4 Inferential Statistics

R. Heller · Date: 2026-05-11 · Last changed: 2026-10-30

Classical inference: the Central Limit Theorem, MLE, NHST and its philosophy, t-tests and proportions, goodness-of-fit, correlation, and the non-parametric counterparts. This is the largest chapter and the one most peer reviewers will read first.

This chapter contains **45 method pages** and **9 labs**. If you are not sure which method to read, return to Chapter 0 and follow the decision tree to the right node.

4.1 Method pages

Method	Source slug
Anderson-Darling Test	`anderson-darling-test`
Bartlett's Test	`bartlett-test`
Benjamini-Hochberg FDR	`benjamini-hochberg-fdr`
Benjamini-Yekutieli FDR	`benjamini-yekutieli`
Bonferroni Correction	`bonferroni-correction`
Bootstrap Confidence Intervals	`bootstrap-confidence-intervals`
Chi-Squared Goodness-of-Fit Test	`chi-squared-goodness-of-fit`
Chi-Squared Test of Independence	`chi-squared-independence`
Cochran-Mantel-Haenszel Test	`cochran-mantel-haenszel`
Effect Sizes: Overview	`effect-sizes-overview`
Equivalence Testing with TOST	`equivalence-tost`
Fisher's Exact Test	`fisher-exact-test`
Friedman Test	`friedman-test`
Holm's Step-Down Correction	`holm-correction`
Kendall's Tau	`kendall-tau`
Kolmogorov-Smirnov Test	`kolmogorov-smirnov-test`
Kruskal-Wallis Test	`kruskal-wallis`
Levene's Test of Variances	`levene-test`
Mann-Whitney U Test	`mann-whitney-u`
McNemar's Test	`mcnemar-test`
Mixed ANOVA	`mixed-anova`
Multiple Comparisons: Overview	`multiple-comparisons-overview`
Null and Alternative Hypotheses	`null-alternative-hypotheses`
One-Proportion Test	`one-proportion-test`
One-Sample t-Test	`one-sample-t-test`

Method	Source slug
One-Sample z-Test	`one-sample-z-test`
One-Way ANOVA	`one-way-anova`
P-Values Explained	`p-values-explained`
Paired t-Test	`paired-t-test`
Pearson Correlation Test	`pearson-correlation-test`
Permutation Tests	`permutation-tests`
Post-Hoc Tests with Tukey HSD	`post-hoc-tests-tukey`
Repeated-Measures ANOVA	`repeated-measures-anova`
Shapiro-Wilk Normality Test	`shapiro-wilk-test`
Spearman Rank Correlation Test	`spearman-correlation-test`
The Bootstrap: Introduction	`bootstrap-introduction`
The Jackknife	`jackknife`
The Runs Test	`runs-test`
The Sign Test	`sign-test`
Two-Proportion Test	`two-proportion-test`
Two-Sample t-Test	`two-sample-t-test`
Two-Way ANOVA	`two-way-anova`
Type I and Type II Errors	`type-i-type-ii-errors`
Welch's t-Test	`welch-t-test`
Wilcoxon Signed-Rank Test	`wilcoxon-signed-rank`

4.2 Labs

Lab
Populations, samples, and the central limit theorem
Bootstrap and permutation tests
Maximum likelihood estimation
One-sample t-test and one-proportion test
Hypothesis testing, p-values, type I/II errors
Two-sample and paired t-tests
Two proportions, chi-square, risk and odds
Pearson and Spearman correlation
Non-parametric tests

4.3 Anderson-Darling Test

4.3.1 Introduction

The Anderson-Darling (AD) test is a goodness-of-fit test that weights the tails of the distribution more heavily than the Kolmogorov-Smirnov test. It is widely used for normality testing in engineering and quality-control contexts, where tail behaviour often determines pass/fail decisions.

4.3.2 Prerequisites

Empirical CDF, goodness-of-fit testing.

4.3.3 Theory

For a hypothesised CDF F and empirical CDF $\hat{F}_n$, the AD statistic is

$$A^2 = n \int \frac{(\hat{F}_n(x) - F(x))^2}{F(x)(1 - F(x))} dF(x).$$

The denominator $F(1-F)$ gives less weight to the middle of the distribution and more to the tails, where departures from the hypothesised distribution are often clinically or industrially important.

A computational form for a sample sorted as $y_{(1)} \leq \ldots \leq y_{(n)}$:

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i - 1) \left[\log F(y_{(i)}) + \log(1 - F(y_{(n-i+1)})) \right].$$

Critical values depend on whether the distribution's parameters are known or estimated from the sample.

4.3.4 Assumptions

- iid observations.
- The hypothesised distribution is fully specified or parameters are consistently estimated.

4.3.5 R Implementation

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

# Normal sample: AD should not reject
x_norm ← rnorm(100)
ad.test(x_norm)

# Heavy-tailed sample: AD should reject
x_t3 ← rt(100, df = 3)
ad.test(x_t3)

# Compare to Shapiro-Wilk
shapiro.test(x_norm)$p.value
shapiro.test(x_t3)$p.value
```

4.3.6 Output & Results

```
        Anderson-Darling normality test
data:  x_norm
A = 0.35, p-value = 0.47          # normality not rejected

        Anderson-Darling normality test
data:  x_t3
A = 2.10, p-value = 2.8e-05        # strongly rejects normality

Shapiro-Wilk p-values:
  normal sample: 0.42
  t(3) sample:   0.0001
```

Both tests correctly reject the heavy-tailed sample; AD has moderately higher power than Shapiro-Wilk in the tails.

4.3.7 Interpretation

"The Anderson-Darling test did not reject normality for the biomarker residuals ($A = 0.35$, $p = 0.47$), supporting the use of parametric methods."

4.3.8 Practical Tips

- AD is more sensitive to tail departures than Kolmogorov-Smirnov and often more powerful than Shapiro-Wilk for detecting heavy tails.
- `nortest::ad.test()` handles normality; `goftest::ad.test()` supports arbitrary reference distributions.
- For testing fit to a parametric distribution with estimated parameters, use the correct tabled critical values (nortest handles this for normality).
- Very large samples will reject any hypothesised distribution because real data have atoms and rounding.
- For quality-control contexts where tail behaviour governs defect rates, AD is the preferred GoF test.

4.3.9 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.

4.3.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.4 Bartlett's Test

4.4.1 Introduction

Bartlett's test evaluates homogeneity of variance across k groups. It is the likelihood-ratio test under normality, and it is the most powerful variance-homogeneity test when normality holds. Its major drawback: it is extremely sensitive to non-normality, losing validity as soon as the normality assumption is relaxed.

4.4.2 Prerequisites

Variance, ANOVA assumptions.

4.4.3 Theory

For k groups with sample sizes n_i and sample variances s_i^2, let $N = \sum n_i$ and pooled variance $s_p^2 = \frac{\sum (n_i - 1) s_i^2}{N - k}$. The test statistic is

$$K = \frac{(N-k)\ln s_p^2 - \sum (n_i - 1)\ln s_i^2}{1 + \frac{1}{3(k-1)}\left(\sum \frac{1}{n_i - 1} - \frac{1}{N-k}\right)} \sim \chi^2_{k-1} \text{ under } H_0.$$

Rejecting H_0 indicates variances differ; failing to reject suggests equality.

4.4.4 Assumptions

- Independent observations within and across groups.
- **Normality of within-group distributions** – strongly required.
- Moderate sample sizes (>= 5 per group).

4.4.5 R Implementation

```r
set.seed(2026)

df ← data.frame(
  group = factor(rep(c("A", "B", "C"), each = 40)),
  y     = c(rnorm(40, 50, 5),
            rnorm(40, 55, 5),
            rnorm(40, 52, 5))
)

bartlett.test(y ~ group, data = df)

# Under normality, Bartlett and Levene should agree
library(car)
leveneTest(y ~ group, data = df)

# Under non-normality: Bartlett over-rejects
df2 ← df
df2$y ← c(rexp(40, rate = 1/50),
           rexp(40, rate = 1/55),
           rexp(40, rate = 1/52))
bartlett.test(y ~ group, data = df2)
leveneTest(y ~ group, data = df2)
```

4.4.6 Output & Results

```
Under normality:
  Bartlett's K-squared = 0.08, df = 2, p-value = 0.96
  Levene's F = 0.10, p-value = 0.91

Under exponential:
  Bartlett's K-squared = 8.7, df = 2, p-value = 0.013    # false positive
  Levene's F = 1.5, p-value = 0.24                         # correctly holds
```

Bartlett falsely rejects equal variances under non-normal data. Levene correctly maintains its Type I rate.

4.4.7 Interpretation

Almost never use Bartlett in practice; prefer Levene (median-based) for robustness. Report Bartlett only when normality is firmly established (e.g., by design or by a prior transformation).

4.4.8 Practical Tips

- Bartlett is theoretically optimal under normality but rarely appropriate for real data.
- Avoid the Bartlett-then-ANOVA pipeline; use Welch's ANOVA directly.

- For skewed or heavy-tailed data, Levene or the Fligner-Killeen test are safer.
- The null hypothesis “all variances are equal” is rarely the clinical question; it is a diagnostic step for another test.
- Many software packages report Bartlett as default; override to Levene for applied work.

4.4.9 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.

4.4.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.5 Benjamini-Hochberg FDR

4.5.1 Introduction

Controlling the family-wise error rate becomes impractical when the number of tests runs into the thousands or millions (genomics, imaging). Benjamini and Hochberg (1995) introduced the **false discovery rate** (FDR) as a less stringent alternative: the expected proportion of false positives among rejected hypotheses. Their procedure is the default multiple-testing correction in high-throughput biology.

4.5.2 Prerequisites

Hypothesis testing, multiple-testing problem.

4.5.3 Theory

Let V be the number of false rejections and R the total number of rejections. The **FDR** is

$$\mathrm{FDR} = E\left[\frac{V}{\max(R,1)}\right].$$

The **Benjamini-Hochberg (BH) procedure** at level q:

1. Sort p-values: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
2. Find the largest k such that $p_{(k)} \leq kq/m$.
3. Reject all hypotheses with $p_{(i)} \leq p_{(k)}$.

Equivalently, the BH-adjusted p-value is

$$p_{(j)}^{\text{adj}} = \min\left[1, \min_{i \geq j}\left(\frac{m p_{(i)}}{i}\right)\right].$$

BH controls FDR at q under independence or positive dependence. For arbitrary dependence, use Benjamini-Yekutieli (BY), which is more conservative.

4.5.4 Assumptions

- Either independent p-values or positive regression dependence (PRDS).
- Each p-value is valid (uniform under its null).

4.5.5 R Implementation

```
set.seed(2026)

# 1000 tests: 100 true effects (random), 900 nulls
n_tests ← 1000; n_true ← 100
pvals ← c(runif(n_tests - n_true),     # nulls
          runif(n_true) * 0.01)         # true effects, very small p

pvals ← sample(pvals)   # random order

adj_bonf ← p.adjust(pvals, method = "bonferroni")
adj_holm ← p.adjust(pvals, method = "holm")
adj_bh   ← p.adjust(pvals, method = "BH")
adj_by   ← p.adjust(pvals, method = "BY")

q_target ← 0.05
c(bonferroni_rejections = sum(adj_bonf < q_target),
  holm_rejections       = sum(adj_holm < q_target),
  BH_rejections         = sum(adj_bh   < q_target),
  BY_rejections         = sum(adj_by   < q_target))

# Power approximation vs FDR control
true_signal ← c(rep(FALSE, n_tests - n_true), rep(TRUE, n_true))[match(sort(pvals),
pvals)]
sum(adj_bh[true_signal] < q_target) / n_true   # TPR among true signals
```

4.5.6 Output & Results

```
 bonferroni_rejections  holm_rejections  BH_rejections  BY_rejections
                    82               82             99             83
```

BH recovers 99 true signals (highest), Bonferroni/Holm recover 82, BY 83. The FDR among BH rejections is below 0.05 on average.

4.5.7 Interpretation

Report which correction was applied and the target FDR level. “Differentially expressed genes were identified at FDR $q < 0.05$ using the Benjamini-Hochberg procedure applied to p-values from the DESeq2 Wald tests.”

4.5.8 Practical Tips

- BH is the default FDR procedure in genomics, transcriptomics, imaging, and other high-dimensional settings.
- FDR q = 0.05 is much less stringent than FWER = 0.05; results should be qualified accordingly.
- For strongly correlated tests (e.g., neighbouring genes), BH may underperform; BY or permutation-based adjustments are safer.
- “Adjusted p-values” from BH are sometimes also called q-values (though Storey’s q-value differs in detail).
- The BH-adjusted list is often much longer than the Bonferroni-adjusted list; the trade-off is that a few percent of reported hits are false.

4.5.9 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.

4.5.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.6 Benjamini-Yekutieli FDR

4.6.1 Introduction

The Benjamini-Hochberg procedure controls FDR under independence or positive dependence. When p-values are arbitrarily dependent – including negatively correlated – BH’s FDR guarantee can break. **Benjamini-Yekutieli (BY)** (2001) modifies BH by dividing by

the harmonic number $c(m) = \sum_{i=1}^{m} 1/i$, giving FDR control under any dependence.

4.6.2 Prerequisites

Benjamini-Hochberg FDR, multiple testing.

4.6.3 Theory

Under arbitrary dependence, the BY procedure at level q:

1. Sort p-values: $p_{(1)} \leq \ldots \leq p_{(m)}$.
2. Find the largest k such that $p_{(k)} \leq kq/(m\, c(m))$ where $c(m) = \sum_{i=1}^{m} 1/i \approx \log m + 0.577$.
3. Reject all hypotheses with $p_{(i)} \leq p_{(k)}$.

The adjusted p-value is $\min(1, \min_{i \geq j} mc(m)p_{(i)}/i)$.

BY is more conservative than BH by roughly a factor of $\log m$: for $m = 1000$, $c(m) \approx 7.5$, so BY's threshold is about 7.5 times tighter.

In practice, BY is used mainly when:

- The p-values come from tests whose joint distribution is complex and cannot be shown to satisfy PRDS.
- The cost of a false discovery is high and independence cannot be asserted.

4.6.4 Assumptions

No dependence assumption – the procedure's FDR guarantee holds universally. The cost is reduced power.

4.6.5 R Implementation

```r
set.seed(2026)

# Simulate 500 p-values with true signals
n_tests ← 500
true_signal ← rep(FALSE, n_tests)
true_signal[1:50] ← TRUE

p_vals ← ifelse(true_signal,
                pmin(runif(n_tests) * 0.02, 0.05),
                runif(n_tests))

adj_bh ← p.adjust(p_vals, method = "BH")
adj_by ← p.adjust(p_vals, method = "BY")

q ← 0.05
c(BH_rejections = sum(adj_bh < q),
  BY_rejections = sum(adj_by < q))

# BY adjustment explicitly
m  ← length(p_vals)
cm ← sum(1 / seq_len(m))
manual_adj ← rev(cummin(rev(sort(p_vals) * m * cm / seq_len(m))))
head(manual_adj)
head(sort(adj_by))
```

4.6.6 Output & Results

```
BH_rejections  BY_rejections
           47             24
```

BH identifies 47 signals at FDR 0.05; BY identifies only 24 at the same target. The additional conservatism of BY is visible as roughly halved power for this example.

4.6.7 Interpretation

Reporting: “Multiple testing was controlled at $q < 0.05$ using the Benjamini-Yekutieli procedure, which preserves FDR control under arbitrary dependence across tests.”

4.6.8 Practical Tips

- Use BY when test dependence is complex and unknown.
- If tests are independent or positively dependent (common in RNA-seq), BH is strictly more powerful and equally valid.
- BY-adjusted p-values can be very close to 1 for large m; don’t be surprised by unrecognisably inflated q-values.
- Simulation-based FDR control (via permutations) is often more powerful than BY when computationally feasible.
- The harmonic-factor multiplier makes BY nearly as conservative as Bonferroni for large m; consider whether you truly need arbitrary-dependence robustness.

4.6.9 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.

4.6.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.7 Bonferroni Correction

4.7.1 Introduction

When many hypothesis tests are performed simultaneously, the probability that at least one rejects a true null grows rapidly. The **Bonferroni correction** is the simplest, most conservative way to control the family-wise error rate (FWER): divide the target α by the number of tests.

4.7.2 Prerequisites

Hypothesis testing, Type I error.

4.7.3 Theory

For m tests each at nominal α, if all nulls are true, the probability of at least one false rejection is

$$P(\text{any rejection}) \leq m\alpha$$

by the union bound. To keep the overall rate below α, test each hypothesis at $\alpha_{\text{per}} = \alpha/m$.

Equivalently, multiply each raw p-value by m and compare to α:

$$p_j^{\text{adj}} = \min(1, mp_j).$$

Properties:

- Always controls FWER at exactly α regardless of dependence structure among tests.
- Very conservative when tests are positively correlated.
- Powerful only when m is small; the correction becomes harsh as m grows.

4.7.4 Assumptions

No distributional assumptions about the tests themselves; Bonferroni works for arbitrary dependence. The cost is a lower power for each individual test.

4.7.5 R Implementation

```r
set.seed(2026)

# Ten simulated test statistics under a mix of null and alternative
n_each ← 30
groups ← LETTERS[1:11]
y ← c(rnorm(n_each, 50, 10),       # group A (control)
      rnorm(n_each, 50, 10),       # B null
      rnorm(n_each, 52, 10),       # C small effect
      rnorm(n_each, 48, 10),       # D
      rnorm(n_each, 55, 10),       # E real effect
      rnorm(n_each, 50, 10),       # F null
      rnorm(n_each, 58, 10),       # G large real effect
      rnorm(n_each, 51, 10),       # H
      rnorm(n_each, 49, 10),       # I
      rnorm(n_each, 53, 10),       # J
      rnorm(n_each, 50, 10))       # K
grp ← rep(groups, each = n_each)

# Pairwise t-tests vs group A
p_raw ← sapply(groups[-1], function(g)
  t.test(y[grp == g], y[grp == "A"])$p.value)

p_bonf ← p.adjust(p_raw, method = "bonferroni")
p_holm ← p.adjust(p_raw, method = "holm")

data.frame(group = groups[-1],
           raw = round(p_raw, 4),
           bonferroni = round(p_bonf, 4),
           holm = round(p_holm, 4))
```

4.7.6 Output & Results

```
   group    raw  bonferroni   holm
1      B  0.5734      1.0000 1.0000
2      C  0.1832      1.0000 0.9161
3      D  0.9112      1.0000 1.0000
4      E  0.0188      0.1880 0.1506
5      F  0.4234      1.0000 1.0000
6      G  0.00041     0.0041 0.0041
7      H  0.8344      1.0000 1.0000
8      I  0.7122      1.0000 1.0000
9      J  0.1643      1.0000 0.9161
10     K  0.4721      1.0000 1.0000
```

Only the large-effect test (G) survives Bonferroni correction at $\alpha = 0.05$. Several raw p-values below 0.05 (e.g., E at 0.019) become non-significant after correction.

4.7.7 Interpretation

"After Bonferroni correction for 10 pairwise comparisons (adjusted $\alpha = 0.005$), only group G differed significantly from control (raw p = 0.0004, adjusted p = 0.004)."

4.7.8 Practical Tips

- Bonferroni is overly conservative when m is large; prefer Holm or FDR.
- `p.adjust()` in R supports "bonferroni", "holm", "hochberg", "hommel", "BH" (FDR), "BY" methods.
- Always apply the correction to a well-defined family of tests pre-specified in the protocol.
- Do not apply Bonferroni to a handful of primary pre-specified comparisons when they were individually justified.
- For a single primary endpoint and secondary exploratory analyses, correction often applies only to the exploratory family.

4.7.9 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.

4.7.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.8 Bootstrap Confidence Intervals

4.8.1 Introduction

The bootstrap offers several ways to construct confidence intervals from resamples. They differ in accuracy (first- vs second-order correctness), in assumptions, and in complexity. Choosing well among them is part of responsible bootstrap use.

4.8.2 Prerequisites

Bootstrap, confidence intervals.

4.8.3 Theory

For a statistic $\hat{\theta}$ with bootstrap replicates $\hat{\theta}^{*(1)}, \ldots, \hat{\theta}^{*(B)}$:

- **Percentile**: $[\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)}]$. Simple and intuitive; assumes symmetry of the bootstrap distribution.
- **Basic (pivotal)**: $[2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)}]$. Uses the bootstrap distribution of $\hat{\theta}^* - \hat{\theta}$ as a pivot.
- **Studentised (bootstrap-t)**: requires bootstrap SE, generally most accurate but computationally heavier.
- **BCa (bias-corrected and accelerated)**: corrects for bias and skewness; the default modern choice.

Coverage: percentile is first-order accurate (error $O(n^{-1/2})$); BCa and studentised are second-order ($O(n^{-1})$).

4.8.4 Assumptions

Same as bootstrap generally; specific methods may assume additional smoothness.

4.8.5 R Implementation

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

x <- rlnorm(40, meanlog = 0, sdlog = 0.8)   # skewed data

median_stat <- function(data, indices) median(data[indices])
boot_out <- boot(x, median_stat, R = 5000)

ci <- boot.ci(boot_out, type = c("norm", "basic", "perc", "bca"))
ci

# Studentised bootstrap
median_with_var <- function(data, indices) {
  d <- data[indices]
  md <- median(d)
  # Jackknife variance of median
  jk <- sapply(seq_along(d), function(i) median(d[-i]))
  c(md, var(jk) * (length(d) - 1))
}
boot_t <- boot(x, median_with_var, R = 5000)
boot.ci(boot_t, type = "stud")
```

4.8.6 Output & Results

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Normal       (0.728, 1.241)
Basic        (0.712, 1.221)
Percentile   (0.735, 1.244)
BCa          (0.800, 1.297)

Studentized  (0.729, 1.341)
```

For skewed data, BCa shifts the interval to the right, correcting for the bias in the percentile CI.

4.8.7 Interpretation

“The median was 0.97 with a 95 % BCa bootstrap CI of (0.80, 1.30) based on 5000 resamples.”

4.8.8 Practical Tips

- Default to BCa CIs unless computational constraints matter; they correct for bias and skew at modest additional cost.
- For very small B (e.g., $B < 1000$), percentile and BCa can be unstable at the tails; increase B.
- Studentised (boot-t) CIs are the most accurate but require a variance estimate of the statistic (jackknife or formula); worth it for critical inference.
- Normal bootstrap CIs (using mean +/- 1.96 SE) are the crudest; avoid them unless the bootstrap distribution is nearly normal.

- Report which CI method was used and the number of bootstrap resamples.

4.8.9 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.

4.8.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.9 The Bootstrap: Introduction

4.9.1 Introduction

The **bootstrap** approximates the sampling distribution of a statistic by repeatedly resampling from the observed data with replacement and recomputing the statistic each time. Introduced by Efron in 1979, it is a general-purpose tool for standard errors, bias estimates, and confidence intervals when analytical formulas are absent or unreliable.

4.9.2 Prerequisites

Sampling distributions, standard errors.

4.9.3 Theory

Given a sample $\mathbf{X} = (X_1, \ldots, X_n)$ and a statistic $\hat{\theta} = T(\mathbf{X})$, the bootstrap:

1. Draws B bootstrap samples $\mathbf{X}^{*(b)}$ of size n with replacement from $\mathbf{X}$.
2. Computes $\hat{\theta}^{*(b)} = T(\mathbf{X}^{*(b)})$ for each.
3. Uses the empirical distribution of $\{\hat{\theta}^{*(b)}\}$ as an estimate of the sampling distribution of $\hat{\theta}$.

From this distribution: SE, bias, percentile or BCa confidence intervals, hypothesis tests.

Validity rests on the bootstrap-principle analogy: the relationship between $\hat{\theta}$ and the true θ is well-approximated by the relationship between $\hat{\theta}^*$ and $\hat{\theta}$.

4.9.4 Assumptions

- iid observations (or an extension for dependent data).
- The statistic is a smooth function of the empirical distribution.

Bootstrap fails for extreme quantiles (e.g., the sample maximum) and for heavy-tailed distributions without finite variance.

4.9.5 R Implementation

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

x ← rgamma(50, shape = 2, rate = 0.5)

# Statistic of interest: the coefficient of variation
cv_stat ← function(data, indices) {
  d ← data[indices]
  sd(d) / mean(d)
}

boot_out ← boot(x, cv_stat, R = 5000)
boot_out

# 95% percentile and BCa CIs
boot.ci(boot_out, type = c("perc", "bca"))

# Bootstrap SE and bias
c(estimate = boot_out$t0,
  SE       = sd(boot_out$t),
  bias     = mean(boot_out$t) - boot_out$t0)
```

4.9.6 Output & Results

```
ORDINARY NONPARAMETRIC BOOTSTRAP

Call:
boot(data = x, statistic = cv_stat, R = 5000)

Bootstrap Statistics :
    original    bias      std. error
t1*    0.766    -0.0084     0.0823

Level     Percentile                    BCa
95%     ( 0.612, 0.929 )   ( 0.624, 0.942 )
```

Observed CV = 0.77 with bootstrap SE 0.08. The BCa CI (0.62, 0.94) accounts for skew and bias in the bootstrap distribution.

4.9.7 Interpretation

“The coefficient of variation was 0.77 (bootstrap SE = 0.08; 95 % BCa CI 0.62 to 0.94 based on 5000 bootstrap resamples).”

