5 Sample Size & Power

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

Sample-size and power calculations for the designs covered in the rest of the book — t-tests, ANOVA, regression, survival, diagnostic accuracy, equivalence and non-inferiority, sequential designs. Each method page includes the exact `pwr` or simulation call.

This chapter contains **30 method pages** and **2 labs**. If you are not sure which method to read, return to Chapter 0 and follow the decision tree to the right node.

5.1 Method pages

Method	Source slug
Effect Size: Cohen's d	`effect-size-cohens-d`
Effect Size: Cohen's h	`effect-size-cohens-h`
Effect Size: Eta-Squared	`effect-size-eta-squared`
Minimum Detectable Effect	`minimum-detectable-effect`
Post-Hoc Power: A Controversy	`post-hoc-power-controversy`
Power Analysis: Introduction	`power-analysis-introduction`
Power for Agreement (Kappa)	`power-agreement-kappa`
Power for Bland-Altman Studies	`power-bland-altman`
Power for Chi-Squared Tests	`power-chi-squared`
Power for Cluster-RCT	`power-cluster-rct`
Power for Correlation Tests	`power-correlation`
Power for Cox Regression	`power-cox-regression`
Power for Crossover Trials	`power-crossover`
Power for Diagnostic Accuracy	`power-diagnostic-accuracy`
Power for Equivalence (TOST)	`power-equivalence-tost`
Power for ICC	`power-icc`
Power for Linear Regression	`power-linear-regression`
Power for Logistic Regression	`power-logistic-regression`
Power for McNemar's Test	`power-mcnemar`
Power for Non-Inferiority Trials	`power-non-inferiority`
Power for One-Proportion Test	`power-one-proportion`
Power for One-Sample t-Test	`power-one-sample-t`
Power for One-Way ANOVA	`power-anova`
Power for Paired t-Test	`power-paired-t`
Power for Repeated-Measures ANOVA	`power-repeated-measures`

Method	Source slug
Power for Stepped-Wedge Trials	`power-stepped-wedge`
Power for the Log-Rank Test	`power-logrank-test`
Power for Two-Proportion Test	`power-two-proportions`
Sample Size for a Two-Sample t-Test	`power-two-sample-t`
Sample Size Sensitivity Analysis	`sensitivity-analysis-sample-size`

5.2 Labs

Lab
Sample size, power, and Quarto reporting
Power: closed-form and simulation

5.3 Effect Size: Cohen's d

5.3.1 Introduction

Cohen's d is the ratio of a mean difference to a standard deviation. It standardises effects for cross-study comparison and power analysis. Although simple in idea, several variants differ in which SD is used as the denominator.

5.3.2 Prerequisites

Means and standard deviations, two-sample t-test.

5.3.3 Theory

Classical definitions:

- **Cohen's** d (two independent groups): $d = (\bar{x}_1 - \bar{x}_2)/s_{\text{pooled}}$, where $s^2_{\text{pooled}} = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)$.
- **Hedges'** g: Cohen's d multiplied by a small-sample bias correction factor.
- **Glass's** Δ: d using only the control group's SD (for unequal variances).
- **Cohen's** d_z (paired): mean paired difference / SD of differences.
- **Cohen's** d_{rm} (repeated measures): uses a specific variance of the mean difference.

Cohen's thresholds: 0.20 / 0.50 / 0.80 small / medium / large. These are starting points; clinical context should override conventions.

5.3.4 Assumptions

- Approximately normal data (for the SD to be meaningful).

- Equal variances for classical Cohen's d; use Glass or Welch-style for unequal.

5.3.5 R Implementation

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

g1 <- rnorm(50, 50, 10)
g2 <- rnorm(50, 55, 10)

cohens_d(g1, g2)
hedges_g(g1, g2)
glass_delta(g1, g2)

# Paired
pre  <- rnorm(30, 50, 10)
post <- pre + rnorm(30, -5, 5)
cohens_d(pre, post, paired = TRUE)

# Unequal variances → Glass' Delta
g1_big <- rnorm(50, 50, 5)
g2_big <- rnorm(50, 55, 15)
cohens_d(g1_big, g2_big)
glass_delta(g1_big, g2_big)
```

5.3.6 Output & Results

For two-sample example: Cohen's $d \approx -0.47$, Hedges' $g \approx -0.47$, Glass's $\Delta \approx -0.50$.

5.3.7 Interpretation

"The intervention reduced anxiety by 5.1 points (SE 2.0, Cohen's $d = 0.51$, 95 % CI 0.11 to 0.90), a medium-sized effect."

5.3.8 Practical Tips

- Use Hedges' g in meta-analysis; it's the small-sample-corrected version that combines across studies fairly.
- Paired designs should use d_z, not two-sample d; the two are not interchangeable.
- Report both the standardised effect and the raw effect in units that matter clinically.
- Cohen's thresholds vary across fields; cite a field-appropriate reference.
- Confidence intervals for d (e.g., from `effectsize`) are asymmetric; do not assume symmetric.

5.3.9 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.

5.3.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.4 Effect Size: Cohen's h

5.4.1 Introduction

Cohen's h is the standard standardised effect size for differences between two proportions. Unlike the raw risk difference (whose interpretation depends on where on the 0-to-1 scale the comparison sits) or the odds ratio (whose multiplicative scale obscures the absolute magnitude), Cohen's h uses an arcsine-square-root variance-stabilising transformation that makes the effect size approximately scale-free and directly suitable as input to Normal-approximation power calculations. It is the natural building block of two-proportion sample-size formulas in the `pwr` package and equivalent tools, and it is the recommended effect-size summary in published power-analysis protocols.

5.4.2 Prerequisites

A working understanding of binomial proportions, the arcsine variance-stabilising transformation, and the role of standardised effect sizes in sample-size and power calculations.

5.4.3 Theory

For two proportions $p_1, p_2 \in [0, 1]$, Cohen's h is

$$h = 2 \arcsin \sqrt{p_1} - 2 \arcsin \sqrt{p_2}.$$

The arcsine-square-root transformation stabilises the variance of a sample proportion at approximately $1/n$, independent of the proportion's value. As a result, h is dimensionless, scale-free, and directly comparable across baseline rates. Cohen's conventional thresholds are 0.20 (small), 0.50 (medium), and 0.80 (large). The mapping between raw proportion difference and h is non-linear: the same $|p_1 - p_2|$ corresponds to very different h values depending on where on the unit interval the comparison sits, with extreme proportions (near 0 or 1) yielding larger h for the same absolute difference than mid-range comparisons.

5.4.4 Assumptions

The two proportions arise from independent Bernoulli outcomes; the standardisation is most useful for moderate sample sizes where the Normal approximation to the binomial holds.

5.4.5 R Implementation

```r
library(pwr)

ES.h(p1 = 0.10, p2 = 0.20)
ES.h(p1 = 0.50, p2 = 0.60)
ES.h(p1 = 0.80, p2 = 0.90)

pwr.2p.test(h = ES.h(0.10, 0.20), sig.level = 0.05, power = 0.80)
pwr.2p.test(h = ES.h(0.40, 0.50), sig.level = 0.05, power = 0.80)
```

5.4.6 Output & Results

The three example pairs all involve a 10 percentage-point raw difference, but produce very different Cohen's h values: $h = -0.28$ for 10-vs-20 %, $h = -0.20$ for 50-vs-60 %, and $h = -0.29$ for 80-vs-90 %. The corresponding required sample sizes per arm at 80 % power are roughly 200, 388, and 186 — illustrating that the same absolute difference is most expensive to detect near the middle of the proportion range and cheapest at the extremes.

5.4.7 Interpretation

A reporting sentence: "Assuming an intervention increases response from 50 % to 60 % (Cohen's $h = 0.20$, classed as small by Cohen's conventions), 388 participants per group are required for 80 % power at $\alpha = 0.05$ two-sided. The same 10-percentage-point increase from 10 % to 20 % corresponds to $h = 0.28$ (still small) but requires only 200 per group, and from 80 % to 90 % similarly requires 186 per group. The non-linear mapping between absolute proportion difference and detection cost is why h rather than raw difference is used for sample-size planning." Always translate h back to clinical proportions.

5.4.8 Practical Tips

- Prefer Cohen's h over the raw risk difference for sample-size and power calculations because h is the natural input to the Normal-approximation formulas and accounts for the non-constant variance of a proportion across the 0-to-1 scale.
- For rare events (near 0) or common events (near 1), even small raw differences correspond to relatively large h values and are detectable with fewer subjects than mid-range proportions; this counterintuitive property is a consequence of the variance-stabilising transformation and explains why screening trials of rare conditions can detect small absolute increases efficiently.
- Cohen's h is symmetric: $h(p_1, p_2) = -h(p_2, p_1)$, and the sign carries the direction of the comparison; always report the sign and the directional interpretation.
- In reporting and interpretation, translate h back to the raw proportions used as inputs; an h of 0.28 means little to a clinical reader without the corresponding "from 10 % to 20 %" framing.
- For 2×2 contingency-table tests of independence (rather than two-proportion comparisons), use Cohen's w rather than h; the two are related but designed for different test contexts.

- When the comparison is between a single proportion and a fixed reference (one-sample), Cohen's h still applies — $\mathrm{ES.h}(p_1, p_0)$ — and feeds into `pwr.p.test()` rather than `pwr.2p.test()`.

5.4.9 R Packages Used

`pwr::ES.h()` for the canonical Cohen's h calculation and `pwr::pwr.p.test()`, `pwr::pwr.2p.test()` for the corresponding power-analysis functions; `effectsize::cohens_h()` as a tidyverse-friendly alternative; `Hmisc::bsamsize()` for direct-proportion sample-size calculations that complement h-based planning; `MESS::power_prop_test()` for fast exact alternatives; `Mediana` for trial-design simulation including proportion-difference power.

5.4.10 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.

5.4.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.5 Effect Size: Eta-Squared

5.5.1 Introduction

Eta-squared (η^2) is the standard effect-size measure for analysis of variance: the proportion of total variance in the outcome attributable to a factor. It complements the omnibus F-test, which reports significance, by quantifying magnitude — how much of the variability in the response is explained by the categorical factor of interest. Several variants of eta-squared are widely used in different ANOVA designs: classical η^2 in one-way designs, partial η_p^2 in factorial between-subjects designs, generalised η_G^2 in mixed designs that combine between- and within-subjects factors, and omega-squared ω^2 as an unbiased alternative preferred in small samples. Reporting the appropriate variant alongside the F-test result is now standard in psychology, neuroscience, and clinical-trial publications.

5.5.2 Prerequisites

A working understanding of one-way and factorial ANOVA, the partitioning of sums of squares into factor and error components, and the distinction between between-subjects and within-subjects factors.

5.5.3 Theory

Classical eta-squared is

$$\eta^2 = \frac{\mathrm{SS}_{\text{factor}}}{\mathrm{SS}_{\text{total}}}.$$

Partial eta-squared, for factorial designs, removes other factors' variance from the denominator:

$$\eta_p^2 = \frac{\mathrm{SS}_{\text{factor}}}{\mathrm{SS}_{\text{factor}} + \mathrm{SS}_{\text{error}}}.$$

Generalised eta-squared (Bakeman, 2005), for mixed designs combining within- and between-subjects factors, uses a denominator that depends on the design and is comparable across between- and within-subjects effects.

Omega-squared ω^2 corrects for the upward bias of η^2 in small samples:

$$\omega^2 = \frac{\mathrm{SS}_{\text{factor}} - \mathrm{df}_{\text{factor}} \cdot \mathrm{MSE}}{\mathrm{SS}_{\text{total}} + \mathrm{MSE}}.$$

Cohen's conventional thresholds for η^2 are 0.01 (small), 0.06 (medium), and 0.14 (large).

5.5.4 Assumptions

The same as the underlying ANOVA: independent observations (within a stratum), Normal residuals, homogeneous variances. The eta-squared family does not impose additional assumptions beyond those of the ANOVA from which the sums of squares are computed.

5.5.5 R Implementation

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

df <- data.frame(
  A = factor(rep(c("A1", "A2"), each = 30)),
  B = factor(rep(c("B1", "B2", "B3"), 20)),
  y = rnorm(60)
)
df$y <- df$y + 0.5 * (df$A == "A2") + 0.3 * (df$B == "B2") + 0.1 * (df$B == "B3")

fit <- aov(y ~ A * B, data = df)

eta_squared(fit)
eta_squared(fit, partial = TRUE)
eta_squared(fit, generalized = TRUE)
omega_squared(fit)
```

5.5.6 Output & Results

`effectsize::eta_squared()` and `omega_squared()` return point estimates and confidence intervals for each ANOVA term. The output reports the appropriate variant for the requested design, with η_p^2 interpretable as the variance explained by each factor relative to itself plus error, holding other factors constant.

