `spec`, a test applied in a population of prevalence = `prev` has a positive predictive value of `round(ppv_bayes, 3)` — that is, fewer than one in five positive results is a true positive. In a simulated population of `format(N, big.mark = ",")`, the PPV recovered by direct counting was `round(ppv_count, 3)`. Negative predictive value was correspondingly high.

When prevalence is low, even a very accurate test produces mostly false positives. This is why population screening decisions are controversial, and why "my test is 95% accurate" is an incomplete statement.

3.45.10 Common pitfalls

- Confusing sensitivity with PPV. $P(\text{T+}|\text{D})$ and $P(\text{D}|\text{T+})$ are not the same.
- Reporting a test's "accuracy" without specifying prevalence.
- Assuming independence of repeated tests without checking.
- Treating the prior (prevalence) as negotiable evidence.

3.45.11 Further reading

- Gigerenzer G. *Reckoning with Risk*.
- Altman DG & Bland JM (1994), *Diagnostic tests 2: predictive values*.

3.45.12 Session info

```
sessionInfo()
```

3.45.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)