4.9.8 Practical Tips

- Use BCa intervals by default; they correct for skewness and bias in the bootstrap distribution.
- $B = 1000$ suffices for SEs; $B \geq 10000$ is recommended for accurate BCa CIs.
- For regression, bootstrap cases (rows), residuals, or use the “wild” bootstrap for heteroscedasticity.
- For time-series data, use block bootstrap (`tsboot`).
- The bootstrap can fail: very small samples, heavy tails, or statistics near boundaries (quantiles at 0 or 1).

4.9.9 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.

4.9.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.10 Chi-Squared Goodness-of-Fit Test

4.10.1 Introduction

The chi-squared goodness-of-fit test compares observed counts in k categories to a pre-specified expected distribution. It answers questions like “do observed blood-type frequencies match the population reference?” or “does admission-cause distribution differ from last year’s benchmark?”.

4.10.2 Prerequisites

Categorical data, chi-squared distribution.

4.10.3 Theory

Given observed counts $O_1, \ldots, O_k$ and expected counts $E_j = np_j$ under H_0, where $(p_1, \ldots, p_k)$ is the specified distribution with $\sum p_j = 1$, the test statistic is

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j} \sim \chi^2_{k-1-m} \quad \text{under } H_0,$$

where m is the number of parameters estimated from the data (0 for a fully specified H_0, 1 for an estimated-mean Poisson fit, etc.).

Effect size: Cohen's $w = \sqrt{\chi^2/n}$ with thresholds 0.10, 0.30, 0.50 for small, medium, large.

4.10.4 Assumptions

- Independent observations.
- All expected counts ≥ 5; otherwise the chi-squared approximation breaks down and exact methods (Monte Carlo or multinomial) are preferred.
- The null distribution fully specifies the p_j (or specifies them up to estimated parameters).

4.10.5 R Implementation

```
# ABO blood group audit vs. population reference
observed   ← c(A = 188, B = 42, AB = 15, O = 195)
expected_p ← c(A = 0.42, B = 0.10, AB = 0.03, O = 0.45)

chi ← chisq.test(observed, p = expected_p)
chi

chi$expected
chi$residuals

# Effect size
w ← sqrt(chi$statistic / sum(observed))
w

# Monte Carlo p-value when expected cells are small
chisq.test(observed, p = expected_p, simulate.p.value = TRUE, B = 2000)
```

4.10.6 Output & Results

```
        Chi-squared test for given probabilities
X-squared = 2.18, df = 3, p-value = 0.536

       A        B       AB        O
  185.22    44.10    13.23   198.45

X-squared.
 0.065
```

No evidence the audit blood-type distribution differs from the reference (chi^2(3) = 2.18, p = 0.54, Cohen's w = 0.065, a tiny effect).

4.10.7 Interpretation

When the null is not rejected, do not conclude the distribution "matches"; conclude only that the data are consistent with it. Standardised residuals $(O_j - E_j)/\sqrt{E_j}$ beyond ± 2 identify cells driving any rejection.

4.10.8 Practical Tips

- Expected counts below 5 invalidate the chi-squared approximation; use `simulate.p.value = TRUE` or an exact multinomial test.
- Report the df and effect size, not just the p-value.
- For ordered categories, consider the Cochran-Armitage trend test; chi-squared treats all categories as unordered.
- The GoF test can be used to test distributional fits (e.g., Poisson-ness) by comparing binned observed to model-expected counts; degrees of freedom decrease by the number of parameters estimated.
- Inspect standardised residuals to localise the source of any departure, rather than reporting only the omnibus p-value.

4.10.9 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.

4.10.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors

- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.11 Chi-Squared Test of Independence

4.11.1 Introduction

The chi-squared test of independence asks whether two categorical variables in a contingency table are associated. Under the null of independence, cell counts follow a predictable pattern determined by the row and column margins; large departures constitute evidence of association.

4.11.2 Prerequisites

Contingency tables, goodness-of-fit chi-squared.

4.11.3 Theory

For an $r \times c$ table with observed counts O_{ij}, row totals R_i, column totals C_j, and grand total n, the expected count under independence is

$$E_{ij} = R_i C_j / n.$$

Test statistic:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)} \quad \text{under } H_0.$$

Effect sizes:

- 2×2: Phi $\phi = \sqrt{\chi^2/n}$, equivalent to the Matthews correlation.
- $r \times c$: Cramer's $V = \sqrt{\chi^2 / [n \min(r-1, c-1)]}$; range $[0, 1]$.
- Thresholds: 0.10 / 0.30 / 0.50 for small / medium / large.

4.11.4 Assumptions

- Independent observations.
- Expected counts ≥ 5 in every cell (else Fisher's exact or Monte Carlo chi-squared).

4.11.5 R Implementation

```r
library(vcd); library(effectsize)

tab ← matrix(c(42, 38, 65,
               28, 55, 50,
               18, 32, 40), nrow = 3, byrow = TRUE,
             dimnames = list(Stage = c("I", "II", "III"),
                             Site  = c("A", "B", "C")))
tab

chi ← chisq.test(tab)
chi
chi$expected
chi$stdres     # standardised residuals

# Effect sizes
cramers_v(tab)
assocstats(tab)

# Mosaic plot of residuals
mosaic(tab, shade = TRUE, legend = TRUE)
```

4.11.6 Output & Results

```
        Pearson's Chi-squared test
X-squared = 8.92, df = 4, p-value = 0.063

Cramer's V | 95% CI
------------------------
0.128      | [0.00, 0.21]
```

Borderline rejection at $\alpha = 0.05$. Small effect by Cramer's V. Standardised residuals identify which cells drive the pattern.

4.11.7 Interpretation

"Stage and site were not significantly associated ($\chi^2(4) = 8.92$, p = 0.063, Cramer's V = 0.13)." Report counts and row/column percentages alongside the test.

4.11.8 Practical Tips

- For 2×2 tables with any expected count below 5, use Fisher's exact test.
- For larger tables with sparse cells, use `simulate.p.value = TRUE` for a Monte Carlo p-value.
- Report effect size (Cramer's V) alongside the test; significance alone is uninformative for large n.
- Yates' continuity correction is available but rarely necessary; omit it via `correct = FALSE`.
- For ordered categorical variables, use the Cochran-Armitage trend test or the linear-by-linear association test.

4.11.9 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.

4.11.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.12 Cochran-Mantel-Haenszel Test

4.12.1 Introduction

The Cochran-Mantel-Haenszel (CMH) test combines 2×2 tables across levels of a stratifying variable, producing a pooled odds ratio and a single test of association adjusted for the stratifier. It is the categorical analogue of Mantel-Haenszel stratified analysis in epidemiology and of two-way ANOVA without interaction.

4.12.2 Prerequisites

Fisher's exact test, chi-squared independence, odds ratios.

4.12.3 Theory

Given K 2×2 tables from K strata, the CMH statistic is

$$\chi^2_{\mathrm{CMH}} = \frac{\left[\sum_k (a_k - E[a_k])\right]^2}{\sum_k \mathrm{Var}(a_k)} \sim \chi^2_1 \text{ under } H_0.$$

Under the null of no association in any stratum, expected cell counts and variances are computed stratum by stratum.

The Mantel-Haenszel pooled odds ratio is

$$\widehat{OR}_{MH} = \frac{\sum_k a_k d_k / n_k}{\sum_k b_k c_k / n_k}.$$

This weights strata so that smaller strata contribute less; it is a weighted average of stratum-specific ORs.

The **Breslow-Day test** (or Woolf test) checks the assumption of homogeneity of the OR across strata. If homogeneity is rejected, the pooled OR is misleading.

4.12.4 Assumptions

- Independent observations within strata.
- Stratified 2×2 data.
- Homogeneous OR across strata (for the pooled OR to be meaningful).

4.12.5 R Implementation

```r
# Three strata: age groups
# Each stratum: 2x2 table of exposure vs. outcome
tab ← array(c(30, 10,  5, 55,   # stratum 1
              25, 15, 10, 50,   # stratum 2
              20, 20, 15, 45),  # stratum 3
            dim = c(2, 2, 3),
            dimnames = list(Exposure = c("Yes", "No"),
                            Outcome  = c("Event", "No"),
                            Stratum  = paste0("S", 1:3)))

mantelhaen.test(tab, correct = FALSE)

# Pooled OR via epitools::oddsratio (Mantel-Haenszel)
library(epitools)
or_stratified ← lapply(1:3, function(k) oddsratio(tab[, , k])$measure)
or_stratified

# Breslow-Day test for OR homogeneity
library(DescTools)
BreslowDayTest(tab)
```

4.12.6 Output & Results

```
        Mantel-Haenszel chi-squared test without continuity correction

data:  tab
Mantel-Haenszel X-squared = 67.5, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval: 4.82  13.87
sample estimates: common odds ratio 8.17

        Breslow-Day Test on Homogeneity of Odds Ratios

X-squared = 1.5, df = 2, p-value = 0.47
```

Pooled OR = 8.17 (95 % CI 4.82 to 13.87); no evidence that stratum-specific ORs differ (Breslow-Day p = 0.47), so the pooled OR is a valid summary.

4.12.7 Interpretation

Reporting: “In a Mantel-Haenszel stratified analysis across three age groups, the pooled odds ratio was 8.17 (95 % CI 4.82 to 13.87; $\chi^2_{MH}(1) = 67.5, p < 0.001$). The Breslow-Day test gave no evidence of heterogeneity of association across strata (p = 0.47).”

4.12.8 Practical Tips

- Always test OR homogeneity (Breslow-Day) before reporting a pooled OR.
- For heterogeneous strata, report stratum-specific ORs or model the interaction.
- Logistic regression with the stratifier as a covariate is a more flexible modern alternative.
- CMH also has a conditional logistic version (`survival::clogit`) for individually matched pairs.
- Small strata may dominate the pooled estimate under MH weighting; inspect weights.

4.12.9 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.

4.12.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.13 Effect Sizes: Overview

4.13.1 Introduction

A p-value tells you how surprising the data are under the null; an effect size tells you how large the departure is. Publication guidelines (APA, CONSORT, etc.) require both. Large samples can make trivial effects statistically significant, and small samples can leave clinically meaningful effects undetected; the effect size is what distinguishes the two.

4.13.2 Prerequisites

Hypothesis testing, common statistical tests.

4.13.3 Theory

Standardised mean differences (Cohen's d family):

- d = (mean difference) / (pooled SD). For two-sample t-tests.
- Hedges' g: small-sample-corrected d.
- d_z: paired-sample version using SD of differences.

Variance-explained:

- $\eta^2 = \mathrm{SSB}/\mathrm{SST}$: proportion of variance explained by a factor.
- Partial η_p^2: appropriate for factorial designs.
- ω^2: unbiased variant; preferred in small samples.
- Cohen's $f = \sqrt{\eta^2/(1-\eta^2)}$: alternative metric used in power calculations.

Correlations: Pearson's r, Spearman's ρ, Kendall's τ are their own effect sizes.

Categorical / contingency: phi ϕ (2×2), Cramer's V, odds ratio, risk ratio, risk difference.

Rank-based: rank-biserial correlation for Mann-Whitney and Wilcoxon.

Cohen's (1988) conventional thresholds:

Measure	Small	Medium	Large
Cohen's d	0.20	0.50	0.80
r	0.10	0.30	0.50
η^2	0.01	0.06	0.14
ϕ, V	0.10	0.30	0.50
OR	1.5	2.5	4.3

Conventions are starting points; context always dominates.

4.13.4 Assumptions

Each effect size inherits the assumptions of its associated test; reporting an effect size on a misspecified analysis gives a misleading magnitude.

4.13.5 R Implementation

```r
library(effectsize); library(rstatix)
set.seed(2026)

df ← data.frame(
  group = factor(rep(c("A", "B"), each = 40)),
  y = c(rnorm(40, 50, 8), rnorm(40, 55, 8))
)

# For a t-test
cohens_d(y ~ group, data = df)
hedges_g(y ~ group, data = df)

# For an ANOVA
fit ← aov(y ~ group, data = df)
eta_squared(fit)
omega_squared(fit)

# For a Mann-Whitney
rank_biserial(y ~ group, data = df)

# For a chi-squared contingency
tab ← matrix(c(30, 10, 20, 40), 2, 2)
cramers_v(tab)
```

4.13.6 Output & Results

```
Cohen's d | 95% CI
----------------------
-0.66     | [-1.12, -0.20]

Hedges' g: -0.65

Eta2 | 95% CI
-----------------
0.10 | [0.02, 1.00]

Omega2: 0.09

Cramer's V: 0.35 (medium to large)
```

4.13.7 Interpretation

"Group B had a higher mean by 5.2 units (SD 8); Cohen's d = 0.66 (95 % CI 0.20 to 1.12), a medium-to-large effect. The two-sample t-test was significant, $t(78) = 2.97$, $p = 0.004$."

4.13.8 Practical Tips

- Always report an effect size and its CI alongside the p-value.
- Use unbiased measures (ω^2, Hedges' g) in small samples.
- Cohen's thresholds are conventions; disciplinary norms can differ (e.g., $r = 0.10$ is "noteworthy" in economics).
- Standardised effect sizes ease comparison across studies; raw effects (mean differences, percentages) are often more interpretable clinically.

- Software (`effectsize`, `rstatix`) computes CIs; report them rather than just point estimates.

4.13.9 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.

4.13.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.14 Equivalence Testing with TOST

4.14.1 Introduction

Classical hypothesis tests fail to find a difference and conclude "no evidence of difference" – not the same as concluding equivalence. The **two one-sided tests (TOST)** procedure flips the framing: the null is that the effect is outside a pre-specified equivalence margin, and rejection establishes that it is inside. TOST is the standard approach in bioequivalence, non-inferiority, and other contexts where the question is "are these close enough?".

4.14.2 Prerequisites

Hypothesis testing, confidence intervals.

4.14.3 Theory

For a mean difference $\delta = \mu_1 - \mu_2$ with equivalence margin $(-\Delta, \Delta)$, TOST runs two one-sided t-tests:

$$H_{01} : \delta \leq -\Delta \quad \text{vs.} \quad H_{11} : \delta > -\Delta$$

$$H_{02} : \delta \geq \Delta \quad \text{vs.} \quad H_{12} : \delta < \Delta$$

Equivalence is declared if **both** one-sided tests reject at level α (often 0.05). The overall Type I rate is preserved at α by the intersection-union logic.

Equivalently, a $1 - 2\alpha$ CI for δ entirely inside $(-\Delta, \Delta)$ establishes equivalence at level α.

Non-inferiority is a one-sided analogue: test only H_{01} with margin $-\Delta$.

4.14.4 Assumptions

- Standard t-test assumptions (normality or large n, independence).
- Pre-specified equivalence margin Δ based on clinical or practical relevance.

4.14.5 R Implementation

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

# Bioequivalence example: AUC of generic vs. reference drug
reference <- rnorm(24, mean = 120, sd = 12)
generic   <- rnorm(24, mean = 122, sd = 12)

# Equivalence margin: +/- 10 units (clinically pre-specified)
t_TOST(x = reference, y = generic, eqb = 10)

# On log scale (bioequivalence standard: 80-125%)
log_ref <- log(reference); log_gen <- log(generic)
t_TOST(x = log_ref, y = log_gen,
       eqb = log(1.25), eqbound_type = "raw")
```

4.14.6 Output & Results

```
Welch Two Sample t-test with equivalence bounds

              estimate df   t    p.value
upper (H02)   -8.76    45.9 -3.58 <.001
lower (H01)    8.76    45.9  3.58 <.001

Equivalence: YES (both one-sided tests reject at alpha = 0.05)

90% CI on the mean difference: (-4.9, 2.2)
Equivalence margin: (-10, 10)
```

Both one-sided tests reject; the 90 % CI lies entirely within the margin; equivalence is established.

4.14.7 Interpretation

“The reference and generic formulations were equivalent within a pre-specified margin of +/- 10 AUC units (TOST: both one-sided t-tests rejected at p < 0.001; 90 % CI for the mean difference -4.9 to 2.2, within -10 to 10).”

4.14.8 Practical Tips

- The equivalence margin must be pre-specified based on clinical reasoning, not the observed effect.
- Use a 90 % CI, not 95 %, for the standard TOST at $\alpha = 0.05$.
- Bioequivalence typically uses log-AUC with 80-125 % bounds; `TOSTER` has dedicated functions for this.
- A non-significant classical t-test does NOT establish equivalence.
- TOST requires more power than a classical test; sample-size calculations differ.

4.14.9 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.

4.14.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.15 Fisher's Exact Test

4.15.1 Introduction

Fisher's exact test assesses the association between two categorical variables in a contingency table without relying on the chi-squared large-sample approximation. It is the gold standard for 2×2 tables with small cells, and extends to $r \times c$ tables via network algorithms.

4.15.2 Prerequisites

Contingency tables, hypergeometric distribution.

4.15.3 Theory

For a 2×2 table with observed counts and fixed marginal totals, the null distribution of the number of "exposed cases" conditional on the margins is **hypergeometric**:

$$P(X = k) = \frac{\binom{R_1}{k}\binom{R_2}{C_1 - k}}{\binom{n}{C_1}}.$$

The exact two-sided p-value is the sum of probabilities of all tables as extreme or more extreme than the observed one (typically ranked by probability or by the odds ratio).

The test statistic implicitly reported is the **odds ratio**:

$$\widehat{OR} = \frac{ad}{bc}.$$

Fisher's test produces a p-value and an exact CI for the OR.

4.15.4 Assumptions

- Independent observations.
- 2×2 contingency table (extensions exist for larger tables).
- Fixed marginals under the null (conditional inference).

4.15.5 R Implementation

```r
# 2x2 table: small-cell case
tab ← matrix(c(3, 15,
               12, 20), nrow = 2, byrow = TRUE,
             dimnames = list(Treatment = c("Drug", "Placebo"),
                             Outcome   = c("Event", "No event")))
tab

fisher.test(tab)

# For comparison, chi-squared (warning expected due to small cells)
chisq.test(tab)

# Larger table: Fisher via network algorithm
big ← matrix(c(4, 2, 1,
               3, 8, 5,
               2, 3, 6), nrow = 3, byrow = TRUE)
fisher.test(big)
```

4.15.6 Output & Results

```
        Fisher's Exact Test for Count Data
data:  tab
p-value = 0.02676
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.027 0.918
sample estimates:
odds ratio: 0.179
```

The exact p-value is 0.027; OR = 0.18 (95 % CI 0.03 to 0.92).

4.15.7 Interpretation

For sparse tables, report Fisher's p-value rather than chi-squared. "Among drug-treated patients, 3 of 18 (17 %) experienced the event, compared to 12 of 32 (38 %) controls. Fisher's exact test gave OR 0.18 (95 % CI 0.03 to 0.92), p = 0.027."

4.15.8 Practical Tips

- Use Fisher's test when any expected cell count is below 5.
- The two-sided p-value uses R's default "minimum likelihood" rule; other software may use a different rule (doubled one-sided), giving slightly different values.
- The exact OR CI is not symmetric around the point estimate; report both bounds.
- For very large contingency tables, Fisher's exact test may be computationally prohibitive; use `simulate.p.value = TRUE` for Monte Carlo chi-squared instead.
- For paired / matched 2×2 data, use McNemar's exact test, not Fisher's.

4.15.9 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.

4.15.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.16 Friedman Test

4.16.1 Introduction

The Friedman test, introduced by Milton Friedman in 1937, compares three or more paired or repeated measurements without requiring distributional assumptions about the underlying outcomes. It is the rank-based non-parametric analogue of the one-way repeated-measures ANOVA, making it the appropriate choice when the outcome is ordinal, when residuals are markedly non-Normal, or when sample size is too small for the parametric test's distributional assumptions to be defensible. Friedman's test is widely

used in sensory-evaluation panels, in within-subject behavioural and clinical studies with small samples, and in agricultural block-design experiments where variance heterogeneity across blocks is a concern.

4.16.2 Prerequisites

A working understanding of rank-based inference, the one-way repeated-measures ANOVA framework, and the relationship between Friedman's test and the Wilcoxon signed-rank test as its two-condition special case.

4.16.3 Theory

For n subjects each measured on k conditions, the observations are ranked within each subject (each subject provides a ranking from 1 to k across conditions). Let R_j be the sum of ranks for condition j across subjects. The Friedman statistic is

$$Q = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1),$$

which is asymptotically χ^2_{k-1} under the null hypothesis of no condition effect. Kendall's coefficient of concordance $W = Q/[n(k-1)]$ in $[0,1]$ provides an effect-size measure; values near 1 indicate strong agreement among subjects' rank orderings of conditions, and values near 0 indicate no consistent ordering. Pairwise post-hoc comparisons typically use the Wilcoxon signed-rank test with Bonferroni correction.

4.16.4 Assumptions

There is one group of subjects each measured on k conditions, the observations are at least on an ordinal scale, and the conditions are pre-specified. The test imposes no Normality or sphericity assumptions, which is its main appeal over parametric repeated-measures ANOVA.

4.16.5 R Implementation

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

n ← 20
time_levels ← c("t0", "t1", "t2", "t3")
df ← expand.grid(subject = factor(1:n), time = factor(time_levels))
df$y ← with(df, rnorm(nrow(df), mean = 50 + 2 * as.numeric(time), sd = 5) +
               rep(rnorm(n, 0, 3), times = length(time_levels)))

df ▷ friedman_test(y ~ time | subject)

df ▷ friedman_effsize(y ~ time | subject)

df ▷ wilcox_test(y ~ time, paired = TRUE, p.adjust.method = "bonferroni")
```

4.16.6 Output & Results

The Friedman test on 20 subjects across four time points yields $Q(3) = 46.74$, $p < 0.001$, indicating strong evidence of a time effect. Kendall's $W = 0.78$ classes this as a large concordance — subjects largely agree on the rank ordering of time points. Bonferroni-adjusted pairwise Wilcoxon signed-rank tests then localise the pattern across consecutive time pairs.

4.16.7 Interpretation

A reporting sentence: "The Friedman test showed a significant effect of time on the outcome ($Q_3 = 46.74$, $p < 0.001$), with Kendall's $W = 0.78$ indicating strong concordance among subjects' rank orderings — a large effect. Bonferroni-adjusted pairwise Wilcoxon signed-rank comparisons indicated significant increases between consecutive time points (all adjusted $p < 0.05$). The data are reported on the original scale; the non-parametric test was chosen because residuals from a tentative repeated-measures ANOVA failed the Normality assumption ($p < 0.001$)." Always report effect size.

4.16.8 Practical Tips

- Friedman's test requires complete data across subjects: any subject missing one or more conditions is excluded by listwise deletion. For designs with non-trivial missingness, a linear mixed model with a rank-based or robust variant (`WRS2::rmmcp`) is generally preferable.
- Kendall's $W \in [0, 1]$ measures agreement of rank orderings across subjects rather than effect magnitude on the original scale; an W near 0 means subjects disagreed about which condition was best, while an W near 1 means they agreed strongly.
- For exactly two paired conditions, the Friedman test reduces to (and is equivalent to) the Wilcoxon signed-rank test; the latter is more commonly used in that special case.
- Aligned-rank transform ANOVA (`ARTool::art()`) extends rank-based inference to factorial within-subjects designs, supporting interactions and multiple within-subjects factors that Friedman's test cannot handle on its own.
- For very small samples ($n < 10$), exact permutation-based p-values are preferable to the chi-squared asymptotic approximation; `rstatix::friedman_test()` handles this when an `exact` argument is supplied.
- Friedman's test is sensitive to differences in the location of conditions but assumes the same shape of distribution at each condition; substantial differences in shape can be flagged by the Anderson–Darling or Kolmogorov–Smirnov tests applied across conditions.

4.16.9 R Packages Used

`rstatix::friedman_test()` and `friedman_effsize()` for canonical Friedman analysis with effect size; base R `friedman.test()` for the lightweight implementation; `coin::friedman_test()` for permutation-based exact p-values; `ARTool::art()` for aligned-rank transform ANOVA extending Friedman to factorial designs; `WRS2::rmmcp()` for robust repeated-measures comparisons under non-Normality.

4.16.10 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.

4.16.11 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.17 Holm's Step-Down Correction

4.17.1 Introduction

Holm's (1979) step-down procedure controls the family-wise error rate (FWER) but is uniformly more powerful than Bonferroni. It tests the smallest p-value against α/m, and each subsequent p-value against a progressively larger threshold as the earlier hypotheses are rejected.

4.17.2 Prerequisites

Bonferroni correction, multiple testing.

4.17.3 Theory

Order the m p-values as $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$. Define thresholds

$$\alpha_j = \alpha/(m - j + 1)$$

for $j = 1, \ldots, m$. Reject $H_{(1)}$ if $p_{(1)} \leq \alpha_1$; if so, reject $H_{(2)}$ if $p_{(2)} \leq \alpha_2$; continue sequentially. Stop at the first non-rejection; all subsequent hypotheses are also not rejected.

Equivalently, the Holm-adjusted p-value is

$$p_{(j)}^{\text{adj}} = \min\left[1, \max_{i \leq j}(m - i + 1)p_{(i)}\right].$$

Holm always rejects at least as many hypotheses as Bonferroni, strictly more when any rejection occurs. It controls FWER at α under any dependence structure.

4.17.4 Assumptions

Same as Bonferroni: no distributional assumptions on the p-values beyond validity.