5.5.7 Interpretation

A reporting sentence: "The factorial ANOVA showed a small main effect of factor A on the outcome (partial $\eta_p^2 = 0.06$, 95 % CI 0.01 to 0.21, $F_{1,54} = 3.4$, $p = 0.07$); other terms were small and non-significant. Reporting partial η_p^2 rather than classical η^2 is appropriate for this between-subjects factorial design because each effect's variance share is computed relative to its own error stratum. Omega-squared values were 0.04 and 0.00 respectively, slightly more conservative as expected." Always specify which η^2 variant.

5.5.8 Practical Tips

- Report partial η_p^2 for between-subjects factorial designs; classical η^2 in factorial contexts conflates a factor's variance share with all other factors and is rarely the right summary.
- Use generalised η_G^2 (Bakeman, 2005) for mixed designs that combine within-subjects and between-subjects factors; it gives effect sizes that are directly comparable across the two types of factor, where partial η_p^2 is not.
- Prefer omega-squared ω^2 in small samples (typically $n < 30$ per cell); η^2 is upward-biased in small samples and ω^2 corrects this. Report both when feasible.
- Always report a confidence interval on the effect size, not just the point estimate; CIs on η^2 are non-central-F-based and are reported by `effectsize::eta_squared()` directly.
- Cohen's thresholds (0.01 / 0.06 / 0.14) come from behavioural research and may not be appropriate benchmarks in biomedical or physical-science contexts where larger effects are routinely expected; interpret thresholds in light of substantive expectations rather than treating them as universal.
- For the related $f^2 = \eta^2/(1-\eta^2)$ effect size used in regression power analysis, the conversion is direct and useful when designing follow-up studies based on existing ANOVA results.

5.5.9 R Packages Used

`effectsize::eta_squared()`, `effectsize::omega_squared()`, and `effectsize::epsilon_squared()` for the canonical effect-size family with confidence intervals; `lsr::etaSquared()` for an alternative interface; `MOTE::eta.full.SS()` for textbook-companion calculations; `BayesFactor` for Bayesian effect-size estimation; `Superpower` for ANOVA-design power simulation that also produces eta-squared estimates.

5.5.10 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.

5.5.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.6 Minimum Detectable Effect

5.6.1 Introduction

When the sample size for a study is constrained by budget, time, or availability of participants, the natural planning question is no longer "what n achieves target power?" but instead "given my fixed n, what is the smallest effect I can reasonably hope to detect?" The minimum detectable effect (MDE) reverses the usual sample-size calculation, solving for the effect size at which power equals the target — typically 80 % — given the fixed n, α, and analysis. Reporting an MDE is increasingly required for grant applications with constrained sample sizes, for pilot and exploratory studies, and for honest power-analysis reporting where a study cannot afford to enrol enough participants for the original target.

5.6.2 Prerequisites

A working understanding of statistical power, the relationship between α, effect size, sample size, and power, and the standard test-specific effect-size measures (Cohen's d, r, f^2, w).

5.6.3 Theory

Power, effect size, significance level, and sample size are linked by the non-central distribution of the test statistic, and solving for any one of these quantities given the other three is the standard inversion in any power-analysis framework. The MDE is defined as the effect size at which the planned test achieves a specified power (usually 0.80) under the planned α and the available sample size. Plotted as a curve of MDE vs. n, this gives a clear picture of the design's sensitivity across resource scenarios.

5.6.4 Assumptions

The same assumptions as the underlying test apply (Normality for t-tests, bivariate Normality for correlation, large expected counts for chi-squared, etc.); the MDE is a property of the planned analysis under the design, not a property of the observed data

after the fact.

5.6.5 R Implementation

```r
library(pwr)

pwr.t.test(n = 40, sig.level = 0.05, power = 0.80,
           type = "two.sample", d = NULL)

pwr.r.test(n = 100, sig.level = 0.05, power = 0.80)

n_grid <- seq(10, 200, by = 5)
d_grid <- sapply(n_grid, function(n)
  pwr.t.test(n = n, sig.level = 0.05, power = 0.80, type = "two.sample")$d)

plot(n_grid, d_grid, type = "l", lwd = 2, col = "#2A9D8F",
     xlab = "n per group", ylab = "MDE (Cohen's d)",
     main = "Minimum detectable effect at power = 0.80")
```

5.6.6 Output & Results

For a two-sample t-test with $n = 40$ per group, the MDE at 80 % power is $d = 0.64$ (medium-to-large); doubling to $n = 200$ per group reduces the MDE to $d = 0.28$ (small-to-medium). For a correlation test with $n = 100$, the MDE is $r = 0.28$. Plotting the MDE-vs-n curve shows the diminishing-returns relationship that planners should understand when negotiating sample-size constraints.

5.6.7 Interpretation

A reporting sentence: “With the available sample of 40 participants per arm, the minimum detectable standardised effect (Cohen’s d) at 80 % power and two-sided $\alpha = 0.05$ is 0.64 — a medium-to-large effect by Cohen’s conventions. Effects smaller than this magnitude would likely remain undetected. The smallest clinically meaningful effect for the primary outcome is $d = 0.40$, so the planned study is underpowered for the clinical target; we report the MDE explicitly to clarify the inferential limits and propose a larger follow-up trial.” Always discuss MDE relative to the clinically meaningful effect.

5.6.8 Practical Tips

- Report the minimum detectable effect explicitly when sample size is constrained; this prevents inflated claims of null effects (“no significant difference”) when the study was simply too small to detect a clinically meaningful effect, a recurring problem in the literature.
- Couple MDE reporting with explicit discussion of whether the MDE is clinically or scientifically meaningful; an MDE that exceeds the smallest meaningful effect tells readers the study could not have detected the truth even if it were present.
- For exploratory pilot studies, report MDE in place of formal hypothesis testing; pilot studies are not powered for inference and the MDE conveys the design’s sensitivity honestly.

- MDE inflates rapidly for sub-group analyses (smaller n per cell), cluster-adjusted analyses (larger effective standard errors), repeated-measures designs with low correlation, and any setting where the effective sample size is reduced from the nominal n. Account for these factors in the MDE calculation, not just the headline figure.
- Simulation-based MDE estimation is recommended for complex designs (mixed-effects models, longitudinal data with informative dropout, generalised linear models with heavy-tailed outcomes) where the standard non-central-distribution formulas do not apply directly.
- Pre-register the MDE in the protocol or analysis plan; reporting MDE only after the fact when results turn out non-significant is methodologically suspect, while a pre-registered MDE protects against that interpretation.

5.6.9 R Packages Used

`pwr` family of functions with the effect-size argument set to `NULL` to solve for MDE; `pwrss` for extended MDE calculations across many test types; `WebPower` for an alternative interface; `simr::powerSim()` for simulation-based MDE in mixed-effects designs; `Superpower` for ANOVA-based MDE simulations.

5.6.10 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.

5.6.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.7 Post-Hoc Power: A Controversy

5.7.1 Introduction

Post-hoc power – also called "observed power" – is the calculation of power using the observed effect size from the study that just ran. It is common in some fields (especially when a study fails to reject) and near-universally criticised by statisticians.

5.7.2 Prerequisites

Power analysis, hypothesis testing.

5.7.3 Theory

Why it's problematic:

- Observed power is a monotone function of the p-value: smaller p always yields higher observed power. It conveys no information beyond the p-value itself (Hoenig and Heisey 2001).
- For non-significant results, observed power is bounded above by about 50 % (exactly 50 % when $p = \alpha$).
- It cannot tell you whether the original design was underpowered for a clinically important effect; it only restates how surprising the data were under the null.

Legitimate alternatives:

- **Sensitivity analysis**: compute power for a pre-specified clinically important effect, given n.
- **Confidence intervals** on the observed effect: communicate the range of effects consistent with the data.
- **Minimum detectable effect** (MDE): the smallest effect detectable at the study's n and power target, before the study ran.

5.7.4 Assumptions

None; the issue is epistemological.

5.7.5 R Implementation

```r
library(pwr)

# Simulated study that fails to reject
set.seed(2026)
x ← rnorm(30, 0.2, 1)
t_res ← t.test(x)
c(t = t_res$statistic, p = t_res$p.value)

# Post-hoc power (what many people do)
d_obs ← mean(x) / sd(x)
pwr.t.test(n = 30, d = d_obs, sig.level = 0.05, type = "one.sample")$power

# Better: 95% CI for the mean and MDE
c(CI_lower = t_res$conf.int[1], CI_upper = t_res$conf.int[2])
pwr.t.test(n = 30, power = 0.80, type = "one.sample", sig.level = 0.05)$d
```

5.7.6 Output & Results

```
       t         p
    1.11     0.277      # non-significant

Observed power: 0.19     # uninformative

95% CI: (-0.18, 0.58)
MDE at n = 30: d = 0.52  # effect ⩾ this detectable with 80% power
```

The CI and MDE together say: the effect could be anywhere from -0.18 to 0.58 SD; to detect a medium effect ($d = 0.52$) with 80 % power, $n = 30$ is adequate, so a negative finding here is informative.

5.7.7 Interpretation

Do not report observed power. Report the confidence interval on the observed effect (communicates uncertainty directly) and, if relevant, the MDE at the design stage (communicates what the study was equipped to detect).

5.7.8 Practical Tips

- Reviewers who request post-hoc power should be politely redirected to CI-based reporting.
- "We had low power because the effect turned out to be small" is circular; "our CI cannot rule out a clinically meaningful effect" is informative.
- Pre-registered power calculations are immune to this critique; they are design-stage, not data-driven.
- For replication studies, the effect of interest is from the original study, not the replication sample.
- When a study is genuinely underpowered, the honest conclusion is "inconclusive", not "no effect".

5.7.9 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.

5.7.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.8 Power for Agreement (Kappa)

5.8.1 Introduction

Cohen's kappa measures agreement beyond chance between two raters on a categorical scale and is the de facto standard for inter-rater reliability with categorical outcomes. Sample-size planning for a kappa-based reliability study addresses one of two distinct questions: how many subjects are needed to estimate kappa with a desired precision (a target 95 % confidence-interval half-width around the expected value), or how many subjects are needed to test a hypothesis about kappa (typically that kappa exceeds a

benchmark value such as 0 or 0.6) with adequate power. Modern reliability-study protocols are increasingly precision-driven rather than hypothesis-driven, because a "kappa is greater than zero" claim conveys little practical information about a clinical instrument.

5.8.2 Prerequisites

A working understanding of Cohen's kappa, the marginal-prevalence dependence of kappa's variance, and the distinction between precision-based and hypothesis-test sample-size questions.