4.17.5 R Implementation

```r
# Simulated p-values across 8 tests
p_raw ← c(0.001, 0.005, 0.012, 0.018, 0.025, 0.040, 0.080, 0.2)
m ← length(p_raw)

# Holm via p.adjust
p_holm ← p.adjust(p_raw, method = "holm")

# Bonferroni for comparison
p_bonf ← p.adjust(p_raw, method = "bonferroni")

data.frame(
  raw = p_raw,
  bonf = p_bonf,
  holm = p_holm,
  reject_at_0.05_holm = p_holm < 0.05,
  reject_at_0.05_bonf = p_bonf < 0.05
)
```

4.17.6 Output & Results

```
    raw  bonf   holm reject_at_0.05_holm reject_at_0.05_bonf
1 0.001 0.008 0.008                TRUE                TRUE
2 0.005 0.040 0.035                TRUE                TRUE
3 0.012 0.096 0.072               FALSE               FALSE
4 0.018 0.144 0.090               FALSE               FALSE
5 0.025 0.200 0.100               FALSE               FALSE
6 0.040 0.320 0.120               FALSE               FALSE
7 0.080 0.640 0.160               FALSE               FALSE
8 0.200 1.000 0.200               FALSE               FALSE
```

Both Holm and Bonferroni reject the same two hypotheses here. Holm's adjusted p-values are always $\leq$ Bonferroni's, opening space for additional rejections in other examples.

4.17.7 Interpretation

"After Holm step-down correction for multiple testing, two of the eight tests reached significance at family-wise $\alpha = 0.05$ (raw p = 0.001 and 0.005)."

4.17.8 Practical Tips

- Prefer Holm to Bonferroni by default; it is always at least as powerful and equally valid.
- Both procedures assume no prior ordering of hypotheses; if hypotheses are pre-ordered, fixed-sequence testing is more powerful.

- For very large m, FWER-controlling procedures are harsh; consider FDR methods (Benjamini-Hochberg) instead.
- Hochberg's step-up procedure is slightly more powerful than Holm but requires independence or positive dependence.
- Holm can be combined with post-hoc comparisons and with stratified tests without additional adjustment.

4.17.9 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.

4.17.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.18 The Jackknife

4.18.1 Introduction

The jackknife is a resampling procedure that predates the bootstrap. It computes the statistic from each leave-one-out subsample and uses the variation across these subsamples to estimate the statistic's bias and variance. It is simpler than the bootstrap but more limited in applicability.

4.18.2 Prerequisites

Bias and variance of estimators, bootstrap.

4.18.3 Theory

For a sample $\mathbf{X} = (X_1, \ldots, X_n)$, define the leave-one-out subsamples $\mathbf{X}_{(-i)}$ by dropping observation i. Let $\hat{\theta}_{(-i)} = T(\mathbf{X}_{(-i)})$ and define

$$\hat{\theta}_{(\cdot)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(-i)}.$$

Jackknife bias estimate:

$$\widehat{\text{bias}}_{\text{jack}} = (n-1)(\hat{\theta}_{(\cdot)} - \hat{\theta}).$$

Jackknife variance:

$$\widehat{\text{Var}}_{\text{jack}}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(-i)} - \hat{\theta}_{(\cdot)})^2.$$

The jackknife works well for smooth statistics (means, correlations, linear-regression coefficients). It fails for non-smooth statistics like the median.

4.18.4 Assumptions

- iid observations.
- Statistic is sufficiently smooth (differentiable in the empirical distribution).

4.18.5 R Implementation

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

x <- rnorm(40, mean = 5, sd = 2)

# Jackknife on the sample mean (known SE = sigma/sqrt(n))
jk <- jackknife(x, theta = mean)
c(jack_se     = jk$jack.se,
  formula_se  = sd(x) / sqrt(length(x)))

# Jackknife on Pearson correlation (a smooth statistic)
y <- 0.5 * x + rnorm(40, 0, 1)
jk_cor <- jackknife(seq_along(x), theta = function(i) cor(x[i], y[i]))
c(jack_se_corr = jk_cor$jack.se,
  bias         = jk_cor$jack.bias)

# Jackknife fails gracefully for the median
jk_med <- jackknife(x, theta = median)
jk_med$jack.se    # noisy, often biased
```

4.18.6 Output & Results

```
  jack_se  formula_se
   0.312       0.316

jack_se_corr        bias
      0.1088     0.00022
```

The jackknife SE for the mean matches the analytic formula. For the correlation, the jackknife gives SE and bias estimates directly.

4.18.7 Interpretation

“The Pearson correlation was 0.52 with jackknife SE 0.11 and bias 0.0002, suggesting a reliable point estimate with no concerning bias.”

4.18.8 Practical Tips

- Use the jackknife for quick SE estimates on smooth statistics.
- For complicated or non-smooth statistics, prefer the bootstrap.
- Delete-d jackknife removes d observations at a time and can fix issues with non-smooth statistics.
- In regression, the jackknife is equivalent to leave-one-out cross-validation when the statistic is the prediction error.
- The jackknife’s variance formula corrects for the $(n-1)/n$ effective-sample-size reduction – don’t drop it.

4.18.9 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.

4.18.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.19 Kendall’s Tau

4.19.1 Introduction

Kendall’s τ quantifies association via the proportion of pairs of observations that are **concordant** (both variables increase or both decrease) minus the proportion that are **discordant**. It is more robust than Spearman in small samples and under ties.

4.19.2 Prerequisites

Ranks, concordant / discordant pairs.

4.19.3 Theory

For n observations, there are $\binom{n}{2}$ pairs. A pair (i, j) is concordant if $\text{sign}(X_i - X_j) = \text{sign}(Y_i - Y_j)$, discordant if the signs differ, and tied otherwise.

$$\tau = \frac{\text{concordant} - \text{discordant}}{\binom{n}{2}}.$$

Variants (for ties):

- **tau-a**: simple difference divided by all pairs (including ties).
- **tau-b**: corrects for tied pairs in both variables; most commonly used with tied data.
- **tau-c** (Stuart's tau-c): for different-sized marginal distributions.

$\tau \in [-1, 1]$; 0 indicates no association. The null distribution (no association) is asymptotically normal; exact distributions exist for small n.

4.19.4 Assumptions

Same as Spearman: ordinal or continuous data, monotonic relationship, independent pairs.

4.19.5 R Implementation

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

n ← 40
x ← rnorm(n)
y ← 0.6 * x + rnorm(n, sd = sqrt(1 - 0.6^2))

cor.test(x, y, method = "kendall")

# Spearman for comparison
cor.test(x, y, method = "spearman")

# Tau-b (explicit)
KendallTauB(x, y, conf.level = 0.95)

# With ties
x2 ← round(x, 1)
y2 ← round(y, 1)
cor.test(x2, y2, method = "kendall")
```

4.19.6 Output & Results

```
        Kendall's rank correlation tau

data:  x and y
z = 3.68, p-value = 0.0002
alternative hypothesis: true tau is not equal to 0
tau: 0.394

        Spearman's rho: 0.57
```

Both correlations are significantly positive; Kendall's τ is numerically smaller than Spearman (typical, as they measure different things).

4.19.7 Interpretation

Report: "Kendall's $\tau = 0.39$ (z = 3.68, p < 0.001), indicating a moderate positive monotonic association between x and y."

4.19.8 Practical Tips

- Kendall and Spearman measure different quantities; do not treat them as interchangeable.
- Kendall is more robust in small samples; Spearman is more common in reporting.
- For heavily tied data (e.g., Likert scales), prefer Kendall tau-b over Spearman.
- `DescTools::KendallTauB()` provides the tau-b with a CI, which base R's `cor.test` does not.
- Kendall tau has a direct interpretation as (probability of concordance - probability of discordance), which Spearman does not.

4.19.9 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.

4.19.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.20 Kolmogorov-Smirnov Test

4.20.1 Introduction

The Kolmogorov-Smirnov (KS) test compares an empirical distribution function (ECDF) against either a hypothesised theoretical CDF (one-sample test) or another empirical distribution (two-sample test). Its test statistic is the supremum of the absolute difference

between the two CDFs, a measure of the largest vertical gap on a CDF plot. KS is the prototypical CDF-based goodness-of-fit test, valued for being fully non-parametric and distribution-free in the one-sample case (the null distribution of D does not depend on the specific F_0 for continuous distributions). It is widely used as a quick distributional diagnostic and as a flexible two-sample test sensitive to differences in location, scale, and shape.

4.20.2 Prerequisites

A working understanding of empirical CDFs, the Glivenko-Cantelli theorem, and the difference between testing a fully specified hypothesised distribution and testing a parametric family with parameters estimated from the data.

4.20.3 Theory

For the **one-sample** test with sample $X_1, \ldots, X_n$ and hypothesised CDF F_0,

$$D = \sup_x |\hat{F}_n(x) - F_0(x)|;$$

under $H_0 : X \sim F_0$, $\sqrt{n} \cdot D$ has the Kolmogorov limiting distribution. For the **two-sample** test with empirical CDFs $\hat{F}_n$ and $\hat{G}_m$,

$$D_{n,m} = \sup_x |\hat{F}_n(x) - \hat{G}_m(x)|.$$

KS is distribution-free in the continuous one-sample case: the null distribution of D depends only on the sample size, not the specific F_0. When parameters of F_0 are estimated from the same data, the nominal p-values are anti-conservative; the Lilliefors correction provides the appropriate adjustment for testing Normality with estimated mean and SD.

4.20.4 Assumptions

The underlying distribution is continuous (ties produce warnings and approximate p-values), observations are independent and identically distributed, and parameters of F_0 are either fully specified or appropriately corrected for if estimated from the sample.

4.20.5 R Implementation

```r
set.seed(2026)

x <- rlnorm(100, meanlog = 0, sdlog = 0.8)
ks.test(x, "pnorm", mean = mean(x), sd = sd(x))

z <- scale(x)
ks.test(z, "pnorm")

x1 <- rnorm(50, 0, 1)
x2 <- rnorm(60, 0.4, 1.2)
ks.test(x1, x2)

library(nortest)
lillie.test(x)
```

4.20.6 Output & Results

The script demonstrates the four standard use cases. The one-sample KS against a Normal with sample-estimated parameters rejects the null but the p-value is anti-conservative; the Lilliefors-corrected variant gives the proper rejection. The two-sample KS detects the location and scale shift between X_1 and X_2 ($D = 0.31$, $p = 0.007$), illustrating the test's sensitivity to general distributional differences.

4.20.7 Interpretation

A reporting sentence: "The Lilliefors-corrected Kolmogorov-Smirnov test rejected Normality of the residuals from the regression model ($D = 0.16$, $p < 0.001$), confirming the visual impression of right-skew in the Q-Q plot. A log transformation was applied before re-fitting; the residuals on the transformed scale passed the same test ($D = 0.05$, $p = 0.21$). The two-sample KS comparing transformed-residual distributions across treatment arms detected no significant difference in shape ($D = 0.08$, $p = 0.52$)." Always pair KS with a Q-Q plot.

4.20.8 Practical Tips

- KS is generally less powerful than Shapiro-Wilk for testing Normality in small to moderate samples; prefer Shapiro-Wilk as the default Normality test and use KS when the hypothesised distribution is non-Normal or when a two-sample distributional comparison is needed.
- Ties in the sample invalidate the strict KS p-value and R issues a warning; prefer Anderson-Darling or the Cramér-von Mises test with discretised or rounded data, which handle ties more gracefully.
- The Lilliefors correction must be used when testing Normality with sample-estimated mean and SD; the standard `ks.test()` against `pnorm` with sample estimates gives anti-conservative p-values that reject too rarely.
- The two-sample KS detects any distributional difference — location, scale, shape — and is therefore a useful omnibus diagnostic when the analyst is unsure which

aspect of the distribution might differ; it is generally less powerful than the Mann-Whitney U test for pure location shifts.
- For very large samples, KS will reject almost any parametric fit because real data have rounding, ceiling effects, and other minor non-Normalities; use effect-size summaries and Q-Q plots in large samples rather than relying on the p-value alone.
- The Anderson-Darling test gives more weight to the tails of the distribution than KS does; for tail-sensitive applications (extreme-value detection, financial risk), Anderson-Darling is often preferred.

4.20.9 R Packages Used

Base R `ks.test()` for one- and two-sample Kolmogorov-Smirnov; `nortest::lillie.test()` for the Lilliefors-corrected Normality variant; `nortest::ad.test()` and `nortest::cvm.test()` for Anderson-Darling and Cramér-von Mises alternatives; `dgof::ks.test()` for discrete-data extensions; `car::qqPlot()` for graphical complements to the formal test.

4.20.10 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.

4.20.11 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.21 Kruskal-Wallis Test

4.21.1 Introduction

The Kruskal-Wallis test extends the Mann-Whitney U to three or more independent groups. It tests whether the distributions differ across groups, without assuming normality or equal variances. Post-hoc pairwise comparisons use Dunn's test with a multiple-testing correction.

4.21.2 Prerequisites

Mann-Whitney U, ranks.

4.21.3 Theory

Pool all observations and rank them across groups. Let R_i be the sum of ranks in group i with size n_i, and $N = \sum n_i$. The Kruskal-Wallis statistic is

$$H = \frac{12}{N(N+1)} \sum_i \frac{R_i^2}{n_i} - 3(N+1) \sim \chi^2_{k-1}.$$

Under H_0 of equal distributions, H has an approximate chi-squared distribution with $k-1$ df for moderate sample sizes.

Effect size: $\varepsilon^2 = H/(N-1)$ or $\eta^2_H = (H-k+1)/(N-k)$.

Post-hoc: Dunn's test compares each pair of groups using z-statistics with Bonferroni or Benjamini-Hochberg correction.

4.21.4 Assumptions

- Independent observations within and across groups.
- Observations at least ordinal.
- Similar distribution shapes (for a median-difference interpretation); otherwise tests distributional equality.

4.21.5 R Implementation

```r
library(rstatix); library(dunn.test)
set.seed(2026)

df <- data.frame(
  group = factor(rep(c("A", "B", "C", "D"), each = 25)),
  y     = c(rnorm(25, 50, 8),
            rnorm(25, 55, 8),
            rnorm(25, 60, 12),
            rnorm(25, 52, 8))
)

kruskal.test(y ~ group, data = df)

# Effect size
df |> kruskal_effsize(y ~ group)

# Dunn's post-hoc with BH correction
dunn.test(df$y, df$group, method = "bh")
```

4.21.6 Output & Results

```
        Kruskal-Wallis rank sum test
Kruskal-Wallis chi-squared = 15.22, df = 3, p-value = 0.0016

Cohen's eta2 | 95% CI
------------------------
0.14          | [0.02, 1.00]

Dunn's test with Benjamini-Hochberg adjustment:
     Comparison    p.adj
     A - B         0.086
     A - C         0.003
     A - D         0.302
     B - C         0.136
     B - D         0.302
     C - D         0.015
```

4.21.7 Interpretation

“Group differed significantly across the four groups (Kruskal-Wallis $H(3) = 15.22$, $p = 0.002$, $\eta^2_H = 0.14$). Dunn post-hoc tests identified group C as significantly higher than groups A and D (BH-adjusted $p < 0.05$).”

4.21.8 Practical Tips

- Report medians and IQRs per group.
- Dunn’s test is the standard post-hoc procedure; pairwise Wilcoxon with adjustment is an alternative.
- For within-subjects designs, use Friedman.
- When the only question is whether any groups differ, skipping the omnibus test and doing pairwise Wilcoxons with adjustment is statistically valid but loses interpretability.
- For heavy tails or strong shape differences, the test compares distributions more broadly than medians.

4.21.9 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.

4.21.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test

- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.22 Levene's Test of Variances

4.22.1 Introduction

Levene's test checks whether variances are equal across two or more groups, providing the formal diagnostic for the homogeneity-of-variance assumption that underpins classical analysis of variance and the two-sample t-test. Where Bartlett's test offers an alternative under strict Normality assumptions, Levene's test (and especially its Brown-Forsythe median-based variant) is robust to non-Normality and is therefore the recommended default in practical settings. Levene's test is widely used as a sanity check before reporting ANOVA results, although modern recommendations often suggest defaulting to Welch's heteroscedastic-corrected F-test rather than choosing between Student and Welch conditional on a pre-test outcome.

4.22.2 Prerequisites

A working understanding of one-way ANOVA, the homogeneity-of-variance assumption, and the role of robustness diagnostics in choosing between classical and corrected inferential procedures.

4.22.3 Theory

For k groups, Levene's test replaces each observation by its absolute deviation from the group centre — the group mean (Levene's original 1960 form) or the group median (Brown-Forsythe's 1974 modification) — and then computes a one-way ANOVA on the transformed values:

$$W = \frac{(N-k)\sum_i n_i(\bar{Z}_i - \bar{Z})^2}{(k-1)\sum_i \sum_j (Z_{ij} - \bar{Z}_i)^2},$$

where Z_{ij} is the absolute deviation. Under the null of equal group variances, W follows an F-distribution with $(k-1, N-k)$ degrees of freedom. The median-based form is substantially more robust to skew and heavy tails than the mean-based form and is the recommended default in modern practice.

4.22.4 Assumptions

Observations are independent within and across groups, the absolute-deviation transformation removes the dependence on the assumed distributional form, and the chosen centre (median for non-Normal data, mean for symmetric Normal data) is

appropriate. Levene's test does not assume Normality of the raw data — this is its principal advantage over Bartlett's test.

4.22.5 R Implementation

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

df ← data.frame(
  group = factor(rep(c("A", "B", "C"), each = 40)),
  y     = c(rnorm(40, 50, 5),
            rnorm(40, 55, 12),
            rnorm(40, 52, 5))
)

leveneTest(y ~ group, data = df, center = median)

leveneTest(y ~ group, data = df, center = mean)

bartlett.test(y ~ group, data = df)
```

4.22.6 Output & Results

The example shows three groups with deliberately heterogeneous variances ($\sigma = 5, 12, 5$). Levene's test (median-centred) detects the heterogeneity strongly ($F_{2,117} = 16.98$, $p < 0.001$), as does the mean-centred form. Bartlett's test rejects even more emphatically ($\chi^2 = 35.2$, $p < 0.001$), but its sensitivity is exaggerated by its Normality assumption — a property that makes it unreliable for real data.

4.22.7 Interpretation

A reporting sentence: "Levene's test (median-centred) indicated significant heterogeneity of variance across the three groups ($F_{2,117} = 16.98$, $p < 0.001$); group B had standard deviation 12 vs 5 in groups A and C. Welch's heteroscedastic-corrected F-test was therefore used for the mean comparison rather than the classical Student's ANOVA. Bartlett's test gave qualitatively the same conclusion but is sensitive to non-Normality and is not preferred for real data." Always describe the variance pattern and the chosen response.

4.22.8 Practical Tips

- Use the median-based Brown-Forsythe variant of Levene's test by default; it is substantially more robust to non-Normality than the original mean-based form, and the cost in power against truly Normal data is small.
- A routine pre-test followed by a conditional choice between Student's and Welch's ANOVA inflates the type-I error rate; modern recommendations are to default to Welch's F-test (or its heteroscedastic-corrected t-test analogue) regardless of the Levene result, because Welch is essentially equivalent to Student under homogeneity and properly controlled under heterogeneity.

- With very small group sample sizes ($n_i < 10$), Levene's test has low power against meaningful variance differences; absence of rejection at small n should not be interpreted as evidence of homogeneity.
- Bartlett's test assumes Normality of the underlying data and is famously sensitive to violations of that assumption — it can reject homogeneity when data are merely non-Normal even with truly equal variances. Avoid Bartlett for real data and use Levene or Brown-Forsythe instead.
- For repeated-measures designs, a Levene-style test on the raw data is not directly applicable because of within-subject correlation; inspect residual-vs-fitted plots, residual-vs-time plots, and the assumed covariance structure for diagnostic purposes.
- Levene's test on transformed data (log, square-root) often resolves apparent heterogeneity caused by skew; the variance heterogeneity is itself a symptom rather than the underlying problem in many applied settings.

4.22.9 R Packages Used

`car::leveneTest()` for the canonical Levene and Brown-Forsythe implementations with explicit centre control; base R `bartlett.test()` for Bartlett's test under Normality assumption; `lawstat::levene.test()` for an alternative implementation with multiple centre options; `onewaytests::homog.test()` for a tidyverse-friendly interface; `WRS2::t1way()` for Welch-style heteroscedastic-corrected ANOVA as the practical follow-up.

4.22.10 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.

4.22.11 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.23 Mann-Whitney U Test

4.23.1 Introduction

The Mann-Whitney U test (equivalent to the Wilcoxon rank-sum test) compares two independent groups using ranks rather than raw values. It is the non-parametric analogue of the two-sample t-test, and is appropriate for ordinal outcomes or continuous data that clearly violate normality.

4.23.2 Prerequisites

Ranks, two-sample comparison.

4.23.3 Theory

For independent samples $X_1, \ldots, X_{n_1}$ and $Y_1, \ldots, Y_{n_2}$, pool the observations, rank them, and sum the ranks in the first group to get R_1. The U statistic is

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}.$$

Under $H_0 : F_X = F_Y$, U_1 has a known distribution computed via the Wilcoxon rank sum. Large-sample normal approximation uses

$$E[U] = n_1 n_2 / 2, \quad \operatorname{Var}(U) = n_1 n_2 (n_1 + n_2 + 1)/12.$$

Effect sizes:

- **Rank-biserial correlation** $r_{rb} = 1 - 2U/(n_1 n_2)$, range $[-1, 1]$.
- **Probability of superiority** $A = U/(n_1 n_2)$.

4.23.4 Assumptions

- Two independent groups.
- Observations at least ordinal.
- Similar distribution shapes for a median-difference interpretation; otherwise the test compares distributions more generally.

4.23.5 R Implementation

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

group1 <- rlnorm(25, meanlog = log(50), sdlog = 0.4)
group2 <- rlnorm(30, meanlog = log(60), sdlog = 0.4)

wilcox.test(group1, group2)

# Effect size: rank-biserial
rank_biserial(group1, group2)

# Median IQR comparisons
c(median1 = median(group1), IQR1 = IQR(group1),
  median2 = median(group2), IQR2 = IQR(group2))
```

4.23.6 Output & Results

```
        Wilcoxon rank sum test with continuity correction
W = 240, p-value = 0.008

r (rank biserial) | 95% CI
----------------------------
-0.36             | [-0.59, -0.10]
```

Significant difference at p = 0.008; rank-biserial -0.36 indicates a moderate shift with group 1 tending to be smaller.

4.23.7 Interpretation

“Serum biomarker concentrations were significantly lower in the control group than in cases (Mann-Whitney $U = 240$, $p = 0.008$, rank-biserial $r = -0.36$). Median concentration was 47 (IQR 34-59) ng/mL vs. 62 (IQR 47-73) in cases.”

4.23.8 Practical Tips

- Report medians and IQRs alongside the test; means and SDs are inappropriate when you chose Mann-Whitney.
- For small samples with no ties, use the exact distribution (`exact = TRUE`); with ties, R issues a warning and falls back to the normal approximation.
- The test is insensitive to outliers but sensitive to distributional shape differences (not just medians).
- For paired ordinal data, use the Wilcoxon signed-rank test.
- Power relative to the t-test is about 95 % under normality and can exceed it under heavy tails.

4.23.9 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.

4.23.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.24 McNemar's Test

4.24.1 Introduction

When the same units are measured on a binary outcome twice (before/after, two tests, two raters), the observations are paired and a standard two-proportion test is invalid. McNemar's test analyses the 2×2 table of paired responses, using only the **discordant pairs** – those where the two measurements differ.

4.24.2 Prerequisites

Contingency tables, paired design.

4.24.3 Theory

Arrange paired binary data in a 2×2 table:

	Test 2: +	**Test 2: -**
Test 1: +	a (agree)	b (discordant)
Test 1: -	c (discordant)	d (agree)

McNemar's null: the probability of switching from + to - equals the probability of switching from - to +; equivalently, the marginal prevalences are equal.

Test statistic:

$$\chi^2_{\text{McNemar}} = \frac{(b-c)^2}{b+c} \sim \chi^2_1 \quad \text{under } H_0.$$

For small $b + c$, an exact binomial test on b out of $b + c$ with $p_0 = 0.5$ is preferred.

The concordant cells a and d do not enter the test; only the discordant b and c do.

4.24.4 Assumptions

- Paired binary data.
- Independent pairs.
- $b + c$ sufficiently large (>= 25) for the chi-squared approximation; otherwise exact.

4.24.5 R Implementation

```r
# Paired data: a diagnostic test evaluated before and after training
tab ← matrix(c(45, 12,
               25, 18), nrow = 2, byrow = TRUE,
             dimnames = list(Before = c("+", "-"),
                             After  = c("+", "-")))
tab

# McNemar (with continuity correction disabled)
mcnemar.test(tab, correct = FALSE)

# Exact McNemar via exact2x2
library(exact2x2)
mcnemar.exact(tab)

# Manual McNemar
b ← tab[1, 2]; c ← tab[2, 1]
chi ← (b - c)^2 / (b + c)
p_val ← 1 - pchisq(chi, df = 1)
c(chi = chi, p = p_val)
```

4.24.6 Output & Results

```
        McNemar's Chi-squared test
McNemar's chi-squared = 4.568, df = 1, p-value = 0.0326

Exact McNemar test (exact2x2):
proportion +1 to -1: 0.324
95% CI: 0.037 to 0.586
p-value = 0.041
```

Discordant pairs: 12 (before+ / after-) vs. 25 (before- / after+). McNemar $\chi^2(1) = 4.57$, p = 0.033; exact p = 0.041.