5.8.3 Theory

For an expected kappa κ_1 with marginal prevalences π_1 and π_2 for the two raters, the asymptotic variance of the sample kappa is given by the Fleiss-Cicchetti-Everitt formula. Precision-based planning targets a 95 % CI half-width w:

$$n \approx \frac{(1-\kappa_1^2)\, z_{0.975}^2}{w^2 (1-p_e)^2},$$

with p_e the chance-expected agreement. Hypothesis-test planning targets the power to reject $H_0 : \kappa = \kappa_0$ in favour of $H_1 : \kappa = \kappa_1$, using a Normal-approximation z-test on the difference between estimated and hypothesised values, scaled by the appropriate standard errors.

5.8.4 Assumptions

Subjects are independent and each is rated once by each of the two raters; the expected kappa and the marginal prevalences are pre-specified from pilot data or substantive expectations; the rating scale is categorical (or ordinal with weighted kappa).

5.8.5 R Implementation

```r
library(irr)

N.cohen.kappa(rate1 = 0.5, rate2 = 0.5,
              k1 = 0.70, k0 = 0,
              alpha = 0.05, power = 0.80,
              twosided = TRUE)

kappa_exp ← 0.70; p_e ← 0.5
w ← 0.05
n_prec ← qnorm(0.975)^2 * (1 - kappa_exp^2) / (w^2 * (1 - p_e)^2)
ceiling(n_prec)
```

5.8.6 Output & Results

`N.cohen.kappa()` returns the sample size for the hypothesis test of κ_1 vs κ_0 given the marginal rates. For an expected $\kappa_1 = 0.70$ against $\kappa_0 = 0$ at balanced marginals, roughly 40 subjects suffice for the hypothesis-test approach; precision-targeted with a 95 % CI

half-width of 0.05 requires roughly 83 subjects, illustrating that precision-based planning typically demands more subjects than hypothesis-test planning.

5.8.7 Interpretation

A reporting sentence: “To estimate inter-rater Cohen’s kappa with a 95 % CI half-width of 0.05 around an anticipated value of 0.70 (assumed marginal prevalences 0.5 in each rater), 83 subjects are required. The protocol enrols 100 to allow for 15 % unevaluable ratings. A hypothesis-test framing — power to reject $\kappa = 0$ at 80 % power — would require only 40 subjects, but the precision-based plan is preferred because the substantive question is the magnitude of agreement, not whether it differs from zero.” Always state the planning paradigm.

5.8.8 Practical Tips

- Kappa’s variance depends substantially on the marginal prevalences; very imbalanced marginals (e.g., one category at 5 % prevalence) inflate the variance and require more subjects than balanced marginals at the same expected kappa value.
- Weighted kappa for ordinal categories has smaller standard error than unweighted kappa for nominal categories with the same number of categories; adjust the sample-size calculation accordingly using the weighted-kappa variance formula.
- Multi-rater kappa (Fleiss’s kappa) requires a different sample-size calculation than two-rater Cohen’s kappa; use `samplesize::kappa.sample.size()` or simulation-based approaches for the multi-rater design.
- Always report the marginal prevalences and the assumed agreement matrix in the protocol, not just the expected kappa estimate; reviewers cannot reproduce the calculation without them, and the marginals materially affect the variance.
- For continuous ratings, the appropriate reliability statistic is the intraclass correlation coefficient (ICC), and a separate sample-size calculation applies; do not dichotomise continuous ratings to compute kappa.
- For the kappa-paradox situation where high observed agreement coincides with low kappa due to imbalanced marginals, consider reporting both kappa and the prevalence- and bias-adjusted kappa (PABAK) for a fuller picture of reliability.

5.8.9 R Packages Used

`irr::N.cohen.kappa()` for the canonical Cohen’s-kappa hypothesis-test sample-size calculation; `samplesize::kappa.sample.size()` for an alternative interface and multi-rater extensions; `kappaSize` for kappa sample-size with multi-category outcomes; `simr` for simulation-based extensions; `psych` and `irr` for kappa analysis after data collection.

5.8.10 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.

5.8.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.9 Power Analysis: Introduction

5.9.1 Introduction

Power analysis turns four quantities – effect size, significance level, power, and sample size – into a single equation: given any three, the fourth is determined. Every responsible study design pins down three of the four in advance to solve for the fourth (usually sample size). Skipping power analysis leaves studies under- or over-powered, with reputational and ethical consequences.

5.9.2 Prerequisites

Hypothesis testing, Type I and II errors.

5.9.3 Theory

The four quantities:

- **Effect size** (d, f, r, OR, RR): the magnitude of the true effect under the alternative. Must be pre-specified based on prior evidence, clinical relevance, or Cohen's conventions.
- **Significance level** α: the Type I error rate, almost always 0.05 two-sided.
- **Power** $1 - \beta$: probability of detecting the true effect. Conventionally 0.80; confirmatory trials use 0.90.
- **Sample size** n: number of observations (or per group).

Given any three, the fourth is computed from the test's non-central distribution under H_1.

Sensitivity and trade-offs:

- Doubling n roughly shrinks the detectable effect by $\sqrt{2}$.
- Reducing α from 0.05 to 0.01 reduces power at the same n.
- Larger SD reduces power; more precise measurement increases it.
- Power for a one-sided test at α equals power for a two-sided test at 2α (when the effect is in the right direction).

5.9.4 Assumptions

The test's sampling distribution under H_1 must be known (or approximable). Power calculations are as valid as the assumed effect-size magnitude; a plausible range is often reported.

5.9.5 R Implementation

```r
library(pwr)

# Solve for n given d, alpha, power
pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.80,
           type = "two.sample", alternative = "two.sided")

# Solve for d given n and power (minimum detectable effect)
pwr.t.test(n = 30, sig.level = 0.05, power = 0.80,
           type = "two.sample")

# Power curves across d
d_grid <- seq(0.1, 1, by = 0.05)
pw_n30  <- sapply(d_grid, function(d) pwr.t.test(n = 30, d = d)$power)
pw_n100 <- sapply(d_grid, function(d) pwr.t.test(n = 100, d = d)$power)

plot(d_grid, pw_n30, type = "l", col = "#F4A261", lwd = 2,
     xlab = "Cohen's d", ylab = "Power",
     main = "Power curves: n = 30 vs n = 100 per group")
lines(d_grid, pw_n100, col = "#2A9D8F", lwd = 2)
abline(h = 0.80, lty = 2)
legend("bottomright", c("n = 30", "n = 100"),
       col = c("#F4A261", "#2A9D8F"), lwd = 2)
```

5.9.6 Output & Results

```
     Two-sample t test power calculation

              n = 63.77
              d = 0.5
      sig.level = 0.05
          power = 0.80
    alternative = two.sided

     Two-sample t test power calculation

              n = 30
              d = 0.738
      sig.level = 0.05
          power = 0.80
```

64 per arm for a medium effect at 80 % power. With only 30 per arm, the minimum detectable effect is $d = 0.74$.

5.9.7 Interpretation

For a grant or protocol: "Assuming a between-group standardised difference of $d = 0.5$ (a medium effect), $\alpha = 0.05$ two-sided, and power = 0.80, the required sample size is 64 per arm (128 total). We plan to enrol 70 per arm to allow for 10 % loss to follow-up."

5.9.8 Practical Tips

- Choose the effect size from published data or pilot studies, not from Cohen's generic thresholds.
- Always inflate for dropout; report the final n as target.
- Sensitivity analysis: provide n for several plausible d values.
- Do not perform post-hoc power analysis with the observed effect; it is uninformative and sometimes misleading.
- For complex designs (cluster, multilevel), use simulation-based power (`simstudy`, custom Monte Carlo) rather than formulas.

5.9.9 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.

5.9.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.10 Power for One-Way ANOVA

5.10.1 Introduction

Power analysis for one-way ANOVA computes the per-group sample size needed to detect a specified pattern of differences among three or more group means with adequate probability. Where the two-sample t-test handles the simple two-group comparison, the one-way ANOVA generalises the comparison to any number of groups, and its power calculation must therefore handle a multi-group effect-size measure rather than a single mean difference. Cohen's f is the standard standardised effect size; it captures the spread of true group means relative to the within-group standard deviation in a single number that maps cleanly to the non-central F distribution under the alternative hypothesis.

5.10.2 Prerequisites

A working understanding of one-way ANOVA, the omnibus F-test, the non-central F distribution, and Cohen's f as the standardised effect size for ANOVA.

5.10.3 Theory

Cohen's f for one-way ANOVA is

$$f = \frac{\sigma_{\text{between}}}{\sigma_{\text{within}}},$$

the SD of the true group means divided by the within-group SD. The relationship to η^2 is $f = \sqrt{\eta^2/(1-\eta^2)}$. Cohen's conventional benchmarks are 0.10 (small), 0.25 (medium), and 0.40 (large). Under the alternative hypothesis, the omnibus F-statistic follows a non-central F with $(k-1, N-k)$ degrees of freedom and non-centrality $\lambda = Nf^2$, where N is the total sample size across k groups; power is the tail probability of this non-central F above the critical value.

5.10.4 Assumptions

The design is balanced (equal n per group; modest imbalance is tolerable but reduces efficiency), within-group variances are approximately equal (homogeneity of variance), and residuals are approximately Normal. Welch-corrected variants accommodate variance heterogeneity at the cost of slightly more complex power calculations.

5.10.5 R Implementation

```r
library(pwr)

pwr.anova.test(k = 4, f = 0.25, sig.level = 0.05, power = 0.80)

pwr.anova.test(k = 3, f = 0.10, power = 0.80)
pwr.anova.test(k = 3, f = 0.25, power = 0.80)
pwr.anova.test(k = 3, f = 0.40, power = 0.80)

group_means <- c(50, 55, 58)
pooled_sd   <- 8
grand_mean  <- mean(group_means)
f_est <- sqrt(mean((group_means - grand_mean)^2)) / pooled_sd
f_est
```

5.10.6 Output & Results

`pwr.anova.test()` returns the per-group sample size required to achieve target power at specified f, α, and number of groups. For $k = 4$ groups at medium effect $f = 0.25$, 45 per group (180 total) achieves 80 % power; computing f directly from substantive expected group means and a pooled within-group SD ties the calculation to scientifically defensible assumptions.

5.10.7 Interpretation

A reporting sentence: "With four groups and an anticipated Cohen's $f = 0.25$ (medium effect by Cohen's convention; equivalent to expected group means differing by approximately 0.5 within-group SDs), $n = 45$ participants per group (180 total) are required for 80 % power at $\alpha = 0.05$ using a one-way ANOVA. The protocol allocates 50

per group to allow for 10 % attrition, and pre-specified pairwise Tukey HSD comparisons follow if the omnibus is significant." Always describe f in terms of the underlying group-mean structure.

5.10.8 Practical Tips

- Compute Cohen's f from substantive expectations about group means and within-group SD wherever possible, rather than relying on Cohen's medium/large benchmarks; tying the calculation to the actual expected pattern is far more defensible than invoking generic effect-size labels.
- For unequal group variances, power is approximate under the standard formulas; use Welch-adjusted simulation (or `WebPower::wp.kanova()` with explicit variance specification) when heterogeneity is substantial.
- Unbalanced designs have lower power than balanced designs at the same total N; the loss is small for modest imbalances but grows quickly for ratios more extreme than 2:1. Plan for balance whenever feasible.
- For many groups with a single pre-specified contrast of primary interest (e.g., linear trend across ordered groups, treatment-vs-control contrast), pre-specified contrast tests have higher power than the omnibus F-test and should be used in preference.
- Simulation-based power (`simr`, `Superpower`, `faux`) is more flexible for complex designs — unbalanced cells, heterogeneous variances, non-Normal outcomes, mixed-effects structures — and is increasingly the default approach in modern protocol development.
- For factorial ANOVA (two or more between-subjects factors), the calculation extends naturally with effect sizes for each main effect and interaction; `Superpower::ANOVA_power()` is the standard tool.

5.10.9 R Packages Used

`pwr::pwr.anova.test()` for canonical balanced one-way ANOVA power; `WebPower::wp.kanova()` for an alternative interface and unequal-variance extensions; `Superpower::ANOVA_power()` and `Superpower::ANOVA_exact()` for factorial-ANOVA power simulation; `pwrss::pwrss.f.ancova()` for ANCOVA-power calculations; `simr::powerSim()` for general simulation-based ANOVA power.

5.10.10 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.

5.10.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting

- Power: closed-form and simulation

5.11 Power for Bland-Altman Studies

5.11.1 Introduction

Bland-Altman analysis is the standard method for comparing two quantitative measurement methods — a new instrument against a reference standard, two raters' continuous scores, two clinical-laboratory assays. The central output is the 95 % limits of agreement (LoA), defined as the mean difference ± 1.96 SD of paired differences, which describe the range within which 95 % of differences between the two methods are expected to fall. Sample-size planning for a Bland-Altman study therefore targets the precision (CI half-width) of the LoA estimates rather than power to reject a null. With too few subjects, the LoA themselves are estimated imprecisely, and clinical-acceptance decisions about whether the methods are interchangeable become unreliable.

5.11.2 Prerequisites

A working understanding of the Bland-Altman analysis framework, the calculation of limits of agreement, and the standard-error theory for sample quantiles and Normal-based confidence intervals.

5.11.3 Theory

The standard error of each limit of agreement is approximately

$$\mathrm{SE}(\mathrm{LoA}) \approx \sqrt{3}\,\sigma_d/\sqrt{n},$$

where σ_d is the SD of the paired differences between methods. The 95 % CI half-width on each LoA is therefore approximately $1.96\sqrt{3}\sigma_d/\sqrt{n}$. To achieve a target half-width w:

$$n \approx \left(\frac{1.96\sqrt{3}\sigma_d}{w}\right)^2 \approx \frac{11.5\,\sigma_d^2}{w^2}.$$

The factor $\sqrt{3}$ arises because the LoA is a linear combination of the sample mean and sample SD; the formula accounts for both components' uncertainty under Normal-distribution theory.

5.11.4 Assumptions

The paired differences are approximately Normally distributed, no systematic bias varies across the measurement range (proportional bias), subjects are independent, and the SD of differences is reasonably known from pilot data. The Bland-Altman 1999 extension handles repeated measurements per subject with a different formula.

5.11.5 R Implementation

```r
sigma_d <- 5; w <- 1
n_req <- 11.5 * sigma_d^2 / w^2
ceiling(n_req)

set.seed(2026)
n <- 288
diff <- rnorm(n, mean = 0, sd = sigma_d)
mean_diff <- mean(diff)
sd_diff   <- sd(diff)
LoA_lower <- mean_diff - 1.96 * sd_diff
LoA_upper <- mean_diff + 1.96 * sd_diff
SE_LoA    <- sd_diff * sqrt(3 / n)
half_ci   <- 1.96 * SE_LoA
c(LoA_lower, LoA_upper, half_ci)
```

5.11.6 Output & Results

The closed-form calculation gives $n \approx 288$ for $\sigma_d = 5$ and target half-width 1 unit on each LoA. The simulation block confirms the empirical CI half-width matches the formula prediction; reporting both the analytical calculation and a Monte Carlo verification is good practice when the underlying distributional assumptions are uncertain.

5.11.7 Interpretation

A reporting sentence: “To estimate the 95 % limits of agreement between the two methods with a 95 % CI half-width of ± 1 unit on each limit (the pre-specified clinical-acceptance criterion), $n = 288$ paired measurements are required, assuming a within-subject difference SD of 5 units from pilot data. The protocol enrols 320 to allow for 10 % unevaluable measurements. Sensitivity analyses across $\sigma_d \in [4, 6]$ yield required n from 184 to 414.” Always report the target half-width and pilot-derived σ_d.

5.11.8 Practical Tips

- Precision of the LoA is dominated by the SD of differences σ_d; pilot data to estimate σ_d are essential, and protocols that rely on guessed σ_d values should report sensitivity analyses across a plausible range.
- For repeated measurements per subject (multiple paired observations within each individual), the analysis and the sample-size formula both change; use the Bland-Altman 1999 extension and its corresponding precision formula, which accounts for the within-subject correlation.
- Non-constant bias across the measurement range (proportional bias) invalidates simple LoA reporting; regress differences on means to detect proportional bias, and if present, compute LoA conditional on the measurement value.
- For one-sided clinical-acceptance criteria (e.g., upper LoA must be below a tolerance threshold), adjust the calculation to target a one-sided 95 % CI upper bound; the sample-size factor changes correspondingly and is well-documented in the precision-based literature.

- Report the target LoA half-width explicitly in the methods section; vague phrases like “adequate precision” are no longer acceptable in regulatory submissions or method-comparison publications, where reviewers expect explicit precision targets.
- For ratio-scale outcomes with proportional measurement error, log-transform before computing LoA; the LoA on the log scale back-transforms to a multiplicative LoA that is more interpretable.

5.11.9 R Packages Used

`MethComp` for canonical Bland-Altman analysis with confidence intervals on LoA; `BlandAltmanLeh` for an alternative interface with ggplot-style plotting; `pwr` and `MBESS` for general precision-based sample-size tools that translate to LoA contexts; `rmcorr` for repeated-measures correlation alternatives in method-comparison data; `agRee` for ICC-based agreement statistics complementing LoA reporting.

5.11.10 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.

5.11.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.12 Power for Chi-Squared Tests

5.12.1 Introduction

Power analysis for chi-squared tests — goodness-of-fit against a hypothesised distribution and independence in a two-way contingency table — uses the non-central chi-squared distribution under the alternative hypothesis. The standardised effect size for these tests is Cohen’s w, which combines the difference between hypothesised and expected cell proportions into a single number that maps directly to the non-centrality parameter of the test statistic. Power calculations are essential in the design of survey research, epidemiological category-association studies, market-segmentation analyses, and any setting where the central question is whether categorical proportions differ from a reference or vary across rows of a contingency table.

5.12.2 Prerequisites

A working understanding of chi-squared goodness-of-fit and contingency-table tests, the chi-squared distribution and its non-central generalisation, and Cohen's w as the standardised effect-size measure.

5.12.3 Theory

Cohen's w is

$$w = \sqrt{\sum_i \frac{(p_{1i} - p_{0i})^2}{p_{0i}}},$$

where p_{0i} are the cell proportions expected under the null hypothesis and p_{1i} under the alternative. Equivalently, $w = \sqrt{\chi^2/n}$ at the expected effect. Under the alternative, the test statistic follows a non-central chi-squared with the test's standard degrees of freedom and non-centrality parameter $\lambda = nw^2$. Conventional benchmarks for w are 0.10 (small), 0.30 (medium), and 0.50 (large).

5.12.4 Assumptions

Observations are independent, expected cell counts are large enough that the chi-squared approximation is valid (typically ≥ 5 per cell, though stricter thresholds apply for small tables), and the cell proportions under the alternative hypothesis are pre-specified.

5.12.5 R Implementation

```r
library(pwr)

pwr.chisq.test(w = 0.3, df = 3, sig.level = 0.05, power = 0.80)

pwr.chisq.test(w = 0.3, df = 4, sig.level = 0.05, power = 0.80)

p0 ← c(0.25, 0.25, 0.25, 0.25)
p1 ← c(0.40, 0.30, 0.20, 0.10)
w  ← sqrt(sum((p1 - p0)^2 / p0))
w
pwr.chisq.test(w = w, df = 3, power = 0.80)
```

5.12.6 Output & Results

`pwr.chisq.test()` returns the required sample size to achieve target power against a specified w, α, and df. For a medium effect $w = 0.3$ with 3 df (4-category goodness-of-fit), $n = 122$; with 4 df (a 3×3 contingency table), $n = 133$. Computing w directly from hypothesised cell proportions is the recommended workflow because it ties the calculation to substantive distributional assumptions.

5.12.7 Interpretation

A reporting sentence: "To detect a deviation from equal proportions across four categories of medium magnitude (Cohen's $w = 0.30$, equivalent to expected proportions $(0.40, 0.30, 0.20, 0.10)$ vs. uniform null) at 80 % power and $\alpha = 0.05$, $n = 122$ observations are required. With 5 % attrition, the protocol enrols 130 participants. Sensitivity analyses across $w \in [0.20, 0.40]$ are reported in the supplement to bracket uncertainty about the true effect size." Always state both w and the underlying proportional hypothesis.

5.12.8 Practical Tips

- Convert real-world quantities — relative risks, odds ratios, differences in proportions — to Cohen's w via the contingency-table formula rather than guessing; using a generic "medium" benchmark when the underlying scientific question implies a different magnitude is a recurring source of mis-powered studies.
- Required sample size grows roughly linearly with degrees of freedom at fixed w; spreading effects across more cells (e.g., 5×5 vs 2×2) reduces the per-cell signal and inflates the sample-size requirement.
- For small expected cell counts that would invalidate the chi-squared approximation, plan to use Fisher's exact test and estimate power by Monte Carlo simulation; the asymptotic chi-squared power calculation is unreliable in the small-sample regime where exact tests are needed.
- If the alternative-hypothesis cell proportions are themselves uncertain, pre-specify a sensitivity range and report power across the range; this is increasingly expected by reviewers and protects against under-powering when the assumed proportions turn out to be optimistic.
- Chi-squared power calculations are approximate (the non-central chi-squared assumes large samples); simulation-based power is safer for complex designs, sparse-data scenarios, or when the test statistic deviates from the standard Pearson form.
- For one-way exact-binomial tests of a single proportion, use specialised tools like `pwr.p.test()` or `binom::power.binom()` rather than the chi-squared approximation; the exact methods are more accurate at small samples.

5.12.9 R Packages Used

`pwr::pwr.chisq.test()` for the canonical Cohen's w chi-squared power calculation; `WebPower::wp.chisq()` for an alternative interface; `MESS::power_prop_test()` for two-proportion specific power; `binom` for exact-binomial sample-size tools; `Mediana` for trial-design simulation including chi-squared-test power across more complex designs.

5.12.10 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.

5.12.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.13 Power for Cluster-RCT

5.13.1 Introduction

In cluster-randomised trials (CRTs), treatment is assigned at the cluster (clinic, school, village) rather than individual level. Within-cluster correlation inflates variance, reducing effective sample size and requiring more individuals than an individually randomised trial.

5.13.2 Prerequisites

Intraclass correlation, CRT design.

5.13.3 Theory

The **design effect** (DE) converts individual-level sample size to cluster-RCT sample size:

$$\mathrm{DE} = 1 + (\bar{m} - 1)\rho_{ICC},$$

where $\bar{m}$ is the average cluster size and ρ_{ICC} is the intraclass correlation.

Required individual-level n for an individually-randomised design, multiplied by DE, gives the cluster-RCT total. Required clusters = total / $\bar{m}$.

5.13.4 Assumptions

- Balanced or nearly balanced cluster sizes.
- Pre-specified ICC (from pilot or literature).
- Analysis adjusts for clustering (mixed models or GEE).