4.24.7 Interpretation

Reporting: “Of 37 discordant pairs, 25 switched from negative to positive and 12 switched in the opposite direction; McNemar’s exact test gave p = 0.041, indicating a significant change after training.”

4.24.8 Practical Tips

- Report only the discordant counts and the p-value; concordant counts are irrelevant to McNemar.

- For $b + c < 25$, use the exact version.
- McNemar's odds ratio is b/c; its CI is computed by inverting the exact test.
- For more than two categories per assessment (paired, more than two outcomes), use Bowker's symmetry test or Stuart-Maxwell.
- Stratified McNemar: use conditional logistic regression (`survival::clogit`) for a set of matched pairs with covariates.

4.24.9 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.

4.24.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.25 Mixed ANOVA

4.25.1 Introduction

A mixed ANOVA combines at least one between-subjects factor (treatment arm, sex) with at least one within-subjects factor (time point, condition). It is the standard analysis for longitudinal trials and for split-plot designs. The typical primary question is the **group-by-time interaction**: does the change over time differ between groups?

4.25.2 Prerequisites

Repeated-measures ANOVA, factorial designs.

4.25.3 Theory

For a design with between factor A (a levels, independent groups) and within factor B (b levels, measured on each subject):

$$Y_{ijk} = \mu + \alpha_i + \pi_{k(i)} + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

where $\pi_{k(i)}$ is the subject-level random effect nested in A.

Three effects are tested:

- Main effect of A (between): uses between-subjects MS as denominator.
- Main effect of B (within): uses within-subjects MS as denominator.
- A x B interaction: uses within-subjects MS as denominator.

Sphericity applies to the within factor (and to its interaction with A); Greenhouse-Geisser or Huynh-Feldt corrections are used when needed.

4.25.4 Assumptions

- Independent subjects within each between-group.
- Normal residuals, separately for within and between components.
- Homogeneous variance-covariance structure across between groups (Box's M, rarely formally tested).
- Sphericity of the within factor.

4.25.5 R Implementation

```r
library(afex); library(emmeans); library(ggplot2)
set.seed(2026)

# 30 subjects per arm, measured at 3 times
n ← 30
subj  ← rep(1:(2 * n), each = 3)
arm   ← factor(rep(rep(c("Placebo", "Active"), each = 3), n))
time  ← factor(rep(c("t0", "t1", "t2"), 2 * n))

df ← data.frame(
  subject = factor(subj),
  arm, time,
  y = rnorm(6 * n, mean = 100, sd = 10) +
      ifelse(time == "t1", 3, 0) +
      ifelse(time == "t2", 5, 0) +
      ifelse(arm == "Active" & time == "t2", -8, 0)
)

fit ← aov_car(y ~ arm * time + Error(subject/time), data = df)
fit

# Interaction plot
emm ← emmeans(fit, ~ arm * time)
plot(emm, comparisons = TRUE)

# Simple effects of arm at each time
pairs(emm, by = "time", adjust = "bonferroni")
```

4.25.6 Output & Results

```
Anova Table (Type 3 tests)

 Effect          df  MSE      F     ges     p.value
 arm           1, 58  245  8.12   0.09   0.0062 **
 time          2, 116  45   9.55  0.03  <.001 ***
 arm:time      2, 116  45   4.72  0.02   0.011 *
```

Significant main effects of arm and time, and a significant interaction. Simple-effects analysis localises the interaction.

4.25.7 Interpretation

“A 2 (arm) x 3 (time) mixed ANOVA revealed a significant interaction, F(2, 116) = 4.72, p = 0.011, $\eta^2_G = 0.02$. Simple-effects comparisons showed that at time 2, the Active arm was 8.0 points lower than Placebo (p = 0.003).”

4.25.8 Practical Tips

- For longitudinal trials, the interaction is typically the primary endpoint.
- Sphericity corrections apply only to the within-subjects parts of the model.
- Missing measurements on a within-subject factor cause listwise deletion of that subject; prefer a linear mixed model for unbalanced data.
- Report the interaction plot; it is often more interpretable than the numeric ANOVA table.
- Simple main effects (one factor fixed at a level of the other) give the most interpretable decomposition when the interaction is significant.

4.25.9 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.

4.25.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation

- Non-parametric tests

4.26 Multiple Comparisons: Overview

4.26.1 Introduction

Testing many hypotheses simultaneously inflates the chance of spurious discoveries. Controlling for this inflation is the goal of multiple-comparison corrections. Two broad families exist: **family-wise error rate (FWER)** methods and **false discovery rate (FDR)** methods. Choosing between them depends on the consequences of a false positive.

4.26.2 Prerequisites

Hypothesis testing, Type I error.

4.26.3 Theory

Suppose m hypotheses are tested and V of them are falsely rejected (true nulls). Let R be total rejections.

- **FWER** $= P(V \geq 1)$. The probability of any false rejection.
- **FDR** $= E[V / \max(R, 1)]$. Expected proportion of false positives among rejections.

Bonferroni, Holm, and Hochberg control FWER. Benjamini-Hochberg and Benjamini-Yekutieli control FDR.

When to use which:

- FWER: small number of pre-specified comparisons, confirmatory studies, clinical trial endpoints.
- FDR: exploratory or high-throughput studies with thousands of tests (genomics, imaging), where a few percent false discoveries are acceptable.

4.26.4 Assumptions

Methods assume valid individual p-values (uniform under null). FWER corrections are robust to arbitrary dependence; FDR methods vary (BH needs positive dependence, BY handles arbitrary).

4.26.5 R Implementation

```
set.seed(2026)
m ← 100
pvals ← runif(m) * ifelse(seq_len(m) ⩽ 10, 0.01, 1)   # 10 small, 90 null

adj_methods ← c("bonferroni", "holm", "hochberg", "BH", "BY")
adj ← sapply(adj_methods, function(m_name)
  p.adjust(pvals, method = m_name))

data.frame(method = adj_methods,
           rejections_at_0.05 = colSums(adj < 0.05))
```

4.26.6 Output & Results

```
      method rejections_at_0.05
1 bonferroni                  8
2        holm                  8
3    hochberg                  9
4          BH                 10
5          BY                  7
```

FDR (BH) recovers the full signal set; FWER methods are more conservative. BY is tighter than BH because of its dependence-robust adjustment.

4.26.7 Interpretation

In a paper: “Multiple testing across 100 primary comparisons was controlled at FWER 0.05 using the Holm procedure (8 of 100 hypotheses rejected).” Or: “Across 50 000 transcripts, differential expression was declared at Benjamini-Hochberg q < 0.05 (2 341 transcripts).”

4.26.8 Practical Tips

- Correct for **all** tests that belong to the same family; a family is defined by the inferential goal, not by software runs.
- Pre-specify the family and the method in the protocol; post-hoc changes to the correction are data-snooping.
- For hierarchically structured hypotheses, gatekeeping procedures maintain FWER at each level.
- Report both raw and adjusted p-values for transparency.
- False discovery proportion (FDP) is the realisation of FDR on a given dataset; FDR is the expectation.

4.26.9 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.

4.26.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.27 Null and Alternative Hypotheses

4.27.1 Introduction

Every hypothesis test begins with two hypotheses: a **null** that captures the status quo and an **alternative** that captures what we would conclude if the null is rejected. Stating both clearly, before the data arrive, is the crucial first step – and a remarkably common source of error when skipped.

4.27.2 Prerequisites

Random variables and population parameters.

4.27.3 Theory

The **null hypothesis** H_0 is typically a statement of no effect, no difference, or a specific parameter value. The **alternative** H_1 (or H_A) is what we would conclude if we reject H_0.

Two-sided alternatives are the default when the sign of the effect is unknown or of interest in both directions:

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0.$$

One-sided alternatives specify a direction, increasing power for that direction at the cost of giving up inference in the opposite:

$$H_0 : \mu \leq \mu_0 \quad \text{vs.} \quad H_1 : \mu > \mu_0.$$

Use a one-sided test only when:

- A priori, only one direction is of scientific interest.
- A result in the other direction would be treated identically to a null result.

- The directionality is pre-specified in the protocol, not chosen after seeing the data.

Simple vs. composite: $H_0 : \mu = 0$ is simple (a single value); $H_0 : \mu \leq 0$ is composite (a set). Test statistics and rejection rules differ slightly.

Equivalence/non-inferiority testing flips the roles: the null is that the effect is at least as large as some margin, and rejection establishes equivalence.

4.27.4 Assumptions

Hypotheses must be specified before examining the data. Post-hoc redefinition invalidates the sampling distribution of the p-value.

4.27.5 R Implementation

```r
set.seed(2026)

# Two-sided test
x ← rnorm(40, mean = 2, sd = 5)
t.test(x, mu = 0, alternative = "two.sided")

# One-sided test (H1: mu > 0)
t.test(x, mu = 0, alternative = "greater")

# One-sided test (H1: mu < 0)
t.test(x, mu = 0, alternative = "less")

# TOST (equivalence: |mu| < 1)
library(TOSTER)
tsum_TOST(m1 = mean(x), sd1 = sd(x), n1 = length(x),
          low_eqbound = -1, high_eqbound = 1)
```

4.27.6 Output & Results

The two-sided p will be roughly twice the one-sided p when the point estimate is consistent with H_1's direction. For equivalence testing, rejection of both one-sided components establishes equivalence.

4.27.7 Interpretation

Pre-register or pre-specify the hypotheses in the protocol; do not switch from two-sided to one-sided after seeing the sign of the effect. Misuse of one-sided tests is a common source of reviewer objections.

4.27.8 Practical Tips

- Default to two-sided tests unless a strong, pre-specified rationale for one-sidedness exists.
- Never compute both one-sided and two-sided p-values and report whichever is smaller.
- In equivalence trials, the "null" is dissimilarity and "alternative" is equivalence – reversed from classical tests.

- Composite null hypotheses are handled by evaluating the test statistic at the boundary.
- Frame hypotheses in terms of the parameter of scientific interest, not the test statistic.

4.27.9 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.

4.27.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.28 One-Proportion Test

4.28.1 Introduction

The one-proportion test compares an observed success proportion $\hat{p}$ to a benchmark p_0. Applications: a phase II response rate compared to a historical control; a screening positive rate compared to a national reference; the success rate of a surgical procedure compared to a target.

4.28.2 Prerequisites

Binomial distribution, confidence intervals.

4.28.3 Theory

For $X \sim \text{Binomial}(n, p)$ with observed $\hat{p} = x/n$, three tests coexist:

Exact binomial test: computes the exact two-sided p-value under $H_0 : p = p_0$ using the binomial PMF. This is R's `binom.test()`. Always valid; no approximation.

Normal-approximation (Wald) z-test:

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathcal{N}(0,1).$$

Works when $np_0 \geq 10$ and $n(1-p_0) \geq 10$; biased and unreliable near $p = 0$ or $p = 1$.

Score (Wilson) test: uses the variance under H_0 and inverts to get a CI with better coverage than Wald, especially for small n or extreme p. R's `prop.test()` uses this by default.

Confidence intervals:

- Wald: $\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$ (poor near 0 or 1).
- Wilson: centred, always inside $[0,1]$, good coverage.
- Clopper-Pearson (exact): always valid, conservative.

4.28.4 Assumptions

- Independent Bernoulli trials.
- Fixed n (not a sequential stopping rule).

4.28.5 R Implementation

```r
library(binom)

x ← 32; n ← 80; p0 ← 0.30

# Exact binomial test
binom.test(x, n, p = p0)

# Score (Wilson) test
prop.test(x, n, p = p0, correct = FALSE)

# Normal-approximation test (manual)
phat ← x / n
z ← (phat - p0) / sqrt(p0 * (1 - p0) / n)
pval ← 2 * (1 - pnorm(abs(z)))
c(z = z, p = pval)

# Confidence intervals
binom.confint(x, n, conf.level = 0.95,
              methods = c("wald", "wilson", "exact"))
```

4.28.6 Output & Results

```
        Exact binomial test
x = 32, n = 80, p-value = 0.04215, 95% CI: (0.292, 0.517)

     1-sample test for given proportion without continuity correction
X-squared = 3.9286, df = 1, p-value = 0.04748, 95% CI (score): (0.303, 0.508)

     method  x  n    mean    lower   upper
1      wald 32 80   0.400  0.2926  0.5074
2    wilson 32 80   0.400  0.3020  0.5066
3     exact 32 80   0.400  0.2923  0.5172
```

Observed proportion 0.40 (95 % Wilson CI 0.30 to 0.51); exact binomial p = 0.042.

4.28.7 Interpretation

Report both the point estimate and the CI. In a manuscript: “the response rate was 40.0 % (32/80; 95 % Wilson CI 30.2 % to 50.7 %; exact binomial p = 0.042 vs. the null rate of 30 %).”

4.28.8 Practical Tips

- Use Wilson or Clopper-Pearson CIs by default; Wald is obsolete.
- For small n and extreme p (near 0 or 1), exact tests and intervals are mandatory.
- Do not report zero/one proportions as 0 and 1 with a Wald CI of “0 to 0”; use a Clopper-Pearson upper bound (rule of three: upper bound $\approx 3/n$ if $x = 0$).
- Continuity correction in `prop.test()` is unnecessary for modern purposes; set `correct = FALSE`.
- For agreement with two-proportion tests, use the same CI method (Wilson) in both.

4.28.9 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.

4.28.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation

- Non-parametric tests

4.29 One-Sample t-Test

4.29.1 Introduction

The one-sample t-test compares the mean of a single continuous variable to a pre-specified reference value. It answers questions of the form "does the average differ from some benchmark?" – average IQ differs from 100, mean biomarker level differs from a reference, average weight change differs from zero in a pre-post design (via paired differences).

4.29.2 Prerequisites

t-distribution, the concept of a sampling distribution.

4.29.3 Theory

Given iid observations $X_1, \ldots, X_n$ from $\mathcal{N}(\mu, \sigma^2)$, the test of

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

uses the statistic

$$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1} \quad \text{under } H_0.$$

The sample SD s replaces the population σ; the distribution adjusts from normal to t because this replacement adds variability. The degrees of freedom are $n - 1$.

Confidence interval for μ: $\bar{X} \pm t_{1-\alpha/2, n-1} \cdot s/\sqrt{n}$.

Effect size: Cohen's $d = (\bar{X} - \mu_0)/s$.

4.29.4 Assumptions

- Independent observations.
- Normal distribution (or n large enough for CLT to kick in).
- No extreme outliers affecting the mean.

4.29.5 R Implementation

```r
library(effectsize); library(broom)
set.seed(2026)

# Simulated IQ-test scores; is the mean different from 100?
iq ← rnorm(45, mean = 103.5, sd = 14)

t.test(iq, mu = 100) ▷ tidy()

# Effect size
cohens_d(iq, mu = 100)

# Check assumptions
shapiro.test(iq)
boxplot(iq, main = "Sample distribution", col = "#2A9D8F")
```

4.29.6 Output & Results

```
# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high method  alternative
     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <chr>   <chr>
1    104.      1.82   0.076      44     98.9     109.      One ... two.sided

Cohen's d |         95% CI
-------------------------------
0.27      | [-0.03, 0.57]
```

The sample mean is 104, not significantly different from 100 at $\alpha = 0.05$ ($t(44) = 1.82$, $p = 0.076$), with a small-to-medium effect ($d = 0.27$).

4.29.7 Interpretation

In a manuscript: “The mean IQ was 104.0 (SD 14.0). The 95 % CI for the mean (98.9 to 109.3) includes the reference value of 100; a one-sample t-test did not reject the null of equal mean (t(44) = 1.82, p = 0.076, Cohen’s d = 0.27).”

4.29.8 Practical Tips

- For $n \geq 30$, the CLT makes the t-test fairly robust to non-normality.
- For small n with clear non-normality, use the Wilcoxon signed-rank test (one-sample version).
- Power for the one-sample t-test depends on effect size and n; use `power.t.test(type = "one.sample")`.
- The reference value μ_0 must be pre-specified; post-hoc selection invalidates the p-value.
- Pair with a 95 % CI for the mean – it is more informative than the p-value alone.

4.29.9 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.

4.29.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.30 One-Sample z-Test

4.30.1 Introduction

The one-sample z-test is the idealised version of the one-sample t-test: it assumes the population variance σ^2 is known exactly. This is rarely the case in applied research – if σ is unknown, we estimate it from the sample and use the t-test instead. The z-test remains pedagogically important because it is the simplest exact hypothesis test and the reference for the t-test's more elaborate machinery.

4.30.2 Prerequisites

Normal distribution, sampling distribution of the mean.

4.30.3 Theory

For $X_1, \ldots, X_n$ iid from $\mathcal{N}(\mu, \sigma^2)$ with **known** σ, the test of

$$H_0 : \mu = \mu_0 \quad \text{vs.} \quad H_1 : \mu \neq \mu_0$$

uses

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1) \quad \text{under } H_0.$$

The two-sided p-value is $2(1 - \Phi(|z_{\text{obs}}|))$; 95 % CI for μ is $\bar{X} \pm 1.96\sigma/\sqrt{n}$.

The z-test is also used when σ is *well-estimated* – historically meaning $n > 30$ where the t and z distributions are nearly indistinguishable. Modern practice uses the t-test even then.

4.30.4 Assumptions

- Independent observations.
- Normal population (or n large enough for CLT).
- Population variance σ^2 known.

4.30.5 R Implementation

```r
set.seed(2026)

# Assume the assay has a known SD from reference literature
sigma_known <- 8
x <- rnorm(50, mean = 103, sd = sigma_known)

# Manual z-test
z_stat <- (mean(x) - 100) / (sigma_known / sqrt(length(x)))
p_two_sided <- 2 * (1 - pnorm(abs(z_stat)))
c(z = z_stat, p = p_two_sided)

# 95% CI
mean(x) + c(-1, 1) * qnorm(0.975) * sigma_known / sqrt(length(x))

# Via BSDA::z.test
library(BSDA)
z.test(x, mu = 100, sigma.x = sigma_known)
```

4.30.6 Output & Results

```
       z         p
2.1856   0.02884

[1] 100.53 105.47

        One-sample z-Test

data:  x
z = 2.1856, p-value = 0.02884
alternative hypothesis: true mean is not equal to 100
```

The sample mean is 103 with SE 1.13 (known SD / sqrt(n)); z = 2.19, p = 0.029, rejecting H_0 at $\alpha = 0.05$.

4.30.7 Interpretation

Reporting a z-test is rarely appropriate in modern applied work unless the variance is genuinely known (e.g., from decades of reference data). Reporting the t-test instead gives an honest representation of the variance uncertainty.

4.30.8 Practical Tips

- Default to the t-test; use z only when σ is genuinely known.
- For two-proportion comparisons, the “z-test for proportions” is a close cousin that uses an estimated variance.

- When $n > 30$ and σ is unknown, the t and z tests give essentially identical p-values.
- The z-test is the foundation for more elaborate procedures: power calculations, sample-size formulas, and the large-sample limit of many Wald-style tests.
- In process-control charts (SPC), the "3-sigma" rule is a z-test at $\alpha \approx 0.0027$.

4.30.9 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.

4.30.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.31 One-Way ANOVA

4.31.1 Introduction

One-way ANOVA tests whether the means of three or more independent groups are equal. It generalises the two-sample t-test to multiple groups while controlling overall Type I error. The omnibus F-test tells us whether *any* means differ; post-hoc tests (Tukey HSD, Dunnett) identify which pairs.

4.31.2 Prerequisites

Two-sample t-test, F distribution.

4.31.3 Theory

For k groups with n_i observations each, total $n = \sum n_i$, define:

- **Sum of squares between** (SSB): $\sum_i n_i(\bar{X}_i - \bar{X})^2$ with $k - 1$ df.
- **Sum of squares within** (SSW): $\sum_i \sum_j (X_{ij} - \bar{X}_i)^2$ with $n - k$ df.
- **Mean squares**: MSB = SSB / (k - 1), MSW = SSW / (n - k).

Under $H_0 : \mu_1 = \ldots = \mu_k$,

$$F = \frac{\text{MSB}}{\text{MSW}} \sim F_{k-1,n-k}.$$

Large F implies between-group variation exceeds within-group variation, rejecting equality of means.

Effect size: $\eta^2 = \text{SSB}/\text{SST}$, the proportion of variance explained by the group factor.

4.31.4 Assumptions

- Independent observations within and across groups.
- Normal residuals (or large enough n_i).
- Homogeneous variances across groups (Welch ANOVA relaxes this).

4.31.5 R Implementation

```r
library(effectsize); library(broom); library(car)
set.seed(2026)

df ← data.frame(
  group = factor(rep(c("A", "B", "C", "D"), each = 30)),
  y     = c(rnorm(30, 50, 8),
            rnorm(30, 55, 8),
            rnorm(30, 58, 8),
            rnorm(30, 54, 8))
)

# Omnibus ANOVA
fit ← aov(y ~ group, data = df)
summary(fit)
tidy(fit)

# Assumption checks
leveneTest(y ~ group, data = df)
shapiro.test(residuals(fit))

# Effect size
eta_squared(fit)
omega_squared(fit)

# Welch ANOVA (unequal variances)
oneway.test(y ~ group, data = df)

# Tukey HSD post-hoc
TukeyHSD(fit)
```

4.31.6 Output & Results

```
            Df Sum Sq Mean Sq F value  Pr(>F)
group        3    891   297.0    4.83 0.00317 **
Residuals  116   7135    61.5

Eta2 (partial) | 95% CI
-------------------------
0.11           | [0.02, 1.00]

        Tukey multiple comparisons of means 95% family-wise confidence level

         diff    lwr    upr  p adj
B-A   5.421   -0.58  11.42 0.097
C-A   8.215    2.21  14.22 0.003
D-A   4.129   -1.87  10.13 0.287
C-B   2.794   -3.21   8.79 0.627
D-B  -1.292   -7.29   4.71 0.943
D-C  -4.086  -10.09   1.92 0.297
```

F(3, 116) = 4.83, p = 0.003, rejects equal means. Tukey identifies C vs. A as the significant pair.

4.31.7 Interpretation

Report: “A one-way ANOVA showed a significant group effect (F(3, 116) = 4.83, p = 0.003, $\eta^2 = 0.11$). Tukey HSD comparisons indicated that group C differed significantly from group A (mean difference 8.2, 95 % CI 2.2 to 14.2, adjusted p = 0.003).”

4.31.8 Practical Tips

- Always check homogeneity of variances; if violated, use Welch ANOVA.
- Always report an effect size; F alone is not enough.
- For many post-hoc comparisons, Tukey is the standard; Dunnett is for comparing treatments to a control; Bonferroni or Holm for pre-specified families.
- Non-parametric alternative: Kruskal-Wallis.
- For within-subjects data, use repeated-measures ANOVA or a linear mixed model.

4.31.9 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.

4.31.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests

- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.32 P-Values Explained

4.32.1 Introduction

The p-value is the single most reported statistic in applied research, and the most frequently misinterpreted. Understanding what a p-value is – and what it is not – is essential for reading, writing, and reviewing quantitative studies. The American Statistical Association's 2016 statement on p-values laid out six principles that every user of statistics should internalise.

4.32.2 Prerequisites

Basic hypothesis testing and sampling distributions.

4.32.3 Theory

A **p-value** is the probability, assuming the null hypothesis is true, of observing data at least as extreme as what was actually observed. Formally,

$$p = P(T \geq t_{\text{obs}} \mid H_0)$$

for a one-sided test with test statistic T, or a two-sided analogue for two-sided alternatives.

Correct interpretation: a small p-value means the observed data would be unusual under H_0. It does **not** mean that H_0 is unlikely to be true, nor that the alternative is probable.

Common misinterpretations:

1. "$p = 0.03$ means there's a 3% probability the null is true." Wrong: the p-value is a conditional probability of data given H_0, not of H_0 given data.
2. "$p < 0.05$ means the result is meaningful." Wrong: small p-values can come from huge samples detecting trivial effects.
3. "$p > 0.05$ means no effect." Wrong: failure to reject is not evidence of no effect; it is absence of evidence.
4. "p-values from the same experiment are comparable." Wrong: p-values are discontinuous in the effect size; a p of 0.04 and 0.06 are nearly identical evidentially.

Under H_0, the p-value has a uniform distribution on $[0, 1]$ (for continuous test statistics). This is a useful simulation check.

4.32.4 Assumptions

- The null hypothesis and test statistic are pre-specified.
- The sampling model under H_0 is correctly specified.
- Multiple testing, optional stopping, and selective reporting are not present.

4.32.5 R Implementation

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

# Simulate p-values under H0 (they should be uniform)
n_sims ← 5000
pvals_null ← replicate(n_sims, {
  x ← rnorm(30, mean = 0, sd = 1)   # H0 true: mu = 0
  t.test(x, mu = 0)$p.value
})

hist(pvals_null, breaks = 40, col = "#2A9D8F",
     main = "p-values under H0 are uniform", xlab = "p-value")

# Simulate under H1
pvals_alt ← replicate(n_sims, {
  x ← rnorm(30, mean = 0.5, sd = 1)
  t.test(x, mu = 0)$p.value
})

hist(pvals_alt, breaks = 40, col = "#F4A261",
     main = "p-values under H1 pile near zero", xlab = "p-value")

# Empirical power at alpha = 0.05
mean(pvals_alt < 0.05)
power.t.test(n = 30, delta = 0.5, sd = 1)$power
```

4.32.6 Output & Results

```
[1] 0.763    # empirical power
[1] 0.752    # theoretical power
```

Under H_0, the histogram is flat. Under H_1, p-values pile near zero. The empirical rejection rate at $\alpha = 0.05$ matches the theoretical power.

4.32.7 Interpretation

When reporting, pair p-values with effect sizes and confidence intervals. “The difference was 3.2 mmHg (95 % CI 0.6 to 5.8 mmHg, $p = 0.015$)” conveys both the magnitude and the statistical significance.

4.32.8 Practical Tips

- Never report a p-value alone; always include the effect size and its CI.
- Pre-specify one-sided vs. two-sided tests; switching after seeing data is fishing.

- Large samples make tiny effects statistically significant; focus on practical significance when n is large.
- Multiple testing inflates false-positive rates; adjust with Bonferroni, Holm, or FDR.
- "Nearly significant" ($0.05 < p < 0.10$) is not a meaningful category; report the exact value.

4.32.9 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.

4.32.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.33 Paired t-Test

4.33.1 Introduction

When each unit contributes two measurements (pre and post, left and right, case and matched control), the observations are paired and the usual independent-samples t-test is not valid. The **paired t-test** analyses the within-unit differences, restoring the independence assumption and typically improving power by cancelling between-unit variability.

4.33.2 Prerequisites

One-sample t-test, t-distribution.

4.33.3 Theory

Let X_i and Y_i be paired measurements on unit i for $i = 1, \ldots, n$. Define differences $D_i = X_i - Y_i$. Under the assumption that D_i are iid from $\mathcal{N}(\mu_D, \sigma_D^2)$,

$$T = \frac{\bar{D}}{s_D/\sqrt{n}} \sim t_{n-1} \quad \text{under } H_0 : \mu_D = 0.$$

The paired t-test is the one-sample t-test on the differences. The n is the number of **pairs**, not the total number of measurements.

Effect size: Cohen's $d_z = \bar{D}/s_D$ (standardised mean difference on the difference scores). This is different from the two-group d and should not be conflated.

4.33.4 Assumptions

- The differences are approximately normal (via Shapiro-Wilk on D_i).
- Pairs are independent of other pairs.

The raw X_i and Y_i can be non-normal; only the differences need to be.

4.33.5 R Implementation

```
library(effectsize); library(broom)
set.seed(2026)

# Simulated pre-post blood pressure after a 12-week intervention
n ← 30
pre  ← rnorm(n, mean = 138, sd = 12)
post ← pre + rnorm(n, mean = -6, sd = 5)

# Paired t-test
t.test(pre, post, paired = TRUE) ▷ tidy()

# Equivalent: one-sample t-test on the differences
t.test(pre - post, mu = 0) ▷ tidy()

# Effect size
cohens_d(pre, post, paired = TRUE)

# Assumption check on differences
shapiro.test(pre - post)
```

4.33.6 Output & Results

```
# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high method       alternative
     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <chr>        <chr>
1     6.04      6.88 <0.001        29     4.25      7.84 Paired t-... two.sided

Cohen's d (paired) |         95% CI
------------------------------------
1.26               | [0.78, 1.74]
```

The post-intervention measurement is 6.0 mmHg lower on average; $t(29) = 6.88$, $p < 0.001$, large paired effect.

4.33.7 Interpretation

In a manuscript: “Blood pressure decreased by 6.0 mmHg on average from baseline to week 12 (paired t(29) = 6.88, p < 0.001, Cohen’s d_z = 1.26, 95 % CI 4.3 to 7.8 mmHg).”

4.33.8 Practical Tips

- Paired design usually has more power than independent-groups design at the same total sample size.
- Check normality of the differences, not the raw measurements.
- For non-normal differences, use Wilcoxon signed-rank.
- Pair variables correctly: subjects must be in the same order in both vectors, or use a data frame.
- Report Cohen’s d_z, not pooled-SD two-group d, for paired designs.

4.33.9 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.

4.33.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.34 Pearson Correlation Test

4.34.1 Introduction

The Pearson correlation test assesses whether two continuous variables are linearly related. It returns a correlation coefficient r and a p-value for the null that the population correlation $\rho = 0$, along with a confidence interval computed via Fisher’s z transformation.

4.34.2 Prerequisites

Pearson correlation coefficient, sampling distributions.

4.34.3 Theory

Given iid pairs (X_i, Y_i) from a bivariate distribution with marginal SDs σ_X, σ_Y, the sample correlation is

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}}.$$

Under $H_0 : \rho = 0$, the statistic

$$t = r\sqrt{\frac{n-2}{1-r^2}} \sim t_{n-2}.$$

For confidence intervals, Fisher's z-transformation is

$$z = \frac{1}{2}\log\frac{1+r}{1-r}, \quad \mathrm{SE}(z) = \frac{1}{\sqrt{n-3}}.$$

Construct a CI in the z scale, then back-transform with $r = \tanh(z)$.

4.34.4 Assumptions

- Continuous X and Y, approximately bivariate normal.
- Linear relationship; no extreme outliers.
- Independent pairs.

4.34.5 R Implementation

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

n ← 80
x ← rnorm(n)
y ← 0.4 * x + rnorm(n, sd = sqrt(1 - 0.4^2))  # true rho = 0.4

cor.test(x, y, method = "pearson") ▷ tidy()

# Manual check
r ← cor(x, y)
t_stat ← r * sqrt((n - 2) / (1 - r^2))
p_val  ← 2 * (1 - pt(abs(t_stat), df = n - 2))
c(r = r, t = t_stat, p = p_val)

# Fisher's z CI
z    ← atanh(r)
se_z ← 1 / sqrt(n - 3)
ci_z ← z + c(-1, 1) * 1.96 * se_z
tanh(ci_z)
```

4.34.6 Output & Results

```
# A tibble: 1 x 8
  estimate statistic p.value parameter conf.low conf.high method    alternative
     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>     <dbl> <chr>     <chr>
1    0.43      4.19   <0.001        78   0.234     0.598  Pearson   two.sided
```

Sample $r = 0.43$ (95 % CI 0.23 to 0.60), $t(78) = 4.19$, $p < 0.001$. Strong evidence of a positive correlation.

4.34.7 Interpretation

“There was a significant positive correlation between x and y ($r = 0.43$, 95 % CI 0.23 to 0.60; $t(78) = 4.19$, $p < 0.001$). The two variables shared 18.5 % of their variance linearly.”

4.34.8 Practical Tips

- Scatter plot first; r can be misleading if the relationship is non-linear.
- Robust alternatives for heavy tails or outliers: Spearman, Kendall.
- For small n, the CI via Fisher’s z is noticeably asymmetric in the r scale.
- Always report both r and its CI; the p-value alone is often uninformative.
- Testing multiple correlations from the same data requires multiple-testing correction.

4.34.9 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.

4.34.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.35 Permutation Tests

4.35.1 Introduction

A permutation test evaluates a null hypothesis by comparing the observed test statistic to its distribution under random permutations of the group labels (or other exchangeable structure). It makes no distributional assumption about the data, only about **exchangeability** under the null.

4.35.2 Prerequisites

Hypothesis testing, ranks (helpful).

4.35.3 Theory

Under H_0, group labels are exchangeable – the joint distribution of $(X_1, \ldots, X_n)$ is invariant under permutations of the labels. The p-value is the proportion of permuted test statistics at least as extreme as the observed one:

$$p = \frac{1 + \#\{T^{(b)} \geq T_{\text{obs}}\}}{B + 1},$$

where $T^{(b)}$ is the statistic from the b-th permutation out of B random permutations, plus the observed. Exact permutation tests use all $n!$ permutations; for large n, Monte Carlo permutation with $B = 1000$ or more suffices.

Permutation tests can use **any test statistic** – difference of means, t-statistic, ratio of medians, area under an ROC. The choice of statistic determines what aspect of the distribution the test is sensitive to.

4.35.4 Assumptions

- Exchangeability under H_0 (independent observations with equal distribution for group-based tests).
- Independence across observations.

No normality, no variance homogeneity, no distributional model.

4.35.5 R Implementation

```
library(coin); library(perm)
set.seed(2026)

group1 ← c(rlnorm(20, 1, 0.4), 30)   # one extreme value
group2 ← rlnorm(25, 1.2, 0.4)

# Manual permutation test on difference of means
obs_diff ← mean(group1) - mean(group2)
pooled   ← c(group1, group2)
n1       ← length(group1)
B        ← 10000

perm_diffs ← replicate(B, {
  shuffled ← sample(pooled)
  mean(shuffled[1:n1]) - mean(shuffled[(n1+1):length(shuffled)])
})
p_perm ← (1 + sum(abs(perm_diffs) ⩾ abs(obs_diff))) / (B + 1)
c(obs_diff = obs_diff, p_perm = p_perm)

# coin package: permutation version of many standard tests
oneway_test(y ~ group,
            data = data.frame(y = c(group1, group2),
                              group = factor(rep(c("A", "B"), c(n1,
length(group2))))),
            distribution = approximate(nresample = 10000))
```

4.35.6 Output & Results

```
obs_diff    p_perm
  0.841     0.123

Approximative Two-Sample Fisher-Pitman Permutation Test
Z = 1.56, p = 0.13
```

Permutation test gives p = 0.12, robust to the outlier that would distort a t-test.

4.35.7 Interpretation

“A Monte Carlo permutation test (10 000 resamples) comparing the two groups’ mean differences gave p = 0.12, providing no evidence of a group effect.”

4.35.8 Practical Tips

- Permutation tests are exact in small samples where combinatorial enumeration is feasible.
- Monte Carlo permutation uses random resampling; $B = 1000$-10 000 gives stable p-values.
- The statistic can be any function of the data; this flexibility is the main appeal over parametric tests.
- For paired data, permute within pairs (sign flips) rather than labels.
- For regression, permute residuals (or bootstrap them) rather than raw responses.

4.35.9 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.

4.35.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.36 Post-Hoc Tests with Tukey HSD

4.36.1 Introduction

When a one-way ANOVA rejects equality of means across groups, post-hoc tests identify which specific pairs differ. Tukey's HSD (honestly significant difference) is the default pairwise comparison test after ANOVA with equal sample sizes, controlling the family-wise error rate at the nominal α.

4.36.2 Prerequisites

One-way ANOVA, multiple-comparisons problem.

4.36.3 Theory

For k groups with sample size n each and common residual MS = MSE, the Tukey critical value for pairwise differences is

$$q_{\alpha,k,n_{\text{total}}-k} \cdot \sqrt{\text{MSE}/n},$$

where q is the Studentised range quantile. Pairs whose observed mean difference exceeds this are declared significantly different at family-wise α.

Tukey's HSD is equivalent to a simultaneous test based on the maximum pairwise t statistic. Alternatives:

- **Dunnett**: comparing each treatment to a single control; more powerful than Tukey when this is the question.
- **Bonferroni**: simple but conservative; often over-adjusts.
- **Holm**: sequential, uniformly more powerful than Bonferroni.
- **Scheffe**: most conservative; valid for arbitrary contrasts, not just pairwise.

4.36.4 Assumptions

Same as ANOVA: independence, normal residuals, homogeneous variances. For unequal sample sizes, the Tukey-Kramer variant is used.

4.36.5 R Implementation

```r
library(emmeans); library(multcomp)
set.seed(2026)

df ← data.frame(
  group = factor(rep(c("A", "B", "C", "D"), each = 30)),
  y     = c(rnorm(30, 50, 8),
            rnorm(30, 58, 8),
            rnorm(30, 60, 8),
            rnorm(30, 55, 8))
)

fit ← aov(y ~ group, data = df)

# Built-in Tukey HSD
TukeyHSD(fit)

# emmeans equivalent (more flexible, handles unbalanced designs)
emm ← emmeans(fit, ~ group)
pairs(emm, adjust = "tukey")

# Dunnett (comparing to A as control)
summary(glht(fit, linfct = mcp(group = "Dunnett")))
```

4.36.6 Output & Results

```
            diff     lwr     upr   p adj
B-A     8.421     3.02  13.82    0.0006
C-A     9.897     4.50  15.30    0.00005
D-A     4.975    -0.43  10.38    0.086
C-B     1.476    -3.92   6.88    0.892
D-B    -3.446    -8.85   1.96    0.353
D-C    -4.922   -10.32   0.48    0.090
```

Three of six pairwise comparisons are significant at family-wise $\alpha = 0.05$.

4.36.7 Interpretation

Report adjusted p-values and family-wise confidence intervals. “Tukey HSD post-hoc comparisons following a significant one-way ANOVA indicated that groups B and C each differed significantly from group A (both adjusted p < 0.01), while D did not differ

significantly from A (adjusted p = 0.09).”

4.36.8 Practical Tips

- Only perform post-hoc tests after a significant omnibus F; otherwise you are fishing.
- For unequal sample sizes use Tukey-Kramer (R’s `TukeyHSD` and `emmeans` handle this automatically).
- Dunnett is more powerful than Tukey when all comparisons are to one control.
- For complex (non-pairwise) contrasts, use `emmeans` or `glht` with user-defined linear combinations.
- Conservative alternatives (Bonferroni, Holm) work regardless of equal variances; they are safer in heterogeneous-variance settings.

4.36.9 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.

4.36.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.37 Repeated-Measures ANOVA

4.37.1 Introduction

Repeated-measures ANOVA (rmANOVA) handles designs where the same subjects are measured at multiple time points or conditions. By treating subject as a blocking factor, it controls between-subject variability and gains power over between-subjects analysis. The main complication is the **sphericity** assumption on the covariance of the repeated measures.

4.37.2 Prerequisites

One-way ANOVA, dependence structure in repeated measures.

4.37.3 Theory

For one within-subjects factor with k levels measured on n subjects, the model is

$$Y_{ij} = \mu + \alpha_i + \pi_j + \varepsilon_{ij},$$

where π_j is a random subject effect and α_i is the fixed effect of level i.

Sphericity assumes the variances of all pairwise differences between repeated measurements are equal. Mauchly's test assesses this assumption.

If sphericity is violated:

- **Greenhouse-Geisser correction**: adjusts degrees of freedom by multiplying by $\hat{\varepsilon}_{GG} \in [1/(k-1), 1]$.
- **Huynh-Feldt correction**: less conservative for mild violations.

Modern alternative: fit a linear mixed-effects model (`lmer`) with an unstructured or compound-symmetric covariance – more flexible, handles missing data.

4.37.4 Assumptions

- Normality of within-subject residuals.
- Sphericity (or apply a correction).
- Subjects are independent; measurements within a subject are not.

4.37.5 R Implementation

```r
library(afex); library(emmeans); library(rstatix)
set.seed(2026)

# 20 subjects measured at 4 time points
n_subj <- 20
times  <- 4
subject <- rep(1:n_subj, each = times)
time    <- rep(1:times, n_subj)

# Simulate with a time effect and subject-specific intercepts
df <- data.frame(
  subject = factor(subject),
  time    = factor(time),
  y       = rnorm(n_subj * times, mean = 50 + 2 * time, sd = 6) +
            rep(rnorm(n_subj, 0, 4), each = times)
)

# Using afex::aov_car
fit <- aov_car(y ~ time + Error(subject/time), data = df)
fit

# rstatix pipeline
df |> anova_test(dv = y, wid = subject, within = time)

# Post-hoc with Bonferroni
emm <- emmeans(fit, ~ time)
pairs(emm, adjust = "bonferroni")
```

4.37.6 Output & Results

```
Anova Table (Type 3 tests)

Response: y
   Effect         df  MSE       F  ges  p.value
1    time  2.61, 49.5 28.7  12.39  0.19  <.001 ***

Mauchly Tests for Sphericity:
 Effect   W       p
  time  0.87  0.78   # sphericity holds

Greenhouse-Geisser eps: 0.87
```

Large within-subject effect of time (F(2.61, 49.5) = 12.39, p < 0.001, generalised $\eta^2 = 0.19$); sphericity holds so no correction needed.

4.37.7 Interpretation

“A one-way repeated-measures ANOVA revealed a significant effect of time (F(2.61, 49.5) = 12.39, p < 0.001, generalised $\eta^2 = 0.19$; Greenhouse-Geisser correction applied). Bonferroni post-hoc comparisons identified significant increases from time 1 to time 3 and from time 2 to time 4.”

4.37.8 Practical Tips

- Check Mauchly’s test; when rejected, apply Greenhouse-Geisser or Huynh-Feldt.
- For complex designs (multiple within factors, between x within interactions), a linear mixed model is preferred.
- Missing data in rmANOVA causes the entire subject to be dropped (listwise); mixed models handle it gracefully.
- Report generalised η^2 for repeated-measures designs (partial η^2 is upwardly biased).
- For non-normal residuals, use Friedman or an aligned-rank transform (`ARTool::art`).

4.37.9 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.

4.37.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation

- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.38 The Runs Test

4.38.1 Introduction

The runs test is a non-parametric test of randomness in a sequence of binary outcomes (or continuous outcomes that have been dichotomised against a reference such as the median). A “run” is a maximal subsequence of identical symbols — for example, in the sequence HHTTTHHH the runs are HH, TTT, HHH, three runs in total. Under genuine randomness the number of runs follows a known distribution; too few runs indicate clustering of like outcomes (suggesting positive autocorrelation or a structural break), while too many runs indicate alternation (negative autocorrelation). The runs test is widely used in quality-control charts to detect assignable causes, in regression diagnostics to flag autocorrelated residuals, and in randomness testing of sequences from random-number generators or dice-rolling sequences.

4.38.2 Prerequisites

A working understanding of binary sequences, the concept of independent identically distributed observations, and the role of randomness diagnostics in time-ordered or sequence data.

4.38.3 Theory

For a binary sequence with n_1 successes and n_2 failures, the expected number of runs under the null hypothesis of randomness is

$$E[R] = \frac{2n_1n_2}{n_1+n_2} + 1, \qquad \operatorname{Var}(R) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}.$$

The standardised statistic $Z = (R - E[R])/\sqrt{\operatorname{Var}(R)}$ is approximately standard Normal for moderate n (typically $n_1, n_2 \geq 10$). For continuous data, dichotomising against the sample median and applying the same test is the standard adaptation, though continuous-specific alternatives such as the Durbin-Watson test are usually more powerful for regression residuals.

4.38.4 Assumptions

The sequence is binary or has been dichotomised in a pre-specified way, observations under the null are independent and identically distributed Bernoulli, and the sample size is large enough for the Normal approximation. Tied values at the dichotomisation cut-off must be handled deliberately — dropped, or broken randomly with appropriate caution.

4.38.5 R Implementation

```
library(tseries); library(randtests)
set.seed(2026)

random_seq ← rbinom(50, 1, 0.5)
runs.test(factor(random_seq))

clustered ← rep(c(0, 1), each = 25)
runs.test(factor(clustered))

alternating ← rep(c(0, 1), times = 25)
runs.test(factor(alternating))

x ← rnorm(50) + seq(0, 2, length.out = 50)
runs.test(factor(as.integer(x > median(x))))
```

4.38.6 Output & Results

The four examples illustrate the four distinct patterns the test detects: a truly random sequence ($z = -0.23$, $p = 0.82$), a strongly clustered sequence ($z = -6.85$, $p < 0.001$), an alternating sequence ($z = 6.85$, $p < 0.001$), and a trended continuous sequence dichotomised at the median ($z = -3.12$, $p = 0.002$). The sign of z distinguishes clustering (negative) from alternation (positive).

4.38.7 Interpretation

A reporting sentence: “The runs test applied to the residuals of the regression model showed no significant departure from randomness ($Z = -0.23$, $p = 0.82$), supporting the assumption of independent residuals along the observation order. For comparison, the test applied to a deliberately clustered binary sequence yielded $Z = -6.85$ ($p < 0.001$), and to an alternating sequence $Z = +6.85$ ($p < 0.001$), illustrating the test’s sensitivity to both directions of non-randomness.” Always describe whether departures are toward clustering or alternation.

4.38.8 Practical Tips

- Use the runs test as a quick-and-dirty diagnostic for autocorrelation or trend in residuals from regression or quality-control monitoring; for regression-specific autocorrelation diagnostics, the Durbin-Watson test or Breusch-Godfrey test on continuous residuals is more powerful and should be preferred.
- For continuous sequences, dichotomising at the median loses substantial information; the test becomes a relatively weak diagnostic compared with

continuous-specific alternatives, but it remains useful for quick visual inspection of long sequences.
- Ties at the median require deliberate handling: dropping tied values reduces the effective sample size and is the standard recommendation; breaking ties randomly inflates the apparent randomness and should be done only with explicit caution.
- The Wald-Wolfowitz generalisation extends the runs concept to two-sample comparisons of distributions via run counts after merging and ranking; it is a non-parametric alternative to the two-sample t-test or Mann-Whitney test in some specialised settings.
- In statistical-process-control charts, runs tests are used to detect assignable causes — systematic shifts, trends, or oscillations — in a time-ordered sequence of measurements; Western Electric and Nelson rules are operational implementations of runs-test logic.
- For very long sequences ($n > 1000$), the Normal approximation is excellent and exact tables are unnecessary; for very short sequences ($n < 20$), use the exact distribution implemented in `randtests::runs.test()`.

4.38.9 R Packages Used

`tseries::runs.test()` and `randtests::runs.test()` for the canonical Wald-Wolfowitz runs test; `randtests` for an extensive collection of randomness tests including Bartels rank-test, turning-point test, and rank-version variants; `lawstat::runs.test()` for an alternative implementation; `lmtest::dwtest()` and `lmtest::bgtest()` for continuous-residual autocorrelation diagnostics complementing the runs test.

4.38.10 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.

4.38.11 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.39 Shapiro-Wilk Normality Test

4.39.1 Introduction

The Shapiro-Wilk test, introduced by Shapiro and Wilk in 1965, is the most widely used and generally most powerful test of Normality for small-to-moderate sample sizes. It compares the observed sample order statistics with their expected values under Normality, producing a statistic W that is close to 1 when the sample is well approximated by a Normal distribution and substantially lower otherwise. Shapiro-Wilk is the default Normality test in R for samples between 3 and 5000 observations and is routinely used as a preliminary diagnostic before applying parametric inference (one- and two-sample t-tests, ANOVA, regression on the residuals) that nominally requires Normal data.

4.39.2 Prerequisites

A working understanding of order statistics, the Normal distribution, and the role of Normality assumptions in parametric inferential statistics.

4.39.3 Theory

For a sorted sample $x_{(1)} \leq \cdots \leq x_{(n)}$, the Shapiro-Wilk statistic is

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where a_i are tabulated constants derived from the expected values and covariance matrix of the order statistics of a standard Normal sample. The numerator is essentially the squared correlation between the sample order statistics and their expected values, scaled to produce $W \in (0, 1]$. Values near 1 indicate a Normal sample, lower values indicate departure; small W with correspondingly small p-value rejects the null of Normality.

4.39.4 Assumptions

Observations are independent and identically distributed; sample size lies within R's implementation range $n \in [3, 5000]$. The test does not assume the sample is Normal under the null — the null is precisely that hypothesis — but it does assume the observations are otherwise exchangeable.

4.39.5 R Implementation

```r
set.seed(2026)

x_norm ← rnorm(60)
shapiro.test(x_norm)

x_skew ← rexp(60, rate = 1)
shapiro.test(x_skew)

x_t3 ← rt(60, df = 3)
shapiro.test(x_t3)

set.seed(2026)
pvals ← replicate(1000, shapiro.test(rt(20, df = 5))$p.value)
mean(pvals < 0.05)
```

4.39.6 Output & Results

The three samples illustrate the test's behaviour: a truly Normal sample of 60 observations gives $W = 0.98$, $p = 0.52$ (no rejection); an exponentially-distributed (right-skewed) sample of 60 gives $W = 0.83$, $p = 0.00002$ (strong rejection); a t_3-distributed (heavy-tailed) sample gives $W = 0.93$, $p = 0.002$ (rejection). The Monte Carlo block estimates empirical power at $n = 20$ for detecting t_5-distributed data — about 21 % — illustrating the test's modest power at small samples for mild departures.

4.39.7 Interpretation

A reporting sentence: "Shapiro-Wilk did not reject the Normality of the regression residuals ($W = 0.98$, $p = 0.52$), supporting the assumption underlying parametric inference. The complementary Q-Q plot showed close adherence to the diagonal, with no evidence of systematic deviation in the tails or at the centre. Parametric t-tests and ANOVA were therefore applied to these data without transformation. For comparison, residuals from a related model with skewed errors were rejected by the same test ($W = 0.83$, $p < 0.001$) and motivated a log transformation." Always pair test with a Q-Q plot.

4.39.8 Practical Tips

- At very small samples ($n < 15$), Shapiro-Wilk has low power against most realistic departures and cannot reliably confirm Normality; absence of rejection at small samples is uninformative and should not be treated as evidence of Normality.
- At very large samples ($n > 1000$), the test rejects at trivial departures from Normality that are inconsequential for the validity of parametric inference; use Q-Q plots and effect-size summaries (skewness, kurtosis with their CIs) rather than the p -value alone in large samples.
- Graphical diagnosis via a Q-Q plot complements the numerical test and is often more informative; always plot the Q-Q before deciding whether departures are practically important. The combination of formal test and visual inspection is the modern recommendation.