5.13.5 R Implementation

```r
library(clusterPower)

# Continuous outcome: detect a 5-unit mean difference, sigma = 15
# ICC = 0.02, 30 subjects per cluster
# Individual n for same power:
n_ind <- 2 * 15^2 * (qnorm(0.975) + qnorm(0.80))^2 / 5^2
n_ind

rho <- 0.02
m_bar <- 30
DE <- 1 + (m_bar - 1) * rho
c(design_effect = DE,
  cluster_total_n = n_ind * DE,
  clusters_per_arm = ceiling(n_ind * DE / m_bar / 2))

# Direct calculation via clusterPower
cpa.normal(alpha = 0.05, power = 0.80, nclusters = NA,
           nsubjects = 30, d = 5 / 15,   # d = delta / sigma
           ICC = 0.02)
```

5.13.6 Output & Results

Individual $n \approx 143$ per arm; DE = 1.58 at ICC 0.02 and cluster size 30, giving 226 per arm (about 8 clusters of 30 per arm).

5.13.7 Interpretation

“With an assumed ICC of 0.02 and 30 subjects per cluster, the design effect is 1.58. To detect a standardised difference of 0.33 (5 units / 15 SD) at 80 % power and two-sided $\alpha = 0.05$, 8 clusters per arm are required (total 480 subjects).”

5.13.8 Practical Tips

- ICC estimates vary widely; use pilot data, literature, or sensitivity ranges (0.001 to 0.1).
- Unequal cluster sizes reduce power; inflate n by 10-20 % as buffer.
- Large m increases DE linearly in ICC; adding clusters is more efficient than more per-cluster recruitment.
- Stratified or matched CRTs (pairs of clusters) reduce DE.
- Analysis model: linear mixed model with random cluster effects; GEE with exchangeable correlation; cluster-level summary.

5.13.9 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.

5.13.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.14 Power for Correlation Tests

5.14.1 Introduction

Power analysis for correlation tests determines the sample size required to detect a specified population correlation ρ different from zero (or different from a non-zero reference) with adequate probability. Correlation studies are pervasive in clinical research, psychometrics, behavioural science, biomarker validation, and any setting where the strength of association between two continuous variables is the central question. The standard analytical approach uses Fisher's z-transformation, which converts the sampling distribution of the Pearson correlation into an approximately Normal distribution with a simple variance formula, supporting clean closed-form sample-size calculations.

5.14.2 Prerequisites

A working understanding of Pearson and Spearman correlation, the Fisher z-transformation, and the relationship between effect size and required sample size in inferential statistics.

5.14.3 Theory

Fisher's z-transformation is

$$z = \tfrac{1}{2}\log\frac{1+r}{1-r} = \operatorname{atanh}(r),$$

with $\mathrm{SE}(z) \approx 1/\sqrt{n-3}$. The required sample size to detect ρ different from 0 at two-sided α with power $1-\beta$ is

$$n \approx \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\operatorname{atanh}(\rho)}\right)^2 + 3.$$

Cohen's effect-size benchmarks for r are 0.10 (small), 0.30 (medium), and 0.50 (large). Spearman's rank correlation has slightly less efficient inference under bivariate Normality (asymptotic relative efficiency $9/\pi^2 \approx 0.91$), so a 9 % inflation in n converts a Pearson sample-size calculation to its Spearman equivalent.

5.14.4 Assumptions

The data are bivariate Normal for the Pearson test, the pairs are independent and identically distributed, and the population correlation is genuinely well-defined. Spearman correlation tolerates monotone non-linearity but its power calculation is approximate.

5.14.5 R Implementation

```r
library(pwr)

pwr.r.test(r = 0.30, sig.level = 0.05, power = 0.80)

pwr.r.test(r = 0.10, power = 0.80)

r_grid ← seq(0.1, 0.6, by = 0.05)
pw ← sapply(r_grid, function(r) pwr.r.test(r = r, n = 50)$power)
data.frame(r = r_grid, power_n50 = round(pw, 2))
```

5.14.6 Output & Results

`pwr.r.test()` returns the required n at specified r, α, and power. For $r = 0.30$ at 80 % power and two-sided $\alpha = 0.05$, $n = 85$; for $r = 0.10$, $n = 782$ — illustrating how dramatically small correlations inflate the required sample size. The power-by-correlation table at fixed n shows the achievable power across the range of plausible correlations and is a useful supplement to the single-point calculation.

5.14.7 Interpretation

A reporting sentence: "To detect a medium correlation of $r = 0.30$ with 80 % power at two-sided $\alpha = 0.05$, a sample of 85 participants is required. With the planned $n = 50$, achievable power for the same effect is 56 %, leaving the study substantially under-powered. The protocol therefore enrols 90 participants to allow for 5 % attrition. Sensitivity analyses across $\rho \in [0.20, 0.40]$ show required n ranging from 47 to 191." Always report sensitivity over a plausible ρ range.

5.14.8 Practical Tips

- Testing ρ against a non-zero reference value requires a modified Fisher-z formula that accounts for the difference between $\operatorname{atanh}(\rho_0)$ and $\operatorname{atanh}(\rho_1)$; use `pwrss::pwrss.z.corr()` or compute manually rather than relying on the standard `pwr.r.test()` which assumes $\rho_0 = 0$.
- For Spearman correlations, inflate the Pearson-based n by approximately 9 % to reflect the slight efficiency loss of the rank-based test under bivariate Normality; for non-Normal underlying data the rank-based test can actually be more efficient.
- Correlations near ± 1 have lower sampling variance than correlations near zero (a consequence of the bounded support); Fisher's z-transformation handles this automatically by mapping r to a scale where the variance is approximately constant.
- When many correlations are tested simultaneously (e.g., a correlation matrix of 10 variables yields 45 pairs), adjust α for multiplicity using Bonferroni or false-discovery-rate procedures, and recompute power against the adjusted threshold.
- Sensitivity analysis across plausible ρ is routine in correlational study planning; protocols with a single point ρ are increasingly flagged by reviewers, who expect a range and a justification of the lower bound used for sample-size determination.
- For paired or longitudinal correlation comparisons (e.g., comparing ρ at two time points), use Steiger's modified Fisher-z test and its dedicated power calculation; the

standard formulas apply only to a single correlation against a fixed reference.

5.14.9 R Packages Used

`pwr::pwr.r.test()` for the canonical Fisher-z-based power calculation; `pwrss::pwrss.z.corr()` for non-zero reference and other extensions; `WebPower::wp.correlation()` for an alternative interface; `simr` and `Superpower` for simulation-based correlation power across complex designs; `boot` for bootstrap-based power estimation when assumptions are uncertain.

5.14.10 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.

5.14.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.15 Power for Cox Regression

5.15.1 Introduction

In survival analysis, statistical power is driven by the number of **events**, not by the number of enrolled subjects. A Cox regression power calculation determines the required events and, from assumed accrual and follow-up, the number of subjects.

5.15.2 Prerequisites

Cox proportional hazards, survival analysis.

5.15.3 Theory

For a continuous predictor with standardised-coefficient effect, the required number of events for Wald test power at $1-\beta$ and two-sided α is

$$D = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\sigma_X^2 \log^2(HR)}.$$

For a binary predictor with allocation π (fraction in one group):

$$D = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\pi(1-\pi)\log^2(HR)}.$$

Required subjects $N = D/p_{\text{event}}$, where p_{event} is the probability of observing the event during follow-up; depends on survival, accrual, and censoring.

5.15.4 Assumptions

- Proportional hazards.
- Independent censoring.
- Known baseline event rate or survival function.

5.15.5 R Implementation

```r
library(powerSurvEpi)

# Two-arm trial: equal allocation, HR = 0.7, 1-year survival = 0.80 in control
ssizeCT.default(power = 0.80, k = 1,
                pE = 1 - 0.70, pC = 1 - 0.80,
                RR = 0.70, alpha = 0.05)

# Number of events for a continuous predictor
# HR per 1-SD increase of 1.4
D ← (qnorm(0.975) + qnorm(0.80))^2 / log(1.4)^2
D
```

5.15.6 Output & Results

For HR = 0.70 with equal allocation and event rates of 20 % (treatment) and 30 % (control), about 500 events and 1000 subjects are required.

5.15.7 Interpretation

“The study requires approximately 500 events to detect a hazard ratio of 0.70 with 80 % power. Assuming the expected event probabilities during follow-up, we need to enrol 1000 patients.”

5.15.8 Practical Tips

- Events, not subjects, are the currency of power in survival; count accumulated events, not randomised subjects.
- Report both target events and target N; specify the assumed accrual period and follow-up.
- For interim analyses, use group-sequential boundaries (O’Brien-Fleming, Pocock) to inflate the final sample accordingly.
- Non-proportional hazards reduce power; use restricted mean survival time or weighted log-rank alternatives.
- For competing risks, use Fine-Gray or cause-specific power calculations.

5.15.9 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.

5.15.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.16 Power for Crossover Trials

5.16.1 Introduction

In a crossover trial, each subject receives multiple treatments in sequence with a washout period between. The paired comparison cancels between-subject variability, often requiring far fewer subjects than parallel-group trials. The classic 2×2 crossover has two treatments and two periods.

5.16.2 Prerequisites

Paired t-test, power analysis, crossover design basics.

5.16.3 Theory

In a 2×2 crossover, the period-adjusted within-subject difference has variance σ_w^2, typically estimated from prior crossover data. Sample size is analogous to the paired t-test with $\sigma_d = \sigma_w\sqrt{2}$:

$$n \approx 2\sigma_w^2(z_{1-\alpha/2} + z_{1-\beta})^2/\delta^2.$$

Compared to parallel-group with pooled σ: approximately half the subjects for the same δ when $\sigma_w < \sigma_{\text{between}}$.

Carryover effects invalidate the simple formula; a washout period is planned to eliminate them.

5.16.4 Assumptions

- No carryover; washout adequate.
- Period effects modelled if present.
- Within-subject variance known or estimable.

5.16.5 R Implementation

```r
library(PowerTOST); library(pwr)

# 2x2 crossover bioequivalence: CV within = 20%, ratio = 0.95
sampleN.TOST(CV = 0.20, theta0 = 0.95,
             theta1 = 0.80, theta2 = 1.25,
             alpha = 0.05, targetpower = 0.80,
             design = "2x2")

# Therapeutic crossover: mean difference = 5, sigma_w = 10
n_pair ← 2 * (qnorm(0.975) + qnorm(0.80))^2 * 10^2 / 5^2
n_pair
# Equivalent paired-t calculation
pwr.t.test(d = 5 / 10, type = "paired", power = 0.80)
```

5.16.6 Output & Results

Bioequivalence 2×2 at 20 % CV, ratio 0.95: 20 subjects. Therapeutic 5-unit effect with $\sigma_w = 10$: 34 subjects.

5.16.7 Interpretation

“The 2×2 crossover trial requires 20 subjects for 80 % power to establish bioequivalence within the 80-125 % range, assuming a 20 % within-subject CV and a true ratio of 0.95.”

5.16.8 Practical Tips

- Crossover saves subjects but doubles measurement commitment per subject; adherence matters.
- Washout must be long enough to eliminate pharmacokinetic and physiological carryover.
- More than two periods (Latin squares, Williams designs) can be more efficient but require careful balancing.
- Exclude subjects with missing period data (or use mixed models).
- For strictly sequential effects (learning, fatigue), random order assignment and period-as-factor in the model help.

5.16.9 For Reviewers

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

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

5.16.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.17 Power for Diagnostic Accuracy

5.17.1 Introduction

Sample-size planning for diagnostic studies typically targets **precision** (CI width) for sensitivity and specificity rather than power against a hypothesis. Separate calculations apply to diseased and non-diseased cohorts.

5.17.2 Prerequisites

Sensitivity, specificity, proportion CIs.

5.17.3 Theory

For a Wilson 95 % CI on sensitivity of anticipated value p_{sens} with desired half-width w:

$$n_{\text{dis}} \approx \frac{z_{0.975}^2 p_{\text{sens}}(1 - p_{\text{sens}})}{w^2}.$$

Analogously for specificity using non-diseased. Total sample = $n_{\text{dis}} + n_{\text{non-dis}}$, with the ratio determined by the prevalence of disease in the source population.

For hypothesis-testing framing (is sensitivity > some reference), the one-proportion power formula applies.

5.17.4 Assumptions

- Anticipated sensitivity / specificity from pilot or literature.
- Independent gold-standard verification for every subject.

5.17.5 R Implementation

```
# Design 95% CI half-width 5% on sensitivity = 0.90
p_sens ← 0.90; w ← 0.05
n_dis ← qnorm(0.975)^2 * p_sens * (1 - p_sens) / w^2
ceiling(n_dis)

# Spec = 0.85, same width
p_spec ← 0.85
n_non ← qnorm(0.975)^2 * p_spec * (1 - p_spec) / w^2
ceiling(n_non)

# Hypothesis-test framing: sensitivity > 0.80
# Expected sens = 0.90, alpha = 0.05, power = 0.80
library(MKmisc)
power.prop1.test(p0 = 0.80, p1 = 0.90, sig.level = 0.05, power = 0.80)
```

5.17.6 Output & Results

For sens 0.90 +/- 5 %: 138 diseased cases. For spec 0.85 +/- 5 %: 196 non-diseased. Hypothesis-test $n \approx 130$ diseased cases to show sensitivity > 0.80.

5.17.7 Interpretation

“To estimate sensitivity of 0.90 with a 95 % CI half-width of 5 %, at least 138 diseased cases are required. Assuming a disease prevalence of 30 % in the source population, a total enrolment of 460 subjects provides adequate non-diseased controls.”

5.17.8 Practical Tips

- Plan separately for sensitivity and specificity; the two calculations have independent sample-size requirements per cohort.
- Use Wilson or exact intervals for the reported CI; Wald intervals can be optimistic.
- STARD reporting guidelines recommend reporting both prevalence and the numbers of diseased / non-diseased subjects.
- Paired designs (both tests on each subject) use McNemar-style within-subject comparisons; they are more efficient than unpaired designs.
- AUC-based sample sizes exist (e.g., Hanley-McNeil); different from sensitivity-only calculations.

5.17.9 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.

5.17.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.18 Power for Equivalence (TOST)

5.18.1 Introduction

Power analysis for equivalence testing via the two-one-sided-tests (TOST) procedure requires that both one-sided null hypotheses — that the true effect lies below the lower equivalence bound and that it lies above the upper equivalence bound — be rejected at level α. The required sample size depends on the equivalence margin (set by clinical or regulatory convention), the anticipated true effect (typically zero or near-zero for an honest equivalence claim), and the variance. TOST sample-size calculations are widely used in bioequivalence studies of generic drugs (where the regulatory margin is fixed at 80–125 % on the log scale), in non-inferiority trials reframed as equivalence, and in therapeutic-equivalence studies of competing devices or formulations.

5.18.2 Prerequisites

A working understanding of TOST equivalence testing, the role of the equivalence margin, and the difference between equivalence (CI within the margin) and traditional superiority testing (CI excluding the null).

5.18.3 Theory

For two groups with true mean difference δ and pooled SD σ, and symmetric equivalence margin $\pm\Delta$, the power to conclude equivalence is approximately

$$\Pr\left(\frac{|\bar{X}_1 - \bar{X}_2 - \delta|}{\sigma\sqrt{2/n}} < z_{1-\alpha} \cdot \frac{\Delta - |\delta|}{\Delta \cdot \sigma\sqrt{2/n}}\right).$$

Two important features fall out of this expression. First, when the true effect is at the centre of the equivalence margin ($\delta = 0$), the required sample size is minimised. Second, when the true effect equals the boundary ($|\delta| = \Delta$), equivalence cannot be established regardless of n — the test is fundamentally underpowered at the boundary.

5.18.4 Assumptions

The equivalence margin is pre-specified and substantively justified, the anticipated true effect δ is realistically estimated, the data are approximately Normal or the sample is large enough for the central limit theorem, and the variance is known or accurately estimated from pilot data.

5.18.5 R Implementation

```r
library(TOSTER); library(PowerTOST)

sampleN.TOST(alpha = 0.05, targetpower = 0.80,
             theta0 = 1.00, theta1 = 0.80, theta2 = 1.25,
             CV = 0.20, design = "2x2")

power_t_TOST(alpha = 0.05, low_eqbound = -2, high_eqbound = 2,
             mu = 0, sd = 5, type = "two.sample", n = 65)
```

5.18.6 Output & Results

`sampleN.TOST()` returns the required sample size for bioequivalence with the regulatory 80–125 % margin on the log scale; for 20 % within-subject CV and true ratio 1.00, $n \approx 24$ in a 2 × 2 crossover. `power_t_TOST()` computes power for a generic two-sample equivalence test with arbitrary symmetric or asymmetric margin: at ± 2 unit margin, true difference 0, and SD 5, $n = 65$ per group achieves 81 % power.

5.18.7 Interpretation

A reporting sentence: “To establish bioequivalence (regulatory 80–125 % equivalence margin on the log scale) with 80 % power at $\alpha = 0.05$, assuming a 20 % within-subject coefficient of variation and a true geometric mean ratio of 1.00, $n = 24$ subjects are required in a 2 × 2 crossover design. The protocol enrols 28 to allow for 15 % dropout. Sensitivity analyses across CV $\in [15\%, 30\%]$ yield required n from 16 to 38, and across true ratio $\in [0.95, 1.05]$ from 24 to 33.” Always state the margin and the assumed true effect.

5.18.8 Practical Tips

- The equivalence margin is a substantive and regulatory choice, not a statistical one; set it based on clinical or regulatory conventions (e.g., the 80–125 % bioequivalence margin) or on a pre-specified clinical-acceptance criterion, and justify the choice in the protocol.
- Power drops sharply as the anticipated true effect approaches the equivalence margin; always include sensitivity analysis over a plausible range of true effects to demonstrate that the design is robust to mild deviations from the central assumption.
- Crossover bioequivalence trials use within-subject variance (CV_w), which is typically smaller than between-subject variance; plan accordingly with the appropriate within-subject CV from pilot or literature.
- Asymmetric equivalence margins are supported by most TOST tools (`low_eqbound` ≠ `high_eqbound`) and are useful when only one direction of effect carries a clinical penalty; for example, a drug that is allowed to be more effective but not less effective than the reference.
- Non-inferiority testing is mathematically a one-sided TOST against a single non-inferiority margin; the required sample size is roughly half of a two-sided equivalence test for the same margin and assumptions, reflecting the simpler hypothesis.
- For highly variable drugs (within-subject CV above 30 %), reference-scaled bioequivalence widens the margin proportionally to the reference variability; standard fixed-margin TOST sample-size tools under-estimate the required n for these compounds.

5.18.9 R Packages Used

`PowerTOST` for canonical bioequivalence and equivalence sample-size and power across crossover and parallel-group designs; `TOSTER` for general TOST analysis and power on raw and standardised effect sizes; `MBESS` and `pwrss::pwrss.t.2means()` for related precision and equivalence calculations; `bear` for end-to-end FDA-compliant bioequivalence workflow; `replicateBE` for replicate-design reference-scaled bioequivalence.

5.18.10 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.

5.18.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.19 Power for ICC

5.19.1 Introduction

The intraclass correlation coefficient (ICC) is the standard reliability statistic for continuous ratings — agreement between raters on quantitative scales, test-retest stability, instrument repeatability across multiple measurements per subject. Sample-size planning for an ICC study depends on the number of subjects, the number of raters or replicate measurements per subject, the target ICC value to detect (or estimate), and whether the goal is hypothesis testing (rejecting a null ICC value) or precision (achieving a target confidence-interval width). Both questions admit closed-form solutions under standard assumptions, and explicit power planning is now expected for any prospective reliability study aiming for FDA, IVDR, or guideline acceptance.

5.19.2 Prerequisites

A working understanding of ICC variants — one-way random ICC(1,1), two-way random ICC(2,1), two-way mixed ICC(3,1), single-measurement vs average-measurement variants — and the distinction between hypothesis-testing and precision-based sample-size questions.

5.19.3 Theory

Following Walter, Eliasziw, and Donner (1998), the sample size for testing $H_0 : \rho \leq \rho_0$ against $H_1 : \rho \geq \rho_1$ with k raters or measurements per subject and target power $1 - \beta$ has a closed-form expression based on the F-distribution under the variance-component decomposition. For precision-based planning, the approximate formula

$$n \approx \frac{(1 - \rho^2)^2 \cdot 8 \cdot (k - 1)}{\epsilon^2 \cdot k}$$

gives the number of subjects needed for a 95 % CI half-width of ϵ around the expected ICC ρ — useful when the goal is a tight reliability estimate rather than rejecting a specific null.

5.19.4 Assumptions

The outcome is continuous and approximately Normal, subjects are independent, and each subject is rated by the same k raters (or measured at k replicate occasions); the ICC variant assumed in the calculation matches the design (one-way random vs two-way random vs two-way mixed).

5.19.5 R Implementation

```r
library(ICC.Sample.Size)

calculateIccSampleSize(p = 0.8, p0 = 0.6, k = 2,
                       alpha = 0.05, tails = 2, power = 0.80)

k ← 2; rho ← 0.75; eps ← 0.1
n_prec ← (1 - rho^2)^2 * 8 * (k - 1) / (eps^2 * k)
ceiling(n_prec)
```

5.19.6 Output & Results

`calculateIccSampleSize()` returns the required number of subjects for the hypothesis test. For $\text{ICC}_1 = 0.8$ vs $\text{ICC}_0 = 0.6$ at $k = 2$ raters, $n \approx 84$. The precision-based calculation gives roughly 19 subjects for a 95 % CI half-width of 0.10 around an expected ICC of 0.75 — a much smaller sample because precision tolerates a wider range than rejecting a specific null.

5.19.7 Interpretation

A reporting sentence: "Each subject will be rated by two independent raters; $n = 84$ subjects are required for 80 % power to detect an ICC of 0.80 against a null hypothesis of ICC = 0.60 at two-sided $\alpha = 0.05$. The protocol enrols 100 subjects to allow for 15 % missing or unevaluable ratings. Sensitivity analyses show $n = 110$ for the more conservative null of 0.65 and $n = 65$ for $k = 3$ raters; the design report includes both alternatives in the supplement." Always state which ICC form and how many raters.

5.19.8 Practical Tips

- More raters per subject reduces the required number of subjects, but at linear cost in rater-hours; the trade-off depends on the relative cost and availability of subjects vs raters and is project-specific.
- Specify which ICC variant matches the design exactly — one-way random (ICC(1,1)) for fully crossed designs without rater identity, two-way random (ICC(2,1)) when raters are sampled from a population, two-way mixed (ICC(3,1)) when raters are fixed and chosen — because the sample-size formula and the analytic ICC differ across variants.
- Small n produces wide ICC confidence intervals; for reliability claims supporting clinical use, aim for a 95 % CI lower bound above 0.75 (good) or 0.90 (excellent), per Koo and Li (2016) recommendations.

- Sensitivity analysis varying the expected ρ across a realistic range is standard practice; protocols with a single point estimate of ρ are increasingly flagged by reviewers, who expect an honest acknowledgement of uncertainty.
- Published reliability studies historically have small samples and produce poorly characterised ICC estimates; planning conservatively (assuming ρ at the lower end of pilot CIs) protects against under-powered ICC validation.
- For agreement on categorical ratings, use Cohen's or Fleiss's kappa rather than ICC; the sample-size frameworks differ but the planning principles are analogous.