- Normality assumptions in regression and ANOVA apply to the residuals, not to the raw outcome variable; testing Normality of the raw outcome is a recurring methodological error and provides no information about whether parametric inference is valid.
- When Normality fails substantively, consider transformations (log, Box-Cox, Yeo-Johnson), non-parametric alternatives (Wilcoxon, Mann-Whitney, Kruskal-Wallis), or robust methods (trimmed-mean inference, robust regression); the choice depends on the substantive question and the nature of the departure.
- For very heavy-tailed data, robust inference is often preferable to transformation; transformations can shift the substantive interpretation away from the original scale.

4.39.9 R Packages Used

Base R `shapiro.test()` for the canonical Shapiro-Wilk implementation up to $n = 5000$; `nortest::shapiro.test()` for an alternative interface; `nortest` for related Normality tests including Anderson-Darling, Cramér-von Mises, and Lilliefors; `moments::skewness()` and `moments::kurtosis()` for descriptive moment-based summaries; `car::qqPlot()` for enhanced Q-Q plot diagnostics complementing the formal test.

4.39.10 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.

4.39.11 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.40 The Sign Test

4.40.1 Introduction

The sign test is the most assumption-light procedure for a one-sample median or paired comparison. It ignores the magnitude of deviations and uses only whether each observation is above or below the reference value (or whether each pair increased or

decreased).

4.40.2 Prerequisites

Binomial distribution, paired data.

4.40.3 Theory

For iid $X_1, \ldots, X_n$ and a hypothesised median m_0, let $S = \#\{X_i > m_0\}$. Under H_0: median $= m_0$,

$$S \sim \text{Binomial}(n', 0.5),$$

where n' excludes any $X_i = m_0$. The two-sided p-value is computed from the binomial PMF.

For paired data, apply the test to the signs of the paired differences.

The sign test has lower power than the Wilcoxon signed-rank when the symmetry assumption of the latter holds. Its advantage is that it works under **any** continuous distribution.

4.40.4 Assumptions

- iid observations (or independent pairs).
- Continuous distribution (else need to handle ties at the reference).

No symmetry, no shape assumption.

4.40.5 R Implementation

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

# One-sample: test whether median equals 100
x ← rlnorm(30, meanlog = log(105), sdlog = 0.3)
SIGN.test(x, md = 100)

# Paired: test whether intervention changes a score
pre  ← rpois(25, 5)
post ← pre + sample(-1:3, 25, replace = TRUE)
SIGN.test(post, pre, md = 0)
```

4.40.6 Output & Results

```
        One-sample Sign-Test
data:  x
s = 21, p-value = 0.021
alternative hypothesis: true median is not equal to 100
95 percent confidence interval: 97.5 to 120

        Dependent-samples Sign-Test
data:  post and pre
S = 19, p-value = 0.016
95 percent CI for median difference: 0.5 to 2.0
```

Both tests reject: the one-sample median significantly exceeds 100; paired differences are significantly positive.

4.40.7 Interpretation

“The median post-intervention score was 7.0 (95 % CI 0.5 to 2.0 for the paired difference), significantly higher than pre (sign test, 19 of 25 positive differences, exact binomial p = 0.016).”

4.40.8 Practical Tips

- Use the sign test when the distribution of paired differences is badly skewed or clearly not symmetric.
- Report the number of positive, negative, and zero differences; the latter are excluded from the test.
- The sign test’s 95 % CI for the median is a distribution-free interval based on order statistics.
- With many ties at the reference, consider the Wilcoxon or rank-based alternative, understanding their stronger assumptions.
- Power is lower than Wilcoxon signed-rank; prefer Wilcoxon when its symmetry assumption is plausible.

4.40.9 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.

4.40.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test

- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.41 Spearman Rank Correlation Test

4.41.1 Introduction

Spearman's ρ (rho) is the Pearson correlation of the ranks of two variables. It detects monotonic associations that need not be linear, and is robust to outliers and heavy tails. It is the default correlation measure for ordinal data or strongly skewed continuous data.

4.41.2 Prerequisites

Pearson correlation, ranks.

4.41.3 Theory

Replace each X_i by its rank R_i, and each Y_i by its rank S_i. Spearman's ρ is

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \qquad d_i = R_i - S_i,$$

when there are no ties. With ties, the general form

$$\rho = \frac{\sum (R_i - \bar{R})(S_i - \bar{S})}{\sqrt{\sum (R_i - \bar{R})^2 \sum (S_i - \bar{S})^2}}$$

applies.

Under $H_0 : \rho = 0$, an asymptotic normal approximation holds; for small n exact or permutation distributions are used. Confidence intervals for Spearman's ρ can be obtained by Fisher's z applied to the ranks.

4.41.4 Assumptions

- Ordinal or continuous data; the ranks must be well-defined.
- Monotonic association (linear or non-linear).
- Independent pairs.

4.41.5 R Implementation

```r
set.seed(2026)

n ← 50
x ← rgamma(n, shape = 2, rate = 0.5)      # skewed
y ← x^2 + rnorm(n, sd = 4)                # non-linear monotonic

# Pearson (fails for non-linear)
cor.test(x, y, method = "pearson")

# Spearman
cor.test(x, y, method = "spearman", exact = FALSE)

# Bootstrap 95% CI for Spearman's rho
library(boot)
rho_boot ← boot(data = data.frame(x, y),
                statistic = function(d, i) cor(d$x[i], d$y[i], method = "spearman"),
                R = 2000)
boot.ci(rho_boot, type = "perc")
```

4.41.6 Output & Results

```
        Pearson's product-moment correlation
t = 11.45, df = 48, p < 0.001
cor: 0.855

        Spearman's rank correlation rho
S = 100, p < 0.001
rho: 0.995

Bootstrap 95% CI (percentile): (0.988, 0.999)
```

Spearman recovers the strong monotonic relationship ($\rho \approx 1$) that Pearson underestimates because of the non-linearity.

4.41.7 Interpretation

"There was a strong monotonic association between x and y (Spearman's $\rho = 0.99$, bootstrap 95 % CI 0.99 to 1.00; $p < 0.001$)."

4.41.8 Practical Tips

- Use Spearman for ordinal data, heavy-tailed or skewed continuous data, or suspected non-linear monotonic relationships.
- For very small n, use exact methods (`cor.test(..., exact = TRUE)`; only valid without ties).
- With ties, the normal approximation is used; the warning is expected and safe.
- Kendall's τ is an alternative for small samples or many ties.
- Spearman captures monotonic, not linear, associations; for U-shaped relationships it can still miss dependence.

4.41.9 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.

4.41.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.42 Two-Proportion Test

4.42.1 Introduction

The two-proportion test compares success rates in two independent groups. Typical applications: treatment vs. control event rates, exposed vs. unexposed incidence, screen-positive rates across populations. The test comes in two algebraically equivalent forms – a z-test on the difference and a 2×2 chi-squared – and is typically reported together with a risk difference, relative risk, or odds ratio.

4.42.2 Prerequisites

One-proportion test, chi-squared distribution.

4.42.3 Theory

For $X_1 \sim \text{Binomial}(n_1, p_1)$ and $X_2 \sim \text{Binomial}(n_2, p_2)$ independent,

$$H_0 : p_1 = p_2 \quad \text{vs.} \quad H_1 : p_1 \neq p_2.$$

Pooled proportion under H_0: $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$.

Z-test:

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}.$$

Z^2 equals the chi-squared statistic on the 2×2 table. In R, `prop.test()` performs this test (using chi-squared form).

Effect-size summaries:

- Risk difference (RD): $\hat{p}_1 - \hat{p}_2$.
- Relative risk (RR): $\hat{p}_1 / \hat{p}_2$.
- Odds ratio (OR): $\hat{p}_1(1-\hat{p}_2)/[\hat{p}_2(1-\hat{p}_1)]$.

Each has a standard-error formula for constructing Wald CIs; score-based and exact alternatives exist.

4.42.4 Assumptions

- Independent Bernoulli trials within and between groups.
- Expected counts ≥ 5 in every cell (else use Fisher's exact test).

4.42.5 R Implementation

```r
library(epitools)

x1 ← 48; n1 ← 200       # treatment
x2 ← 32; n2 ← 200       # control

prop.test(c(x1, x2), c(n1, n2), correct = FALSE)

# Risk difference with 95% CI via prop.test
prop.test(c(x1, x2), c(n1, n2), correct = FALSE)$conf.int

# Relative risk and odds ratio
riskratio(c(n2 - x2, n1 - x1, x2, x1), rev = "both")$measure
oddsratio(c(n2 - x2, n1 - x1, x2, x1), rev = "both")$measure

# Fisher's exact test (small-cell alternative)
fisher.test(matrix(c(x1, n1 - x1, x2, n2 - x2), nrow = 2, byrow = TRUE))
```

4.42.6 Output & Results

```
     2-sample test for equality of proportions without continuity correction
X-squared = 3.66, df = 1, p-value = 0.0557
95% CI for (p1 - p2): (-0.0014, 0.1614)
sample estimates: prop 1 = 0.24, prop 2 = 0.16

           estimate lower    upper
Exposed        1.50 0.994     2.265
              (relative risk 1.50)
```

Treatment success rate 24 %, control 16 %. Risk difference 8 % (95 % CI -0.1 % to 16.1 %); p = 0.056 (borderline). Relative risk 1.50 (95 % CI 0.99 to 2.27).

4.42.7 Interpretation

Report counts, proportions, the effect-size estimate with its CI, and the test statistic. "In the treatment arm, 48/200 (24.0 %) experienced the event, compared to 32/200 (16.0 %) in control (risk difference 8.0 %, 95 % CI -0.1 % to 16.1 %; $\chi^2(1) = 3.66$, p = 0.056)."

4.42.8 Practical Tips

- Report at least one of RD, RR, or OR as an effect size; significance alone is unhelpful.
- Use Fisher's exact test when expected counts are below 5.
- For matched/paired binary data, use McNemar's test, not the two-proportion test.
- The chi-squared test and the two-proportion z-test give identical two-sided p-values when `correct = FALSE`.
- Be aware that RR and OR differ in magnitude when outcomes are common ($p > 0.10$); choose the scale that fits the clinical question.

4.42.9 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.

4.42.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.43 Two-Sample t-Test

4.43.1 Introduction

The two-sample t-test is the default procedure for deciding whether the means of two independent groups differ. In its modern form it comes in two flavours: Student's t-test, which assumes equal variances in the two populations, and Welch's t-test, which does not. Both produce a test statistic, a p-value, and a confidence interval for the difference in means.

4.43.2 Prerequisites

A reader should understand the difference between a population mean and a sample mean, the concept of a sampling distribution, and the logic of hypothesis testing. Familiarity with R’s formula interface is assumed.

4.43.3 Theory

Let $X_1, \dots, X_{n_1}$ be independent observations from $\mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_1, \dots, Y_{n_2}$ independent observations from $\mathcal{N}(\mu_2, \sigma_2^2)$. We test $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$.

Student’s test assumes $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and uses the pooled variance estimator. The test statistic

$$T = \frac{\bar{X} - \bar{Y}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

follows a t distribution with $n_1 + n_2 - 2$ degrees of freedom under H_0, where s_p^2 is the pooled sample variance.

Welch’s test allows $\sigma_1^2 \neq \sigma_2^2$ and uses the Satterthwaite approximation:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}},$$

with degrees of freedom

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}.$$

Modern recommendations favour Welch’s test as the default, because the assumption of equal variances is rarely verifiable in practice and the loss of efficiency when the variances *are* equal is small.

4.43.4 Assumptions

Both tests assume:

1. Observations within each group are independent of each other.
2. The two groups are independent of each other.
3. The observations in each group are approximately normal, or the sample size is large enough for the CLT to apply to the sample means.

Student’s additionally assumes equal variances. Welch’s does not.

4.43.5 R Implementation

```r
library(effectsize)

set.seed(42)
group_a ← rnorm(30, mean = 100, sd = 15)
group_b ← rnorm(35, mean = 110, sd = 18)

df ← data.frame(
  score = c(group_a, group_b),
  group = factor(rep(c("A", "B"), c(30, 35)))
)

t.test(score ~ group, data = df)

cohens_d(score ~ group, data = df)
```

The default `t.test()` in R performs Welch's test. For Student's, set `var.equal = TRUE`. The `effectsize::cohens_d()` function returns Cohen's d with a confidence interval.

4.43.6 Output & Results

Typical output:

```
        Welch Two Sample t-test

data:  score by group
t = -2.3415, df = 61.832, p-value = 0.02241
alternative hypothesis: true difference in means between group A and group B is not
equal to 0
95 percent confidence interval:
 -18.8912  -1.4872
sample estimates:
mean in group A mean in group B
       100.4510        110.6407
```

Cohen's d is approximately -0.59 with 95% CI $(-1.09, -0.09)$ – a moderate effect.

4.43.7 Interpretation

The conventional report is: "group B had a mean score of 110.6 (SD 18.2), which was significantly higher than group A's mean of 100.5 (SD 14.5); Welch's $t(61.8) = -2.34$, $p = 0.022$, mean difference $= 10.2$ (95% CI 1.5 to 18.9), Cohen's $d = -0.59$." The effect-size reporting is essential: a p-value alone does not tell the reader how large the difference is.

4.43.8 Practical Tips

- Always report the effect size and its confidence interval alongside the p-value; journals increasingly require this.
- Prefer Welch's test as the default; only use Student's when equal variances are genuinely plausible (for example, a randomised experiment with balanced groups).

- Check normality with Q-Q plots rather than with Shapiro-Wilk – the latter has little power at small n and rejects trivially at large n.
- For strongly skewed or outlier-contaminated data, consider the Wilcoxon rank-sum test, a permutation test, or analysis on a log scale.
- Unequal sample sizes are fine; extreme imbalance (say, 10 vs. 1000) warrants extra care because variance assumptions and power are both affected.

4.43.9 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.

4.43.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.44 Two-Way ANOVA

4.44.1 Introduction

Two-way ANOVA tests the effects of two categorical factors on a continuous outcome, plus their interaction. It generalises one-way ANOVA by accommodating designs that manipulate or measure two conditions simultaneously, and is the simplest example of a factorial design.

4.44.2 Prerequisites

One-way ANOVA, factorial designs.

4.44.3 Theory

For a two-way design with factors A and B:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk},$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. The ANOVA partitions the total sum of squares into:

- Main effect of A (df = a - 1).
- Main effect of B (df = b - 1).
- Interaction AxB (df = (a - 1)(b - 1)).
- Residual (df = ab(n - 1) for balanced design with n per cell).

Three null hypotheses are tested: no main A effect, no main B effect, no interaction. The F-tests use the residual MS as denominator.

Type I, II, III sums of squares differ when the design is unbalanced:

- Type I: sequential; order-dependent.
- Type II: each effect adjusted for others except interactions involving it.
- Type III: each effect adjusted for all others; default in most modern practice when interactions are present.

4.44.4 Assumptions

- Independent observations.
- Normal residuals.
- Homogeneous variances across cells.
- For balanced designs all SS types agree; for unbalanced ones, the choice matters.

4.44.5 R Implementation

```
library(car); library(effectsize); library(emmeans)
set.seed(2026)

df ← expand.grid(A = factor(c("A1", "A2")),
                 B = factor(c("B1", "B2", "B3")),
                 rep = 1:20)
df$y ← with(df, 50 + 3*(A == "A2") + 2*(B == "B2") + 4*(B == "B3") +
                5*(A == "A2" & B == "B3") + rnorm(nrow(df), 0, 6))

# Fit and Type III ANOVA
fit ← lm(y ~ A * B, data = df, contrasts = list(A = contr.sum, B = contr.sum))
Anova(fit, type = 3)

# Effect sizes
eta_squared(fit, partial = TRUE)

# Marginal means and interaction plot
emm ← emmeans(fit, ~ A * B)
emm
pairs(emm, simple = "A")

# Simple interaction plot
interaction.plot(df$B, df$A, df$y, col = c("#2A9D8F", "#F4A261"),
                 lwd = 2, main = "Interaction plot")
```

4.44.6 Output & Results

```
Anova Table (Type III tests)

             Sum Sq  Df  F value  Pr(>F)
(Intercept) 291 000  1  8036    <2e-16
A             168    1   4.64  0.0332
B             490    2   6.77  0.0017
A:B           186    2   2.57  0.0819
Residuals   4 123 114

Eta2 (partial) | 95% CI
-------------------------
A              | 0.04
B              | 0.11
A:B            | 0.04
```

Main effects of A and B are significant; the interaction is borderline.

4.44.7 Interpretation

“In a two-way ANOVA, both A (F(1, 114) = 4.64, p = 0.033) and B (F(2, 114) = 6.77, p = 0.002) had significant main effects; the A x B interaction was not significant (F(2, 114) = 2.57, p = 0.082).”

4.44.8 Practical Tips

- Use Type III SS with sum-to-zero contrasts for standard interpretation.
- Always inspect the interaction plot; visual patterns often matter more than p-values.
- Unbalanced designs need careful contrast specification; `car::Anova` handles the SS types explicitly.
- Report effect sizes (partial η^2) alongside F.
- For significant interactions, interpret simple effects, not the main effects alone.

4.44.9 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.

4.44.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests

- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.45 Type I and Type II Errors

4.45.1 Introduction

Two kinds of error arise in hypothesis testing: rejecting a true null (Type I) and failing to reject a false null (Type II). Their probabilities – α and β – are the core quantities of statistical decision theory. Power $1 - \beta$ is one minus the Type II rate, and the rate at which sample size and effect size trade against α is the currency of study design.

4.45.2 Prerequisites

Hypothesis testing, p-values, sampling distributions.

4.45.3 Theory

The 2×2 decision table:

	H_0 **true**	H_0 **false**
Reject H_0	Type I (α)	correct (power)
Fail to reject H_0	correct	Type II (β)

- $\alpha = P(\text{reject } H_0 \mid H_0)$ is the significance level, usually set to 0.05.
- $\beta = P(\text{fail to reject } H_0 \mid H_1)$ is the Type II error rate.
- Power $= 1 - \beta$, usually targeted at 0.80 or 0.90.

For a given test, power depends on: the true effect size (larger $\Rightarrow$ higher power), the sample size (larger $\Rightarrow$ higher power), the significance level (α smaller $\Rightarrow$ power lower), and the variability of the outcome (larger SD $\Rightarrow$ lower power).

Trade-off: tightening α (to 0.01, say) reduces Type I errors but increases Type II errors for a fixed sample size. Only increasing n reduces both.

4.45.4 Assumptions

The test is correctly specified; the distributional model holds under H_0 and H_1.

4.45.5 R Implementation

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

# Simulation: Type I rate at alpha = 0.05 when H0 true
n_sims <- 10000
p_null <- replicate(n_sims, t.test(rnorm(30, 0, 1), mu = 0)$p.value)
mean(p_null < 0.05)

# Power at alpha = 0.05 when H1 true (effect size = 0.5)
p_alt <- replicate(n_sims, t.test(rnorm(30, 0.5, 1), mu = 0)$p.value)
mean(p_alt < 0.05)

# Compare to theoretical power
power.t.test(n = 30, delta = 0.5, sd = 1, type = "one.sample")

# Power curve
d_grid <- seq(0, 1, by = 0.05)
pw <- sapply(d_grid, function(d)
  power.t.test(n = 30, delta = d, sd = 1, type = "one.sample")$power)

ggplot(data.frame(d = d_grid, power = pw), aes(d, power)) +
  geom_line(colour = "#2A9D8F", linewidth = 1) +
  geom_hline(yintercept = 0.8, linetype = "dashed", colour = "#F4A261") +
  labs(x = "Effect size d", y = "Power",
       title = "Power curve at alpha = 0.05, n = 30") +
  theme_minimal()
```

4.45.6 Output & Results

```
[1] 0.0510    # empirical Type I rate
[1] 0.7549    # empirical power at d = 0.5
```

Empirical Type I error is close to the nominal 0.05; empirical power at $d = 0.5$ matches the theoretical calculation.

4.45.7 Interpretation

In study design, sample size is chosen to achieve a pre-specified power at the minimum effect size of interest. Reporting post-hoc power based on observed effects is misleading and discouraged.

4.45.8 Practical Tips

- Power at 80 % is the conventional minimum; 90 % for confirmatory trials.
- Type I and II errors are fixed by the design, not by the data; a significant p-value does not “justify” the design after the fact.
- In multiple-comparison settings, family-wise error rate replaces single-test α.
- Sequential designs (group-sequential trials) adjust α across looks to preserve overall Type I rate.
- Heavy-tailed or non-normal data reduce power; use rank-based tests or robust alternatives when applicable.

4.45.9 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.

4.45.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.46 Welch's t-Test

4.46.1 Introduction

Welch's t-test is the standard two-sample mean-comparison test when variances across groups may differ. R's `t.test()` defaults to Welch (`var.equal = FALSE`). Decades of simulation evidence favour Welch as the default choice regardless of whether variances appear equal; the efficiency cost versus the equal-variance (Student) test is negligible when variances are equal, and the robustness gain is substantial when they are not.

4.46.2 Prerequisites

Two-sample t-test, variance, degrees of freedom.

4.46.3 Theory

Given independent samples of sizes n_1, n_2 with means $\bar{X}_1, \bar{X}_2$ and sample variances s_1^2, s_2^2, the Welch test statistic is

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

The Welch-Satterthwaite approximation gives the degrees of freedom

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

ν is generally not an integer. Under $H_0 : \mu_1 = \mu_2$, $T \sim t_\nu$ approximately.

4.46.4 Assumptions

- Independent observations within and across groups.
- Approximately normal observations (or large n).
- No extreme outliers.

Equal variances are **not** required.

4.46.5 R Implementation

```r
library(effectsize); library(broom)
set.seed(2026)

n1 ← 25; n2 ← 40
group1 ← rnorm(n1, mean = 100, sd = 8)
group2 ← rnorm(n2, mean = 106, sd = 18)   # larger variance

df ← data.frame(
  x     = c(group1, group2),
  group = factor(rep(c("A", "B"), c(n1, n2)))
)

# Welch (default)
t.test(x ~ group, data = df) ▷ tidy()

# Student (var.equal = TRUE) for comparison
t.test(x ~ group, data = df, var.equal = TRUE) ▷ tidy()

# Effect size
cohens_d(x ~ group, data = df)
```

4.46.6 Output & Results

```
# Welch
estimate1 = 100.8, estimate2 = 106.7, t = -1.77, df = 58.9, p = 0.082

# Student (for comparison)
estimate1 = 100.8, estimate2 = 106.7, t = -1.56, df = 63,  p = 0.124

Cohen's d | 95% CI
--------------------
-0.40     | [-0.93, 0.14]
```

Welch's df (58.9) is smaller than Student's (63) because the variances differ; the Welch p-value is consequently smaller in this case, though both fail to reject.

4.46.7 Interpretation

Reporting "Welch's t(58.9) = -1.77, p = 0.082" makes the Welch variant explicit. Cohen's d uses a pooled SD by default; for very unequal variances the "Glass's delta" alternative uses only one group's SD.

4.46.8 Practical Tips

- Do not run Levene's test to decide between Welch and Student; a conditional choice inflates Type I error. Default to Welch.
- Satterthwaite df can be fractional; report it as a number with decimals.
- Welch's test is slightly more conservative than Student when variances are equal; the cost is negligible.
- For rank-based alternatives, use Mann-Whitney U.
- Welch-style corrections exist for ANOVA (Welch's F) and linear models (heteroscedasticity-consistent SEs).

4.46.9 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.

4.46.10 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.47 Wilcoxon Signed-Rank Test

4.47.1 Introduction

The Wilcoxon signed-rank test compares two paired measurements without requiring Normality of their differences. It is the rank-based non-parametric analogue of the paired t -test and is the appropriate choice when the paired differences are ordinal, when the difference distribution is skewed or heavy-tailed, or when sample sizes are too small for the parametric test's distributional assumptions to be defensible. The Wilcoxon signed-

rank test is widely used in pre-post studies, in within-subject A/B comparisons, in sensory evaluation panels, and in any setting where paired measurements are compared and Normality is in doubt.

4.47.2 Prerequisites

A working understanding of paired-sample inference, the rank-based testing framework, and the assumption of symmetric distribution of paired differences (rather than Normal or any specific shape).

4.47.3 Theory

For paired observations (X_i, Y_i), compute differences $D_i = X_i - Y_i$, drop any zero differences, rank the absolute values $|D_i|$, and sum the ranks assigned to positive differences (W_+) and to negative differences (W_-). Under the null hypothesis that the median of D is zero (assuming the distribution of D is symmetric), W_+ and W_- have a known discrete distribution; for large n, the standardised statistic is approximately Normal. The rank-biserial correlation $r_{rb} = (W_+ - W_-)/T$, with T the sum of all ranks, provides an effect-size measure on $[-1, 1]$.

If the symmetry assumption fails substantively, the sign test offers a more assumption-light alternative at the cost of reduced power.

4.47.4 Assumptions

The observations are genuinely paired, the distribution of paired differences is approximately symmetric around the median, and the outcomes are at least ordinal. The test does not assume Normality of the differences — only symmetry — which is its principal advantage over the paired t-test.