5.19.9 R Packages Used

`ICC.Sample.Size::calculateIccSampleSize()` for canonical hypothesis-test sample-size calculation; `irr::icc()` and `psych::ICC()` for ICC analysis after data collection; `ICC` for advanced ICC inference and CI methods; `simr` for simulation-based ICC power across complex designs; `MBESS::ci.reliability()` for confidence-interval-driven sample-size planning across reliability statistics.

5.19.10 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.

5.19.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.20 Power for Linear Regression

5.20.1 Introduction

Power analysis for linear regression determines the sample size required to detect either an overall non-zero R^2 (via the omnibus F-test of the regression) or the incremental R^2 contributed by adding one or more predictors to a baseline model. Both questions are central to the design of observational and experimental studies that use multiple regression for inference: the omnibus power calculation tells you whether the model as a whole has a chance to be detected, and the incremental power calculation tells you whether a specific block of predictors of scientific interest will be detectable above and beyond control variables. Cohen's f^2 effect size unifies these calculations into a single non-central F-distribution framework.

5.20.2 Prerequisites

A working understanding of multiple linear regression, the relationship between R^2 and the omnibus F-statistic, and Cohen's f^2 as the standardised effect-size measure for regression.

5.20.3 Theory

Cohen's f^2 for the omnibus regression test is

$$f^2 = \frac{R^2}{1 - R^2},$$

and for the incremental contribution of a new block of predictors,

$$f^2 = \frac{R^2_{\text{full}} - R^2_{\text{reduced}}}{1 - R^2_{\text{full}}}.$$

Conventional benchmarks are 0.02 (small), 0.15 (medium), and 0.35 (large). The omnibus or partial F-test follows a non-central F-distribution under H_1 with non-centrality $\lambda = f^2(u + v + 1)$, where u is the numerator df (number of predictors tested) and v is the residual df. Power is the probability that this non-central F exceeds the critical value at α.

5.20.4 Assumptions

The standard linear-regression assumptions (linearity, homoscedasticity, Normal residuals, independent observations), the number of predictors in the full model and the incremental block are pre-specified, and the effect size f^2 is chosen on the basis of the smallest scientifically meaningful incremental R^2.

5.20.5 R Implementation

```r
library(pwr)

pwr.f2.test(u = 5, v = NULL, f2 = 0.15, sig.level = 0.05, power = 0.80)

pwr.f2.test(u = 2, v = NULL, f2 = 0.02, power = 0.80)
```

5.20.6 Output & Results

`pwr.f2.test()` returns the residual degrees of freedom v required to achieve target power; the total sample size is $n = v + u + 1$. For a 5-predictor model at medium effect $f^2 = 0.15$, $v = 85$ and $n = 91$; for a small incremental effect $f^2 = 0.02$ with 2 added predictors after 5 baseline predictors, $v = 489$ and $n = 497$ — illustrating how dramatically small incremental effects increase sample-size requirements.

5.20.7 Interpretation

A reporting sentence: "With 5 predictors in the regression and an anticipated $f^2 = 0.15$ (medium-effect convention; equivalent to $R^2 = 0.13$), a total sample of $n = 91$ is required to achieve 80 % power for the omnibus F-test at $\alpha = 0.05$. To detect an incremental $f^2 = 0.02$ from 2 additional predictors of primary interest after adjusting for the 5 baseline covariates, the protocol requires $n = 497$. The trial therefore enrols 550 participants to allow for 10 % attrition." Always state u, f^2, and the inflation factor.

5.20.8 Practical Tips

- Convert anticipated R^2 expectations to f^2 using the formulas above; the conversion is non-linear and a "small R^2" of 0.05 corresponds to $f^2 \approx 0.053$, while a "medium R^2" of 0.13 corresponds to $f^2 \approx 0.15$.
- Small incremental effects ($f^2 < 0.05$) require very large samples to detect; weigh the cost-benefit honestly and consider whether a smaller incremental effect is genuinely scientifically interesting.
- Interactions, polynomial terms, and dummy-coded categorical variables each add to the predictor count u and inflate the standard errors of individual coefficients; the sample-size calculation must reflect the actual model structure.
- Pre-specify the minimum detectable incremental effect of interest in the protocol; avoid post-hoc rationalisation of the chosen f^2 to match the achieved power, which is a form of HARKing.
- Include a 15–20 % buffer above the calculated n for residual non-Normality, minor violations of homoscedasticity, and missing data; the calculated n is the minimum, not a target.
- For prediction-focused studies (rather than inferential studies), consider using a learning-curve or repeated-cross-validation framework to determine the sample size required for stable out-of-sample R^2, which is a different question than power for the omnibus test.

5.20.9 R Packages Used

`pwr::pwr.f2.test()` for canonical f^2-based linear-regression power calculation; `WebPower::wp.regression()` for an alternative interface mirroring popular online calculators; `pwrss::pwrss.f.reg()` for extended regression-power tools; `Superpower` for ANOVA and factorial-design power simulation; `simr` for simulation-based power on regression with complex covariance structures.

5.20.10 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.

5.20.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.21 Power for Logistic Regression

5.21.1 Introduction

Power analysis for logistic regression is less standardised than for linear models. Three approaches: (1) events-per-variable rule of thumb, (2) closed-form formulas for a single predictor, (3) simulation for complex models.

5.21.2 Prerequisites

Logistic regression, odds ratio, events-per-variable (EPV).

5.21.3 Theory

EPV rule (Peduzzi et al.): at least 10 events per predictor coefficient is a working minimum. Harrell's 20 EPV is safer.

Single-predictor formula (Demidenko 2007) uses $\log(OR)$ and the baseline event rate:

$$n = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{p_0(1-p_0)\log^2(OR) \cdot \sigma_X^2/\pi(1-\pi)},$$

with simplifications for dichotomous X.

Simulation: specify the full data-generating process, fit the intended model, compute power across many replicates. More reliable than formulas for multiple predictors.

5.21.4 Assumptions

- Binary outcome.
- Logit-linear predictors.
- Sufficient separation of groups.

5.21.5 R Implementation

```r
library(WebPower)

# Single dichotomous predictor: baseline event rate p0 = 0.2, OR = 1.5
wp.logistic(n = NULL, p0 = 0.2, p1 = 0.2 * 1.5 / (1 + 0.2 * (1.5 - 1)),
            alpha = 0.05, power = 0.80,
            family = "Bernoulli", parameter = NULL, alternative = "two.sided")

# Continuous predictor: detect OR = 1.5 per 1-SD increase
wp.logistic(n = NULL, p0 = 0.2, p1 = 0.2,
            alpha = 0.05, power = 0.80,
            family = "normal",
            alternative = "two.sided",
            parameter = log(1.5))

# EPV rule: 15 predictors, min events = 10*15 = 150
# If baseline event rate is 0.20: n = 150 / 0.20 = 750
```

5.21.6 Output & Results

Detecting a modest OR of 1.5 typically requires hundreds of observations; continuous predictors need somewhat less n than dichotomous, given comparable effect magnitude.

5.21.7 Interpretation

“To detect an OR of 1.5 per unit of the standardised predictor at 80 % power and two-sided $\alpha = 0.05$, 250 participants are required (assuming 20 % baseline event rate).”

5.21.8 Practical Tips

- Prefer simulation over formulas when predictors are correlated or the model is complex.
- Rare-event outcomes dramatically inflate required n; power is approximately symmetric around $p_0 = 0.5$.
- For ordinal or multinomial logistic, scale up by roughly $k - 1$ (number of contrasts).
- Penalised logistic (ridge, Firth) handles small samples with separation.
- Missing data and measurement error further increase required n.

5.21.9 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.

5.21.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.22 Power for the Log-Rank Test

5.22.1 Introduction

The log-rank test is the standard non-parametric test of equality of two (or more) survival functions and the most widely used inference tool in survival analysis. A central feature of its power calculation distinguishes it from sample-size analyses for t-tests, ANOVAs, or proportion comparisons: power depends on the total number of observed events across the two groups, not on the sample size directly. Event-based sample-size planning therefore separates the recruitment question (how many subjects to enrol) from the inferential question (how many events to observe), and links them through assumptions about event rates, accrual time, and follow-up duration. This decoupling is essential to realistic planning of survival trials.

5.22.2 Prerequisites

A working understanding of the log-rank test, the Kaplan-Meier survival estimator, the proportional-hazards assumption, and the relationship between hazard ratio, event rate, and follow-up time.

5.22.3 Theory

Under the alternative hypothesis $\mathrm{HR} = \theta$ with equal allocation, the required total number of events is

$$D = \frac{4(z_{1-\alpha/2} + z_{1-\beta})^2}{\log^2 \theta}.$$

For unequal allocation with arm fractions π_1, π_2 summing to one, the multiplier 4 is replaced by $1/(\pi_1 \pi_2)$. The required sample size n depends on the event probability over the trial duration, which in turn depends on the underlying hazard rates, the planned accrual period, and the post-accrual follow-up period. The relationship is captured in standard formulas (Schoenfeld, Freedman, or Lakatos) and implemented in dedicated power-analysis packages.

5.22.4 Assumptions

The proportional-hazards assumption holds (approximately) over the trial duration, censoring is independent and non-informative, the accrual pattern and follow-up duration are pre-specified, and the underlying hazard rates are reasonably well-estimated from prior data.

5.22.5 R Implementation

```r
library(powerSurvEpi); library(gsDesign)

D ← 4 * (qnorm(0.975) + qnorm(0.80))^2 / log(0.65)^2
D

ssizeCT.default(power = 0.80, k = 1,
                pE = 0.40, pC = 0.55,
                RR = 0.65, alpha = 0.05)

gs ← gsDesign(k = 3, test.type = 2, alpha = 0.025, beta = 0.20)
gsBoundSummary(gs)
```

5.22.6 Output & Results

The closed-form calculation gives roughly 88 events required to detect HR = 0.65 with 80 % power at two-sided $\alpha = 0.05$. With assumed event rates of 40 % in the experimental arm and 55 % in the control arm over the planned follow-up window, this corresponds to roughly 185 subjects (just under 100 per arm). The group-sequential design extends this with multiple looks and appropriate alpha-spending.

5.22.7 Interpretation

A reporting sentence: “The trial requires 88 events to detect a hazard ratio of 0.65 with 80 % power at two-sided $\alpha = 0.05$ under the log-rank test. Assuming 24 months of uniform accrual and 12 months of additional follow-up, with expected event rates of 55 % (control) and 40 % (treatment) over the trial duration, the planned enrolment is 180 subjects (90 per arm). The protocol triggers the primary analysis when 88 events have been observed, regardless of calendar time, ensuring the planned power is achieved.” Always state the event-driven analysis trigger.

5.22.8 Practical Tips

- Power depends on the total number of events observed, not on the number of subjects enrolled; plan accrual and follow-up to deliver the required event count, and trigger the primary analysis when that count is reached rather than at a fixed calendar time. This is why survival trials report “event-driven” stopping rules.
- Non-proportional hazards (delayed effects, crossing survival curves, time-varying treatment effects) reduce the log-rank test’s power; Fleming-Harrington weighted log-rank tests improve detection of early or late differences and can be pre-specified for trials where non-proportionality is expected.
- For interim monitoring, use alpha-spending group-sequential designs (`gsDesign`, `rpact`) with information-fraction-based analysis timing — typically every $D/4$ events for a four-stage design — rather than fixed calendar-time analyses.
- Report the assumed accrual pattern and expected follow-up duration explicitly; they determine the event yield from a given enrolment target, and trial outcomes vary substantially when these assumptions are wrong.
- For competing risks, the log-rank test is not the appropriate inferential tool; use the Fine-Gray subdistribution-hazard test and its corresponding power-analysis tools

(`cmprsk`, `crrSC::power.crr()`).
- Sensitivity analysis over a range of plausible hazard ratios is standard; protocols with a single point HR are increasingly flagged by reviewers, who expect a defensible bracket.