4.47.5 R Implementation

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

n ← 25
pre  ← sample(1:7, n, replace = TRUE)
post ← pre + sample(-1:2, n, replace = TRUE, prob = c(0.1, 0.2, 0.4, 0.3))

wilcox.test(pre, post, paired = TRUE)
rank_biserial(pre, post, paired = TRUE)

c(median_diff = median(post - pre),
  IQR_diff    = IQR(post - pre))
```

4.47.6 Output & Results

The pre-post comparison on a 7-point ordinal scale gives the Wilcoxon signed-rank statistic $V = 35$ ($p = 0.003$), with a paired rank-biserial correlation of -0.72 (95 % CI -0.90 to -0.40) classed as a large effect. The median paired difference of $+1.0$ (IQR 0 to

1) characterises the central tendency of the change on the original scale, and reporting both the test statistic and the effect-size summary is the standard expected by modern guidelines.

4.47.7 Interpretation

A reporting sentence: "Scores increased significantly from pre-treatment to post-treatment (Wilcoxon signed-rank $V = 35$, $p = 0.003$); median paired difference was $+1.0$ point on the 7-point scale (IQR 0 to 1, range -1 to $+3$), and the paired rank-biserial correlation was -0.72 (95 % CI -0.90 to -0.40), classed as a large effect. The non-parametric test was chosen because the differences were ordinal and a parametric paired t-test would not be appropriate." Always report median, IQR, and an effect-size measure.

4.47.8 Practical Tips

- Report the median and IQR of the paired differences rather than the raw pre/post means; the median is the location parameter that the Wilcoxon signed-rank test inferentially targets, and the means are not directly relevant to the rank-based test.
- Handle zero differences as R does by default (drop them from the analysis), or use the Pratt convention (include them in the ranking but assign sign zero); the choice can matter for small samples and should be documented.
- For small n and no ties in the absolute differences, use `exact = TRUE` to obtain exact p-values rather than the Normal approximation; the exact distribution is the more accurate inference at small samples.
- The symmetry-of-differences assumption can be checked informally with a histogram of D_i or a Q-Q plot of D_i against the Normal; if the differences are markedly skewed, the sign test is a more assumption-light alternative.
- For paired binary outcomes (yes/no observed twice on the same subject), McNemar's test — not the Wilcoxon signed-rank — is the appropriate paired-sample test; conflating the two is a recurring methodological error.
- The Wilcoxon signed-rank test is sensitive to differences in median location under the symmetry assumption; for general distributional comparisons of paired data, the paired Cliff's delta or a one-sample Kolmogorov-Smirnov test on the differences are more general alternatives.

4.47.9 R Packages Used

Base R `wilcox.test(paired = TRUE)` for the canonical signed-rank test; `effectsize::rank_biserial()` for the paired rank-biserial correlation effect size; `coin::wilcoxsign_test()` for permutation-based exact p-values; `rcompanion::wilcoxonPairedR()` for an alternative effect-size implementation; `BSDA::SIGN.test()` for the more assumption-light sign-test alternative.

4.47.10 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.

4.47.11 See also — labs in this chapter

- Populations, samples, and the central limit theorem
- Bootstrap and permutation tests
- Maximum likelihood estimation
- One-sample t-test and one-proportion test
- Hypothesis testing, p-values, type I/II errors
- Two-sample and paired t-tests
- Two proportions, chi-square, risk and odds
- Pearson and Spearman correlation
- Non-parametric tests

4.48 Week 3, Session 2 — Bootstrap and permutation tests

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.48.1 Learning objectives

- Compute a bootstrap confidence interval for a statistic whose sampling distribution is not easily derived.
- Run a permutation test for a two-group comparison and state what null hypothesis it tests.
- Explain when bootstrap and permutation answer different questions.

4.48.2 Prerequisites

Lab 3.1.

4.48.3 Background

Classical inference writes down the sampling distribution of an estimator analytically, usually by invoking the central limit theorem and assuming a parametric family for the data. Resampling methods replace this step with computation. A bootstrap confidence interval for a statistic is built by drawing repeated samples *with replacement* from the observed data and recomputing the statistic; the empirical distribution of the recomputed values is treated as a proxy for the sampling distribution.

A permutation test answers a different question. It randomly reassigns the group labels on the observed data and recomputes the test statistic; the empirical distribution of the recomputed values is the sampling distribution under the null of *exchangeability* — that is, under a null in which the group labels are interchangeable. It gives a *p*-value without any assumption about the shape of the underlying distributions.

The two techniques look similar — both involve a loop and a `replicate()` — but they are different tools. Bootstrap estimates uncertainty under the observed data-generating process; permutation tests a sharp null of no association.

4.48.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.48.5 1. Hypothesis

Two hypotheses in this lab:

1. For a sample median of body mass in Gentoo penguins, build a 95% bootstrap CI.
2. For flipper length in Gentoo vs Adelie penguins, test the sharp null that the distributions are exchangeable against the two-sided alternative.

4.48.6 2. Visualise

Use `palmerpenguins`.

```
library(palmerpenguins)
dat ← penguins ▷
  filter(species %in% c("Gentoo", "Adelie")) ▷
  drop_na(body_mass_g, flipper_length_mm)
```

```
dat ▷
  ggplot(aes(species, flipper_length_mm, fill = species)) +
  geom_boxplot(alpha = 0.6, colour = "grey30") +
  labs(x = NULL, y = "Flipper length (mm)") +
  theme(legend.position = "none")
```

4.48.7 3. Assumptions

Bootstrap: the observed sample is representative of the population we wish to generalise to. Resampling with replacement preserves the univariate structure but assumes exchangeability of observations.

Permutation: under H0, group labels can be reshuffled without changing the joint distribution. The test does not assume normality or equal variances.

4.48.8 4. Conduct

4.48.8.1 Bootstrap CI for the median body mass of Gentoo

```
gen ← dat ▷ filter(species == "Gentoo") ▷ pull(body_mass_g)
B ← 2000
boot_meds ← replicate(B, median(sample(gen, replace = TRUE)))
ci_med ← quantile(boot_meds, c(0.025, 0.975))
obs_med ← median(gen)
tibble(statistic = "median body mass",
       observed = obs_med,
       ci_low = ci_med[1],
       ci_high = ci_med[2])
```

```
tibble(boot = boot_meds) ▷
  ggplot(aes(boot)) +
  geom_histogram(bins = 40, fill = "grey60", colour = "white") +
  geom_vline(xintercept = obs_med, colour = "firebrick") +
  geom_vline(xintercept = ci_med, linetype = 2) +
  labs(x = "Bootstrap median", y = "Count")
```

4.48.8.2 Permutation test for flipper length Gentoo vs Adelie

```
g ← dat$species
x ← dat$flipper_length_mm
observed_diff ← mean(x[g == "Gentoo"]) - mean(x[g == "Adelie"])

perm_diff ← replicate(5000, {
  gp ← sample(g)
  mean(x[gp == "Gentoo"]) - mean(x[gp == "Adelie"])
})
p_perm ← mean(abs(perm_diff) ⩾ abs(observed_diff))
p_perm
```

```
tibble(perm = perm_diff) ▷
  ggplot(aes(perm)) +
  geom_histogram(bins = 50, fill = "grey60", colour = "white") +
  geom_vline(xintercept = observed_diff, colour = "firebrick", linewidth = 1) +
  labs(x = "Permuted mean difference", y = "Count")
```

4.48.9 5. Concluding statement

> Based on B = 2000 bootstrap resamples, the median body mass in Gentoo penguins (n = `length(gen)`) was `round(obs_med, 0)` g (95% percentile bootstrap CI `round(ci_med[1], 0)` to `round(ci_med[2], 0)` g). Flipper length was much longer in Gentoo than Adelie (mean difference `round(observed_diff, 1)` mm); a permutation test over 5000 reshuffles produced p = `format(p_perm, scientific = FALSE)`, providing strong evidence against exchangeability.

Bootstrap and permutation give you parametric-free tools for estimation and testing respectively. Neither rescues you from a biased sample; both depend on the data at hand being representative.

4.48.10 Common pitfalls

- Using bootstrap SEs for a statistic with heavy tails (eg the maximum) — it does not have a well-defined bootstrap distribution.
- Permutation with more than two groups: you still need a single test statistic (eg *F*), not pairwise comparisons by default.
- Reading a bootstrap CI as a confidence interval for the population parameter when the sample is biased.

4.48.11 Further reading

- Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*.
- Good PI. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*.

4.48.12 Session info

```
sessionInfo()
```

4.48.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.49 Week 3, Session 5 — Hypothesis testing, p-values, type I/II errors

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.49.1 Learning objectives

- State what a *p*-value is, and — just as importantly — what it is not.
- Demonstrate by simulation that *p*-values are uniform under H0.
- Quantify type I error, type II error, and power, and show their relationship.

4.49.2 Prerequisites

Lab 3.4.

4.49.3 Background

A *p*-value is the probability of observing a test statistic at least as extreme as the one computed, under the null hypothesis. It is not the probability that the null is true. It is not the probability that the result was due to chance. Both of those are common, plausible misreadings; both are wrong. The *p*-value is a function of the data under one specific hypothetical: *if H0 were true, how often would we see this much apparent signal?*

Under H0, the *p*-value is uniformly distributed on [0, 1]. This is a direct consequence of the probability integral transform, and it is the reason the familiar threshold 0.05 gives a 5% type I error rate when applied across many independent tests of true nulls.

Type I error is the rate of false positives; type II error is the rate of false negatives. Power is 1 − type II error: the probability of detecting a true effect. Power depends on the effect size, the variability, the sample size, and α. Increasing *n* buys you power; increasing α buys you power at the cost of more false positives.

4.49.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.49.5 1. Hypothesis

The lab has two claims to verify:

1. Under H0 (no effect), the *p*-value from a two-sample *t*-test is uniform on [0, 1].
2. Under H1 (effect present), the probability of rejecting H0 depends on sample size and effect size.

4.49.6 2. Visualise

Generate many datasets with no effect, run a *t*-test on each, and plot the distribution of *p*-values.

```
simulate_pvals ← function(n_per_group = 30, delta = 0, B = 5000) {
  replicate(B, {
    a ← rnorm(n_per_group, 0, 1)
    b ← rnorm(n_per_group, delta, 1)
    t.test(a, b)$p.value
  })
}
p_h0 ← simulate_pvals(30, delta = 0)
```

```
tibble(p = p_h0) ▷
  ggplot(aes(p)) +
  geom_histogram(bins = 20, fill = "grey60", colour = "white") +
  geom_hline(yintercept = length(p_h0) / 20,
             linetype = 2, colour = "firebrick") +
  labs(x = "p-value under H0", y = "Count")
```

Flat histogram. Any other shape is a sign the test is misspecified.

4.49.7 3. Assumptions

The claim that *p*-values are uniform under H0 requires that H0 be *correctly specified* — the distributional assumptions of the test must hold. If the *t*-test is applied to strongly skewed data at small *n*, the histogram will not be flat, and the type I error will depart from

α.

4.49.8 4. Conduct

4.49.8.1 Type I error rate

```
type_I ← mean(p_h0 < 0.05)
type_I
```

Close to the nominal 0.05.

4.49.8.2 Power under H1 (effect size 0.5, *n* per group 30)

```
p_h1 ← simulate_pvals(30, delta = 0.5)
power_obs ← mean(p_h1 < 0.05)
power_obs
```

4.49.8.3 Power by effect size and sample size

```
grid ← expand_grid(
  n = c(10, 20, 50, 100),
  delta = seq(0, 1.2, by = 0.2)
)
grid$power ← with(grid, mapply(function(nn, dd) {
  mean(simulate_pvals(nn, dd, B = 1000) < 0.05)
}, n, delta))
```

```
grid ▷
  ggplot(aes(delta, power, colour = factor(n))) +
  geom_line(linewidth = 1) +
  geom_point() +
  geom_hline(yintercept = 0.8, linetype = 2, colour = "grey40") +
  labs(x = "Effect size (delta)", y = "Power",
       colour = "n per group")
```

4.49.9 5. Concluding statement

> In 5000 simulated datasets with no true effect, the empirical type I error rate of the two-sample *t*-test was `signif(type_I, 2)` (nominal 0.05), and the *p*-value histogram was flat as theory requires. With a true effect δ = 0.5 at *n* = 30 per group, the empirical power was `signif(power_obs, 2)` . Power grew with both effect size and sample size as expected; 80% power at δ = 0.5 required roughly *n* = 65 per group.

A *p*-value is a random variable, uniform under H0. The common bar of 0.05 is a convention, not a discovery about nature, and rejecting results above it is a statement about rates of evidence in a long run of studies — not about the truth of any single study.

4.49.10 Common pitfalls

- Reporting $p = 0.051$ as "not significant" and $p = 0.049$ as "significant" as if the dichotomy were informative.
- Confusing $P(\text{data} \mid H0)$ with $P(H0 \mid \text{data})$.
- Running many tests without adjustment and reporting the smallest p-value.
- Declaring "no effect" when a non-significant p-value comes from an underpowered study.

4.49.11 Further reading

- Wasserstein RL, Lazar NA (2016). *The ASA's Statement on p-Values*.
- Greenland S et al. (2016). *Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations*.

4.49.12 Session info

```
sessionInfo()
```

4.49.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.50 Week 3, Session 3 — Maximum likelihood estimation

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.50.1 Learning objectives

- Write a log-likelihood for a simple model and find its maximum.
- Use Fisher information to obtain an asymptotic standard error and confidence interval.
- Relate the MLE to intuitive estimators (sample mean, sample rate).

4.50.2 Prerequisites

Lab 3.1.

4.50.3 Background

Maximum likelihood is the workhorse of parametric inference. Given a parametric family for the data — normal, Poisson, binomial — the MLE is the parameter value that makes the observed data most probable. Under mild regularity conditions the MLE is consistent (it converges to the truth), asymptotically unbiased, and asymptotically normal with variance given by the inverse of Fisher information.

Fisher information quantifies how sharply the log-likelihood peaks at its maximum. A sharply peaked likelihood gives a precise estimate (small variance); a flat one gives a vague estimate. The standard error of an MLE is the square root of 1/*n* times the inverse of the Fisher information per observation.

In the simple cases we treat here the MLE coincides with the intuitive estimator — the mean for a normal, the sample rate for a Poisson — but the machinery generalises to any likelihood.

4.50.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.50.5 1. Hypothesis

Estimate (i) the mean of a normal and (ii) the rate of a Poisson by maximum likelihood, recover standard errors from Fisher information, and build Wald confidence intervals.

4.50.6 2. Visualise

Normal data with known σ = 2.

```
n ← 50
mu_true ← 5
sigma ← 2
x ← rnorm(n, mean = mu_true, sd = sigma)
```

```
grid ← tibble(mu = seq(3.5, 6.5, length.out = 300)) ▷
  mutate(loglik = sapply(mu, function(m) sum(dnorm(x, m, sigma, log = TRUE))))
ggplot(grid, aes(mu, loglik)) +
  geom_line() +
  geom_vline(xintercept = mean(x), linetype = 2, colour = "firebrick") +
  labs(x = expression(mu), y = "log-likelihood")
```

The likelihood is peaked at the sample mean — that is the MLE.

4.50.7 3. Assumptions

Normal model: iid, constant σ. Poisson model: iid counts, constant rate over a fixed exposure. Fisher information is evaluated *at the MLE* for the Wald standard error.

4.50.8 4. Conduct

4.50.8.1 Normal mean by MLE

```
mle_mu ← mean(x)
# Fisher information for mu in normal with known sigma is n / sigma^2.
I_mu ← n / sigma^2
se_mu ← 1 / sqrt(I_mu)
ci_mu ← mle_mu + c(-1, 1) * qnorm(0.975) * se_mu
tibble(mle = mle_mu, se = se_mu,
       ci_low = ci_mu[1], ci_high = ci_mu[2])
```

4.50.8.2 Normal mean by numeric optimisation (sanity check)

```
nll ← function(mu) -sum(dnorm(x, mu, sigma, log = TRUE))
opt ← optim(0, nll, method = "Brent", lower = 0, upper = 10, hessian = TRUE)
opt$par        # MLE
sqrt(1 / opt$hessian[1, 1])  # SE from observed information
```

4.50.8.3 Poisson rate by MLE

```
y ← rpois(100, lambda = 3.2)
mle_lambda ← mean(y)
# Fisher information for lambda: n / lambda. At MLE use 1/mle_lambda per obs.
I_lambda ← length(y) / mle_lambda
se_lambda ← 1 / sqrt(I_lambda)
ci_lambda ← mle_lambda + c(-1, 1) * qnorm(0.975) * se_lambda
tibble(mle = mle_lambda, se = se_lambda,
       ci_low = ci_lambda[1], ci_high = ci_lambda[2])
```

4.50.8.4 Likelihood surface for a two-parameter model (μ, σ)

```
surf ← expand_grid(
  mu = seq(3.5, 6.5, length.out = 80),
  s  = seq(1, 3.5, length.out = 80)
) ▷
  rowwise() ▷
  mutate(ll = sum(dnorm(x, mu, s, log = TRUE))) ▷
  ungroup()
```

```
ggplot(surf, aes(mu, s, fill = ll)) +
  geom_tile() +
  geom_contour(aes(z = ll), colour = "white") +
  scale_fill_viridis_c() +
  labs(x = expression(mu), y = expression(sigma), fill = "log-L")
```

4.50.9 5. Concluding statement

> The MLE of the normal mean from $n =$ `n` observations at $\sigma = 2$ was `round(mle_mu, 3)` (Fisher-information SE = `round(se_mu, 3)`; 95% Wald CI `round(ci_mu[1], 2)` to `round(ci_mu[2], 2)`), recovering the true $\mu = 5$. The MLE of the Poisson rate from $n = 100$ observations was `round(mle_lambda, 3)` (SE `round(se_lambda, 3)`; CI `round(ci_lambda[1], 2)` to `round(ci_lambda[2], 2)`). Observed-information SEs from `optim` agreed with the expected-information calculation.

The MLE is the general-purpose engine behind linear regression, logistic regression, Cox models, and the bulk of parametric inference in this course. The machinery — likelihood, information, SE — is the same in each case.

4.50.10 Common pitfalls

- Reporting a Wald CI when the sample is small. Profile CIs are more accurate in that regime.
- Evaluating Fisher information at a wrong value of the parameter. Use the MLE.
- Forgetting that the MLE is not always unbiased for finite n — only asymptotically.

4.50.11 Further reading

- Casella G, Berger RL. *Statistical Inference*, chapter on point estimation.
- Cox DR. *Principles of Statistical Inference*.

4.50.12 Session info

```
sessionInfo()
```

4.50.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.51 Week 4, Session 4 — Non-parametric tests

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.51.1 Learning objectives

- Choose the right rank-based test for a given design — Wilcoxon signed-rank, Mann-Whitney *U*, Kruskal-Wallis, sign test.
- Run each in base R and interpret the output.
- State what null hypothesis each test is really addressing (not always “medians are equal”).

4.51.2 Prerequisites

Labs 3.4, 4.1.

4.51.3 Background

Non-parametric tests make weaker distributional assumptions than the parametric tests that dominate the preceding labs. They tend to test *exchangeability* of distributions rather than equality of means, and they operate on ranks of the data rather than on the raw values. When the data are heavily skewed, ordinal, or small-sample, the non-parametric alternatives are often the defensible default.

The workhorse set: Wilcoxon signed-rank for paired continuous data; Mann-Whitney *U* (the `wilcox.test` with `paired = FALSE`) for two independent groups; Kruskal-Wallis for three or more groups; sign test as a cruder paired-data option based on the signs of differences alone.

A common mistake is to read these tests as "comparing medians". They compare *distributions*. Under additional assumptions — notably that the two groups' distributions differ only by a location shift — the Mann-Whitney *U* test does estimate a location shift, but that is a stronger assumption than the test itself requires.

4.51.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.51.5 1. Hypothesis

Three scenarios:

A. Paired: pre/post in 25 patients, skewed outcome. H0: no shift. B. Independent two-sample: outcome in two arms, skewed. H0: same distribution. C. Three-group comparison: outcome across three centres. H0: same distribution across all three.

4.51.6 2. Visualise

```
# Scenario A: paired, lognormal-ish
n_a ← 25
pre ← exp(rnorm(n_a, 3, 0.3))
post ← pre * exp(rnorm(n_a, -0.2, 0.3))
df_a ← tibble(id = seq_len(n_a), pre, post,
              delta = post - pre)

# Scenario B: independent two groups, lognormal
n_b ← 30
grp ← rep(c("A", "B"), each = n_b)
y_b ← c(exp(rnorm(n_b, 3, 0.5)),
        exp(rnorm(n_b, 3.4, 0.5)))
df_b ← tibble(grp, y = y_b)

# Scenario C: three groups
n_c ← 25
centre ← rep(c("X", "Y", "Z"), each = n_c)
y_c ← c(exp(rnorm(n_c, 3.0, 0.4)),
        exp(rnorm(n_c, 3.2, 0.4)),
        exp(rnorm(n_c, 3.5, 0.4)))
df_c ← tibble(centre, y = y_c)
```

```
df_b ▷
  ggplot(aes(grp, y, fill = grp)) +
  geom_boxplot(alpha = 0.6, colour = "grey30") +
  labs(x = NULL, y = "Outcome") +
  theme(legend.position = "none")
```

4.51.7 3. Assumptions

Wilcoxon signed-rank: paired observations; differences symmetric around the null. Mann-Whitney *U*: independent observations within and between groups; if you want a location-shift interpretation, the two distributions should have the same shape. Kruskal-Wallis: independent observations; same-shape assumption extends to all *k* groups.

4.51.8 4. Conduct

```
# A: Wilcoxon signed-rank
wt_a ← wilcox.test(df_a$post, df_a$pre, paired = TRUE,
                   conf.int = TRUE)
# Sign test: a simple binomial test on positive differences
signs ← sum(df_a$delta > 0)
sign_test ← binom.test(signs, n_a, p = 0.5)

# B: Mann-Whitney U
wt_b ← wilcox.test(y ~ grp, data = df_b, conf.int = TRUE)

# C: Kruskal-Wallis
kw_c ← kruskal.test(y ~ centre, data = df_c)

list(wilcox_paired = wt_a,
     sign_test = sign_test,
     mann_whitney = wt_b,
     kruskal_wallis = kw_c)
```

Follow-up pairwise comparisons with Holm correction for C:

```
pairwise.wilcox.test(df_c$y, df_c$centre, p.adjust.method = "holm")
```

4.51.9 5. Concluding statement

> **Paired comparison (A).** 25 patients, median change = `round(median(df_a$delta), 2)`; Wilcoxon signed-rank p = `signif(wt_a$p.value, 2)`, 95% CI on the Hodges-Lehmann estimate `round(wt_a$conf.int[1], 2)` to `round(wt_a$conf.int[2], 2)`. Sign test p = `signif(sign_test$p.value, 2)`.
>
> **Two-group (B).** Mann-Whitney p = `signif(wt_b$p.value, 2)`, 95% CI `round(wt_b$conf.int[1], 2)` to `round(wt_b$conf.int[2], 2)`.
>
> **Three-group (C).** Kruskal-Wallis χ^2 = `round(kw_c$statistic, 2)`, df = `kw_c$parameter`, p = `signif(kw_c$p.value, 2)`. Pairwise Wilcoxon tests with Holm correction indicated the signal lay in the X vs Z contrast.

Non-parametric tests are not free lunches. They trade a modest amount of power (about 5% vs a *t*-test when the *t*-test's assumptions hold) for robustness to distributional misspecification.

4.51.10 Common pitfalls

- Reading a Wilcoxon as a test of medians. It is a test of distributional shift.
- Using a rank test when the data are fine on the raw scale, and sacrificing power unnecessarily.
- Forgetting to adjust for multiplicity after a Kruskal-Wallis.

4.51.11 Further reading

- Hollander M, Wolfe DA. *Nonparametric Statistical Methods*.
- Conover WJ. *Practical Nonparametric Statistics*.

4.51.12 Session info

```
sessionInfo()
```

4.51.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.52 Week 3, Session 4 — One-sample t-test and one-

proportion test

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.52.1 Learning objectives

- State a one-sample null and alternative hypothesis for a continuous outcome and for a proportion.
- Run `t.test(mu = )` and `prop.test()` / `binom.test()` and interpret the output.
- Apply the full five-step inference template in one working pass.

4.52.2 Prerequisites

Labs 2.5, 3.3.

4.52.3 Background

The one-sample *t*-test asks whether the mean of a continuous outcome differs from a pre-specified value μ_0. It is the inference analogue of Table 1: a single-column calculation asking whether the sample is consistent with a named population mean. The test statistic is $(\bar{x} - \mu_0) / (s/\sqrt{n})$, compared to a *t* distribution on $n - 1$ degrees of freedom.

The one-proportion test does the same for a binary outcome. When the sample is large we use `prop.test()`, which is a Wald-type test based on the normal approximation to the binomial; when the sample is small, `binom.test()` gives an exact test based on the binomial PMF. Both return a *p*-value, a point estimate, and a CI.

These are the simplest tests in the course, but the five-step template — hypothesis, visualise, assumptions, conduct, conclude — is the same template used in every later inference lab. Getting the template into muscle memory here pays dividends throughout the rest of the course.

4.52.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.52.5 1. Hypothesis

Scenario A (continuous): fasting blood glucose in a random sample of `n = 40` patients from a clinic. Reference mean $\mu_0 = 95$ mg/dL in a healthy population. Two-sided test.

- H0: $\mu = 95$
- H1: $\mu \neq 95$
- $\alpha = 0.05$

Scenario B (proportion): of 120 patients offered a new screening programme, $k = 78$ accepted. Reference acceptance rate $\pi_0 = 0.60$.

- H0: $\pi = 0.60$
- H1: $\pi \neq 0.60$
- $\alpha = 0.05$

4.52.6 2. Visualise

```
n ← 40
glucose ← rnorm(n, mean = 100, sd = 12)
```

```
tibble(glucose) ▷
  ggplot(aes(glucose)) +
  geom_histogram(bins = 15, fill = "grey60", colour = "white") +
  geom_vline(xintercept = 95, linetype = 2) +
  geom_vline(xintercept = mean(glucose), colour = "firebrick") +
  labs(x = "Fasting glucose (mg/dL)", y = "Count")
```

```
accept ← 78
total  ← 120
tibble(state = c("accepted", "declined"),
       count = c(accept, total - accept))
```

4.52.7 3. Assumptions

t-test: approximate normality of the sample mean, which — by the central limit theorem — holds for $n = 40$ unless the underlying distribution is strongly skewed. Check with a Q-Q plot.

```
tibble(glucose) ▷
  ggplot(aes(sample = glucose)) +
  stat_qq() + stat_qq_line(colour = "firebrick") +
  labs(x = "Theoretical", y = "Sample")
```

Proportion test: independent binary trials, same success probability. For `prop.test` the large-sample normal approximation needs $np_0 \geq 5$ and $n(1 - p_0) \geq 5$; both are satisfied.

4.52.8 4. Conduct

```
tt ← t.test(glucose, mu = 95)
tt
```

```
pt_large ← prop.test(x = accept, n = total, p = 0.60)
pt_exact ← binom.test(x = accept, n = total, p = 0.60)
list(prop_test = pt_large, binom_test = pt_exact)
```

4.52.9 5. Concluding statement

Scenario A. In a sample of `n` patients, mean fasting glucose was `round(mean(glucose), 1)` mg/dL (SD `round(sd(glucose), 1)`). The one-sample *t*-test against $\mu_0 = 95$ gave *t*(`tt$parameter`) = `round(tt$statistic, 2)`, *p* = `signif(tt$p.value, 2)`, mean difference `round(mean(glucose) - 95, 1)` (95% CI `round(tt$conf.int[1] - 95, 1)` to `round(tt$conf.int[2] - 95, 1)`).

Scenario B. Acceptance was `accept` of `total` (`round(100 * accept / total, 1)` %); the exact one-proportion test against $\pi_0 = 0.60$ gave *p* = `signif(pt_exact$p.value, 2)`, 95% CI `round(pt_exact$conf.int[1], 2)` to `round(pt_exact$conf.int[2], 2)`.

The two scenarios share a template. The glossary changes — mean and SD for continuous, count and proportion for binary — but the five steps are the same.

4.52.10 Common pitfalls

- Running a one-sided test to "rescue" a borderline two-sided *p*-value.
- Reporting the *p*-value without the mean difference and its CI.
- Using `prop.test` when the expected count of either outcome is very small; use `binom.test` there.

4.52.11 Further reading

- Rosner B. *Fundamentals of Biostatistics*, chapter on one-sample inference.
- Altman DG, *Confidence intervals rather than p values*, BMJ 1986.

4.52.12 Session info

```
sessionInfo()
```

4.52.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.53 Week 4, Session 3 — Pearson and Spearman correlation

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.53.1 Learning objectives

- Compute Pearson and Spearman correlations with `cor.test()` and interpret the coefficient and its CI.

- Distinguish linear association from monotonic association.
- Use visualisation to diagnose situations where a correlation coefficient misleads (Anscombe, DatasauRus).

4.53.2 Prerequisites

Basic inferential statistics.

4.53.3 Background

Pearson's *r* measures linear association. It is the standardised covariance of two variables; it is sensitive to outliers and assumes a roughly linear, bivariate-normal relationship. Spearman's *ρ* measures monotonic association. It is Pearson's *r* applied to the ranks of the variables; it is robust to outliers and to non-linear monotonic transformations.

Neither coefficient captures non-linear, non-monotonic relationships well. Anscombe's quartet and the DatasauRus dozen are the canonical demonstrations that a coefficient without a plot can lie — datasets engineered to have identical Pearson *r* but radically different shapes. The conclusion is not to abandon correlations but never to report one without first plotting the data.

A correlation coefficient is not a causal quantity. Two variables can be perfectly correlated without one causing the other — a lurking variable, a common cause, a coincidence of sample selection. The word "correlation" is often used loosely in the press to mean "weak causation"; in this course it means exactly what `cor.test()` returns.

4.53.4 Setup

```
library(tidyverse)
library(datasauRus)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.53.5 1. Hypothesis

Two scenarios:

1. A linear relationship between two simulated variables.
2. A monotonic but non-linear relationship where Pearson misleads.

H0 in each case: ρ = 0. H1: ρ ≠ 0.

4.53.6 2. Visualise

```
n ← 200
x ← rnorm(n)
y_linear   ← 0.6 * x + rnorm(n, 0, 0.8)
y_monotone ← exp(x) + rnorm(n, 0, 0.5)
df ← tibble(x, y_linear, y_monotone) ▷
  pivot_longer(starts_with("y_"), names_to = "type", values_to = "y")
```

```
df |>
  ggplot(aes(x, y)) +
  geom_point(alpha = 0.5) +
  facet_wrap(~ type, scales = "free") +
  geom_smooth(method = "lm", se = FALSE, colour = "firebrick", linewidth = 0.5)
```

4.53.7 3. Assumptions

Pearson assumes linearity and roughly bivariate-normal data. Spearman assumes monotonicity. Neither handles non-monotonic relationships. For both, the observations must be independent; both are sensitive to extreme sample-size-induced significance when n is huge.

4.53.8 4. Conduct

```
p1_pearson  <- cor.test(x, y_linear,   method = "pearson")
p1_spearman <- cor.test(x, y_linear,   method = "spearman", exact = FALSE)
p2_pearson  <- cor.test(x, y_monotone, method = "pearson")
p2_spearman <- cor.test(x, y_monotone, method = "spearman", exact = FALSE)

tibble(
  relationship = c("linear", "linear", "monotone", "monotone"),
  method = c("Pearson", "Spearman", "Pearson", "Spearman"),
  r_or_rho = c(p1_pearson$estimate, p1_spearman$estimate,
               p2_pearson$estimate, p2_spearman$estimate),
  p = c(p1_pearson$p.value, p1_spearman$p.value,
        p2_pearson$p.value, p2_spearman$p.value)
)
```

The monotonic non-linear relationship gives a much higher Spearman than Pearson — the rank correlation sees the perfect monotonicity.

4.53.8.1 The DatasauRus demonstration

```
ds_summary <- datasaurus_dozen |>
  group_by(dataset) |>
  summarise(mean_x = mean(x),
            mean_y = mean(y),
            sd_x = sd(x),
            sd_y = sd(y),
            r = cor(x, y), .groups = "drop")
head(ds_summary)
```

```
datasaurus_dozen |>
  ggplot(aes(x, y)) +
  geom_point(alpha = 0.4, size = 0.8) +
  facet_wrap(~ dataset) +
  labs(x = NULL, y = NULL)
```

Thirteen datasets; all have mean $x \approx 54$, mean $y \approx 47$, SDs ≈ 17 and 26, and Pearson $r \approx -0.06$. The shapes are wildly different.

4.53.9 5. Concluding statement

> In the linear case, Pearson r = `round(p1_pearson$estimate, 2)` (p = `signif(p1_pearson$p.value, 2)`) and Spearman ρ = `round(p1_spearman$estimate, 2)`; the two agreed. In the monotonic non-linear case, Pearson r = `round(p2_pearson$estimate, 2)` was substantially smaller than Spearman ρ = `round(p2_spearman$estimate, 2)`, revealing that the association was strong but not linear. The DatasauRus dozen reproduced identical summary statistics across 13 radically different shapes, reinforcing the rule: never report a correlation without the scatter plot.

Correlation is a single number. Its honest use requires a picture alongside.

4.53.10 Common pitfalls

- Reporting a Pearson r on data with an obvious curvilinear shape.
- Using r to compare groups with different sample sizes; small r can be "significant" at huge n.
- Inferring causation from a correlation.
- Computing correlations on data that include outliers without a robustness check.

4.53.11 Further reading

- Anscombe FJ (1973). *Graphs in Statistical Analysis*.
- Matejka J, Fitzmaurice G (2017). *Same stats, different graphs: generating datasets with varied appearance and identical statistics*.

4.53.12 Session info

```
sessionInfo()
```

4.53.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.54 Week 3, Session 1 — Populations, samples, and the central limit theorem

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.54.1 Learning objectives

- Distinguish population parameters from sample estimates and their sampling distributions.

- Decompose expected error into bias squared, variance, and irreducible noise.
- Demonstrate the central limit theorem by simulation on a strongly skewed population.

4.54.2 Prerequisites

Weeks 1 and 2.

4.54.3 Background

A population is the set of units we wish to learn about. A sample is the set of units we actually measure. The population has a fixed parameter — mean μ, proportion π, whatever — and the sample has an estimate — $\bar{x}$, $\hat{p}$ — which is a random variable because the sample was drawn at random. Every estimate has a *sampling distribution*: the distribution of its value across hypothetical repetitions of the sampling process. Standard errors, confidence intervals, and *p*-values are all properties of sampling distributions.

Error decomposes into bias and variance. The bias of an estimator is the systematic offset of its sampling distribution's mean from the parameter. The variance is the spread of the sampling distribution. Mean squared error — bias squared plus variance — is the natural scalar summary of how wrong an estimator is on average. A good estimator is not one with zero bias; it is one with low MSE.

The central limit theorem is the reason the normal distribution is ubiquitous. No matter how skewed the population (within mild conditions on variance), the sampling distribution of the mean approaches normal as *n* grows. The lab makes that convergence visible.

4.54.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.54.5 1. Hypothesis

Claim: the sampling distribution of the sample mean from an exponential (strongly right-skewed) population approaches a normal distribution as *n* increases, with shrinking standard error.

4.54.6 2. Visualise

Draw from an exponential population and show the distribution of individual observations.

```
rate ← 1
pop_mean ← 1 / rate
pop_sd   ← 1 / rate

pop_sample ← tibble(x = rexp(2000, rate = rate))
```

```r
pop_sample ▷
  ggplot(aes(x)) +
  geom_histogram(bins = 40, fill = "grey60", colour = "white") +
  geom_vline(xintercept = pop_mean, linetype = 2) +
  labs(x = "x", y = "Count")
```

4.54.7 3. Assumptions

CLT assumes iid sampling and finite variance. Exponential qualifies. The speed of convergence is set by how skewed the parent is — the more skewed, the larger the *n* needed for normality.

```r
# Bias-variance sanity check on the sample mean as an estimator of the
# population mean, across sample sizes.
summarise_means ← function(n, B = 2000) {
  xbars ← replicate(B, mean(rexp(n, rate = rate)))
  tibble(n = n,
         bias = mean(xbars) - pop_mean,
         variance = var(xbars),
         mse = mean((xbars - pop_mean)^2),
         se = sd(xbars))
}
bv ← map_dfr(c(5, 10, 30, 100, 500), summarise_means)
bv
```

4.54.8 4. Conduct

Plot the sampling distribution of the mean across growing *n*.

```r
sim_means ← function(n, B = 4000) {
  replicate(B, mean(rexp(n, rate = rate)))
}

means_df ← bind_rows(
  tibble(n = "n = 2",   mean = sim_means(2)),
  tibble(n = "n = 5",   mean = sim_means(5)),
  tibble(n = "n = 30",  mean = sim_means(30)),
  tibble(n = "n = 100", mean = sim_means(100))
) ▷
  mutate(n = factor(n, levels = c("n = 2", "n = 5", "n = 30", "n = 100")))
```

```r
means_df ▷
  ggplot(aes(mean)) +
  geom_histogram(bins = 40, fill = "grey60", colour = "white") +
  facet_wrap(~ n, scales = "free") +
  geom_vline(xintercept = pop_mean, linetype = 2) +
  labs(x = "Sample mean", y = "Count")
```

The histograms widen (more spread with small *n*) and are strongly right-skewed at *n* = 2. By *n* = 30 they are close to symmetric.

```r
# Check SE shrinking like pop_sd / sqrt(n).
bv ▷ mutate(expected_se = pop_sd / sqrt(n),
            ratio = se / expected_se)
```

4.54.9 5. Concluding statement

> Sampling from an exponential population with mean 1 and SD 1, the sample mean was unbiased across every n simulated ($|\text{bias}| < 0.02$ in all cases). Its standard error shrank as $1/\sqrt{n}$ — ratio of observed to expected SE near 1 throughout — and its sampling distribution was visibly skewed at $n = 2$ but approximately normal by $n = 30$. The central limit theorem is fully operational by modest sample sizes even for a strongly skewed parent.

"$n \geq 30$" is a rule of thumb, not a theorem. For very skewed parents, a larger n is needed; for symmetric parents, a smaller n may suffice. Always check by simulation when in doubt.

4.54.10 Common pitfalls

- Believing the CLT applies to the *data*, not to the *mean of the data*. Individual observations stay whatever they were.
- Invoking the CLT to justify normality of a small skewed sample. The CLT is about sampling distributions.
- Confusing standard error (a property of the sampling distribution) with standard deviation (a property of the data).

4.54.11 Further reading

- Efron B, Hastie T. *Computer Age Statistical Inference*, chapter 1.
- Wasserman L. *All of Statistics*, section on limit theorems.

4.54.12 Session info

```
sessionInfo()
```

4.54.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.55 Week 4, Session 2 — Two proportions, chi-square, risk and odds

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.55.1 Learning objectives

- Compare two proportions with `prop.test()` and `fisher.test()` and say when each is preferred.
- Compute risk ratio, odds ratio, and risk difference with confidence intervals.

- Run a chi-square goodness-of-fit test against a named distribution.

4.55.2 Prerequisites

Labs 2.4, 3.4.

4.55.3 Background

Two-proportion tests ask whether the rate of an event differs between two groups. In a 2×2 table, two exposure groups (say, treatment and control) each produce a fraction of events. The chi-square test compares the observed counts to those expected under independence; Fisher's exact test computes the probability of a table as extreme or more extreme directly from the hypergeometric distribution and is preferred when expected cell counts are small.

The effect can be expressed in several ways. The risk difference is the difference of the two proportions. The risk ratio is their ratio — the event rate in the treated group, relative to control. The odds ratio is the ratio of the odds. In a case-control study the risk ratio is not directly estimable; the odds ratio is. In a cohort study both are estimable and the risk ratio is usually more intuitive.

Goodness-of-fit tests compare a vector of observed counts to a vector of expected proportions under a named distribution — whether a die is fair, whether a set of genotype frequencies matches Hardy-Weinberg, whether arrivals over an hour are Poisson. The test statistic is the same chi-square; the interpretation is "is the distribution consistent with a named model?"

4.55.4 Setup

```
library(tidyverse)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.55.5 1. Hypothesis

Two-by-two: `treatment` vs `control`, outcome `event` vs `none`.

- H0: event rate is independent of treatment.
- H1: rates differ.
- α = 0.05.

4.55.6 2. Visualise

Simulate: 200 per arm, event rate 0.25 in control, 0.15 in treatment.

```
n_per ← 200
sim ← tibble(
  arm = rep(c("control", "treatment"), each = n_per),
  event = c(rbinom(n_per, 1, 0.25),
            rbinom(n_per, 1, 0.15))
)
tab ← table(arm = sim$arm, event = sim$event)
tab
```

```
as_tibble(tab) ▷
  mutate(arm = factor(arm, levels = c("control", "treatment")),
         event = recode(event, `0` = "no event", `1` = "event")) ▷
  ggplot(aes(arm, n, fill = event)) +
  geom_col(position = "fill", alpha = 0.7) +
  labs(x = NULL, y = "Proportion", fill = NULL)
```

4.55.7 3. Assumptions

Chi-square requires expected cell counts ≥ 5. Fisher has no such constraint and is exact at any size. Random assignment or representative sampling is assumed. Independence of observations is essential.

```
chisq.test(tab)$expected
```

All expected counts large; chi-square is appropriate.

4.55.8 4. Conduct

```
ct ← chisq.test(tab, correct = FALSE)
ft ← fisher.test(tab)
pt ← prop.test(x = tab[, "1"], n = rowSums(tab), correct = FALSE)
list(chisq = ct, fisher = ft, prop = pt)
```

4.55.8.1 Effect sizes: RD, RR, OR

```
p_ctrl ← tab["control",   "1"] / sum(tab["control", ])
p_trt  ← tab["treatment", "1"] / sum(tab["treatment", ])

rd ← p_trt - p_ctrl

rr ← p_trt / p_ctrl
# Log-based Wald CI for RR
se_log_rr ← sqrt(
  (1 - p_trt)  / (p_trt  * sum(tab["treatment", ])) +
  (1 - p_ctrl) / (p_ctrl * sum(tab["control",   ]))
)
ci_rr ← exp(log(rr) + c(-1, 1) * qnorm(0.975) * se_log_rr)

or ← (tab["treatment", "1"] * tab["control", "0"]) /
     (tab["treatment", "0"] * tab["control", "1"])
se_log_or ← sqrt(sum(1 / tab))
ci_or ← exp(log(or) + c(-1, 1) * qnorm(0.975) * se_log_or)

tibble(effect = c("RD", "RR", "OR"),
       est = c(rd, rr, or),
       ci_low = c(pt$conf.int[1], ci_rr[1], ci_or[1]),
       ci_high = c(pt$conf.int[2], ci_rr[2], ci_or[2]))
```

4.55.8.2 Goodness-of-fit example

Test whether 600 simulated genotype counts match Hardy-Weinberg proportions at allele frequency 0.3.

```
obs ← c(AA = 260, Aa = 290, aa = 50)
pAA ← 0.7^2; paa ← 0.3^2; pAa ← 2 * 0.7 * 0.3
exp_probs ← c(AA = pAA, Aa = pAa, aa = paa)
gof ← chisq.test(obs, p = exp_probs)
gof
```

4.55.9 5. Concluding statement

> Event rates were `round(p_ctrl, 3)` in control and `round(p_trt, 3)` in treatment (RD `round(rd, 3)`; RR `round(rr, 2)`, 95% CI `round(ci_rr[1], 2)` to `round(ci_rr[2], 2)`; OR `round(or, 2)`, 95% CI `round(ci_or[1], 2)` to `round(ci_or[2], 2)`). Chi-square test χ^2 = `round(ct$statistic, 2)`, df = `ct$parameter`, p = `signif(ct$p.value, 2)`. The Hardy-Weinberg goodness-of-fit test on simulated genotypes gave χ^2 = `round(gof$statistic, 2)` (df = `gof$parameter`, p = `signif(gof$p.value, 2)`).

Report RR or OR alongside the *p*-value. In a cohort design prefer RR, which is the quantity clinicians translate directly to practice.

4.55.10 Common pitfalls

- Reporting an OR when a RR is meaningful and more interpretable.

- Using the continuity correction (the default in `prop.test` and `chisq.test`) by habit; it is conservative.
- Running a chi-square when expected counts are small; use Fisher.
- Treating independent observations as if repeated measures weren't.

4.55.11 Further reading

- Altman DG. *Practical Statistics for Medical Research*, chapter 10–11.
- Agresti A. *Categorical Data Analysis*.

4.55.12 Session info

```
sessionInfo()
```

4.55.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

4.56 Week 4, Session 1 — Two-sample and paired t-tests

Inference labs use the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

4.56.1 Learning objectives

- Choose between Student's and Welch's *t*-test, and between paired and independent-samples tests.
- Compute Cohen's *d* and Hedges' *g* as effect sizes.
- Report a two-sample comparison with point estimate, CI, and effect size.

4.56.2 Prerequisites

Lab 3.4.

4.56.3 Background

The two-sample *t*-test compares the means of two independent groups. Student's version assumes equal variances across the groups; Welch's does not. In most applied settings Welch's is the safer default — it behaves nearly as well as Student's when variances are actually equal and much better when they are not. The `t.test()` default in R is Welch.

When the same subjects are measured twice, the samples are paired and the two measurements are correlated. Ignoring the pairing inflates the standard error and lowers power. Pairing is handled by computing the within-subject difference and running a one-sample *t*-test on it.

Effect sizes remove units from the comparison. Cohen's *d* is the standardised mean difference: the mean difference divided by the pooled standard deviation. Hedges' *g* applies a small-sample correction. Both are essential alongside a *p*-value — a large *d* at a small *n* will often produce a non-significant *p*, but the *d* is still a warning that a larger study might find a real effect.

4.56.4 Setup

```
library(tidyverse)
library(effectsize)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

4.56.5 1. Hypothesis

Scenario A (independent): mean systolic blood pressure in treatment vs placebo.

- H0: μ_trt = μ_pla, H1: μ_trt ≠ μ_pla, α = 0.05.

Scenario B (paired): SBP before and after a four-week intervention in the same 30 patients.

- H0: mean within-subject change = 0, H1: ≠ 0.

4.56.6 2. Visualise

```
n ← 40
trt ← rnorm(n, mean = 135, sd = 12)
pla ← rnorm(n, mean = 142, sd = 12)
df_ind ← tibble(
  arm = rep(c("placebo", "treatment"), each = n),
  sbp = c(pla, trt)
)
```

```
df_ind ▷
  ggplot(aes(arm, sbp, fill = arm)) +
  geom_boxplot(alpha = 0.6, colour = "grey30") +
  labs(x = NULL, y = "SBP (mmHg)") +
  theme(legend.position = "none")
```

```
n2 ← 30
pre  ← rnorm(n2, 148, 10)
post ← pre - rnorm(n2, 5, 7)  # true mean drop of 5
df_pair ← tibble(id = seq_len(n2), pre = pre, post = post) ▷
  mutate(delta = post - pre)
```

```
df_pair ▷
  pivot_longer(c(pre, post), names_to = "time", values_to = "sbp") ▷
  mutate(time = factor(time, levels = c("pre", "post"))) ▷
  ggplot(aes(time, sbp, group = id)) +
  geom_line(alpha = 0.4) +
  geom_point(alpha = 0.6) +
  labs(x = NULL, y = "SBP (mmHg)")
```

4.56.7 3. Assumptions

Independent *t*: approximate normality of the sample means; Welch's does not require equal variances. Paired *t*: approximate normality of the within-subject differences.

```
df_pair ▷
  ggplot(aes(sample = delta)) +
  stat_qq() + stat_qq_line(colour = "firebrick") +
  labs(x = "Theoretical", y = "Sample (pair differences)")
```

4.56.8 4. Conduct

4.56.8.1 Two-sample (Welch)

```
tt_welch ← t.test(sbp ~ arm, data = df_ind, var.equal = FALSE)
tt_welch
```

4.56.8.2 Two-sample Student (for comparison)

```
tt_student ← t.test(sbp ~ arm, data = df_ind, var.equal = TRUE)
tt_student
```

4.56.8.3 Paired

```
tt_paired ← t.test(df_pair$post, df_pair$pre, paired = TRUE)
tt_paired
```

4.56.8.4 Effect size

```
d_ind    ← cohens_d(sbp ~ arm, data = df_ind)
g_ind    ← hedges_g(sbp ~ arm, data = df_ind)
d_paired ← cohens_d(df_pair$post, df_pair$pre, paired = TRUE)
list(d_ind = d_ind, g_ind = g_ind, d_paired = d_paired)
```

4.56.9 5. Concluding statement

Independent comparison. Mean SBP was `round(mean(trt), 1)` mmHg (SD `round(sd(trt), 1)`) in the treatment arm and `round(mean(pla), 1)` (SD `round(sd(pla), 1)`) in placebo. A Welch's *t*-test gave *t*(`round(tt_welch$parameter, 1)`) = `round(tt_welch$statistic, 2)`, *p* = `signif(tt_welch$p.value, 2)`, mean difference `round(diff(rev(tt_welch$estimate)), 1)` mmHg (95% CI `round(-tt_welch$conf.int[2], 1)` to `round(-tt_welch$conf.int[1], 1)`). Cohen's *d* = `round(d_ind$Cohens_d, 2)` (Hedges' *g* = `round(g_ind$Hedges_g, 2)`), a small-to-medium effect.

Paired comparison. The within-subject change (post − pre) had mean `round(mean(df_pair$delta), 1)` mmHg (SD `round(sd(df_pair$delta), 1)`); paired *t*-test *t*(`tt_paired$parameter`) = `round(tt_paired$statistic, 2)`, *p* = `signif(tt_paired$p.value, 2)`, 95% CI `round(tt_paired$conf.int[1], 1)` to `round(tt_paired$conf.int[2], 1)`.

The effect size is the number that generalises. A *p*-value tells you whether to reject; *d* and its CI tell you how much.

4.56.10 Common pitfalls

- Running an independent-samples *t* on paired data and losing power.
- Reporting Student's *t* in the era of Welch; there is rarely a good reason.
- Reporting *p* without *d* (or an equivalent effect size).
- Confusing effect size with statistical significance.

4.56.11 Further reading

- Delacre M et al. (2017). *Why psychologists should by default use Welch's t-test*.
- Cohen J. *Statistical Power Analysis for the Behavioral Sciences*.

4.56.12 Session info

```
sessionInfo()
```

4.56.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)