5.22.9 R Packages Used

`powerSurvEpi::ssizeCT.default()` and `powerSurvEpi::powerCT.default()` for canonical log-rank sample-size and power calculations; `gsDesign` and `rpact` for group-sequential survival-trial design with alpha-spending; `nph` for non-proportional-hazards weighted log-rank power; `Hmisc::cpower()` for an alternative interface; `survival::survdiff()` for log-rank analysis after data collection.

5.22.10 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.

5.22.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.23 Power for McNemar's Test

5.23.1 Introduction

Power for McNemar's test depends on the rate of discordant pairs, not on the overall sample size alone. The concordant pairs contribute nothing to the test, so studies with high agreement need proportionally more pairs.

5.23.2 Prerequisites

McNemar's test, paired binary data.

5.23.3 Theory

Let p_{10} = probability of (+, -) pairs and p_{01} = probability of (-, +). The null hypothesis is $p_{10} = p_{01}$. Under H_1, the total proportion of discordant pairs is $p_{\text{disc}} = p_{10} + p_{01}$; the odds ratio of a "+ on test 1" given discordance is p_{10}/p_{01}.

For a two-sided test at α and power $1 - \beta$:

$$n \approx \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{p_{\text{disc}} \cdot \left(\frac{p_{10}-p_{01}}{p_{10}+p_{01}}\right)^2}.$$

As discordance drops, n grows rapidly.

5.23.4 Assumptions

- Independent paired observations.
- Pre-specified p_{10}, p_{01} from pilot.

5.23.5 R Implementation

```r
library(pwrss)

# Expected p10 = 0.15, p01 = 0.05, alpha = 0.05, power = 0.80
pwrss.z.mcnemar(p10 = 0.15, p01 = 0.05,
                alpha = 0.05, power = 0.80)

# Manual calculation
p10 ← 0.15; p01 ← 0.05
p_disc ← p10 + p01
OR ← (p10 - p01) / (p10 + p01)
n_manual ← (qnorm(0.975) + qnorm(0.80))^2 / (p_disc * OR^2)
n_manual
```

5.23.6 Output & Results

$n \approx 79$ pairs required. If discordance is lower (say $p_{10} = 0.10$, $p_{01} = 0.05$), required n roughly doubles.

5.23.7 Interpretation

“With an expected proportion of (+, -) pairs of 0.15 and (-, +) pairs of 0.05, McNemar’s test requires 79 paired observations for 80 % power at two-sided $\alpha = 0.05$.”

5.23.8 Practical Tips

- Plan for the total sample, not just discordant pairs; concordant pairs are expected but non-informative.
- The formula is sensitive to the assumed discordance rates; sensitivity analysis is essential.
- For very high agreement (concordant pairs >> discordant), large samples are needed; consider redesigning the comparison.
- Exact McNemar is more conservative in small samples; simulate if exactness matters.
- Extension to Bowker or Stuart-Maxwell for multi-category paired data requires simulation-based power.

5.23.9 Reporting

Always report the assumed discordance rates alongside the resulting sample size. Reviewers and ethics committees expect to see how the number of pairs was derived from the marginal probabilities, because the same total sample can yield very different power depending on the split between p_{10} and p_{01}. Where the pilot estimate of discordance is uncertain, present a small grid of n across plausible values rather than a single point estimate, and state which value drove the final target. If a sequential design is anticipated, note that the formula above is for a single fixed analysis and that group-sequential or adaptive variants require separate boundary calculations.

5.23.10 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.

5.23.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.24 Power for Non-Inferiority Trials

5.24.1 Introduction

Non-inferiority trials aim to establish that a new treatment is "not meaningfully worse" than an active comparator. The non-inferiority margin is a pre-specified loss that defines the boundary. Power calculations use a one-sided test at the margin.

5.24.2 Prerequisites

Equivalence testing, confidence intervals.

5.24.3 Theory

For continuous outcomes with true difference δ (new minus reference), margin $-\Delta$ (loss), and pooled SD σ, the non-inferiority null is $\mu_{\text{new}} - \mu_{\text{ref}} \leq -\Delta$. Power at α one-sided is

$$1 - \beta = \Phi\left(\frac{\delta + \Delta}{\sigma\sqrt{2/n}} - z_{1-\alpha}\right).$$

For binary outcomes (proportions), analogous formulas using Cohen's h or risk-difference variance apply.

Sample size increases with tighter margin, larger allowed effect, or reduced power.

5.24.4 Assumptions

- Pre-specified non-inferiority margin based on regulatory or clinical guidance.
- Assumed true δ (often zero, assuming equivalence).
- Normal or large-sample approximations.

5.24.5 R Implementation

```r
library(gsDesign)

# Continuous outcome: margin -5 units, true delta = 0, sigma = 15
# One-sided alpha = 0.025, power = 0.80
delta  <- 0; margin <- 5; sigma <- 15; alpha <- 0.025; beta <- 0.20
n_per_arm <- 2 * sigma^2 * (qnorm(1 - alpha) + qnorm(1 - beta))^2 / (delta + margin)^2
n_per_arm

# Binary outcome: reference rate 70% cure, margin -10%, true new = 70%
pC <- 0.70; pE <- 0.70; margin_prop <- 0.10
# Farrington-Manning or score-based calculations are standard; simplified Wald here:
p_bar <- (pC + pE) / 2
n_bin <- (qnorm(1 - alpha) + qnorm(1 - beta))^2 *
         (pC*(1-pC) + pE*(1-pE)) / (pE - pC + margin_prop)^2
n_bin

# Group-sequential non-inferiority design
ni <- gsDesign(k = 2, test.type = 1, alpha = 0.025, beta = 0.20,
               sfu = "OF")
gsBoundSummary(ni)
```

5.24.6 Output & Results

Continuous: $n \approx 140$ per arm. Binary: $n \approx 196$ per arm.

5.24.7 Interpretation

“To establish non-inferiority with a 5-unit margin, assuming a true zero difference and SD 15, at 80 % power and one-sided $\alpha = 0.025$, 140 patients per arm are required.”

5.24.8 Practical Tips

- Margin selection is the most important decision in non-inferiority design; regulators require explicit justification.
- Use one-sided $\alpha = 0.025$ to keep consistency with two-sided 0.05 superiority conventions.
- Report per-protocol analysis alongside intention-to-treat; dilution effects in ITT make non-inferiority easier to show artificially.
- Include interim futility stops for ethical reasons.
- Conservative margins and high power requirements inflate n quickly.

5.24.9 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.

5.24.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.25 Power for One-Proportion Test

5.25.1 Introduction

Power analysis for the one-proportion test computes the sample size required to detect a departure from a null proportion p_0 with adequate probability. The one-proportion test arises constantly in clinical research — testing whether a single-arm response rate exceeds a historical control benchmark, whether a quality-control defect rate departs from a tolerance threshold, whether a survey-derived proportion differs from a reference. Two complementary approaches are widely used: the Normal-approximation power formula, which gives clean closed-form expressions in terms of Cohen's h effect size, and exact-binomial power, which enumerates the binomial distribution and is preferred for small samples or extreme proportions where the Normal approximation is unreliable.

5.25.2 Prerequisites

A working understanding of the one-proportion test, the binomial distribution, the Normal approximation to the binomial, and Cohen's h as the standardised effect-size measure for proportion tests.

5.25.3 Theory

For the test of $H_0 : p = p_0$ versus $H_1 : p = p_1$, Cohen's standardised proportion-difference effect size is

$$h = 2 \arcsin \sqrt{p_1} - 2 \arcsin \sqrt{p_0},$$

which arc-sine transforms each proportion onto a scale where its variance is approximately constant. The Normal-approximation sample-size formula at two-sided α and power $1 - \beta$ is

$$n \approx \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{h} \right)^2.$$

Cohen's conventional benchmarks for h are 0.20 (small), 0.50 (medium), and 0.80 (large). Exact binomial power evaluates the binomial CDF under p_1 and sums probabilities in the rejection region defined by the binomial CDF under p_0.

5.25.4 Assumptions

Trials are independent Bernoulli with constant success probability; for the Normal-approximation formula, $np_0 \geq 10$ and $n(1-p_0) \geq 10$ to ensure the approximation is adequate; for exact binomial power these conditions are not needed.

5.25.5 R Implementation

```
library(pwr)

h ← ES.h(p1 = 0.40, p2 = 0.25)
pwr.p.test(h = h, sig.level = 0.05, power = 0.80, alternative = "two.sided")

exact_binom_power ← function(n, p0, p1, alpha = 0.05) {
  k_reject ← sum(dbinom(0:n, n, p0) ⩽ alpha)
  crit_lo ← qbinom(alpha/2, n, p0)
  crit_hi ← qbinom(1 - alpha/2, n, p0)
  power_lo ← pbinom(crit_lo - 1, n, p1)
  power_hi ← 1 - pbinom(crit_hi, n, p1)
  power_lo + power_hi
}
exact_binom_power(n = 80, p0 = 0.25, p1 = 0.40)
```

5.25.6 Output & Results

`pwr.p.test()` returns the sample size required under the Normal-approximation formula; the custom exact-binomial function then verifies the calculation by direct enumeration. For the example, the Normal approximation gives $n = 82$ and the exact calculation confirms 80 % power at $n = 80$ — close agreement that is typical when the Normal-approximation conditions are satisfied.

5.25.7 Interpretation

A reporting sentence: "To detect a response rate of 40 % against a historical-control null of 25 % with 80 % power at two-sided $\alpha = 0.05$, $n = 82$ participants are required (Normal-approximation calculation, Cohen's $h = 0.32$). Exact binomial power at $n = 80$ is 80 %, confirming the Normal-approximation result. The protocol enrols 90 to allow for 10 % attrition. Sensitivity analyses across $p_1 \in [0.35, 0.45]$ are reported in the supplement." Always report sensitivity over plausible p_1.

5.25.8 Practical Tips

- For small n (typically < 30) or extreme proportions (very close to 0 or 1), prefer exact binomial power calculations to the Normal approximation; the approximation is unreliable in these regimes and can give misleading sample-size recommendations.
- Cohen's h ranges in $[0, \pi]$ with conventional benchmarks 0.20 (small), 0.50 (medium), and 0.80 (large); these are guidelines for translating effect-size language

into sample-size calculations and should be tied to substantive clinical or scientific meaning.
- One-sided tests require less sample size than two-sided tests if the direction of effect is pre-specified and scientifically justified; the trade-off is that a result in the unexpected direction cannot be reported as significant.
- Sensitivity analysis over a range of plausible p_1 values is standard; protocols with a single point estimate of p_1 are increasingly flagged by reviewers, who expect a defensible bracket and a justification of the lower bound used for the sample-size determination.
- For p_0 near 0.5, the test is approximately symmetric and the variance is maximised, so power per unit of n is at its maximum; for p_0 near 0 or 1 the variance is small and the test can be very efficient at detecting small absolute differences.
- Continuity-correction options (Yates correction) marginally affect Normal-approximation power calculations; modern software defaults vary, so check what your sample-size tool is computing.

5.25.9 R Packages Used

`pwr::pwr.p.test()` and `pwr::ES.h()` for canonical Cohen's h Normal-approximation calculations; `binom::power.binom()` for exact binomial sample-size and power; `Hmisc::bpower()` for an alternative implementation; `MESS::power_prop_test()` for fast exact and approximate calculations; `Superpower` and `simr` for simulation-based extensions.

5.25.10 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.

5.25.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.26 Power for One-Sample t-Test

5.26.1 Introduction

Power analysis for the one-sample t-test determines the sample size required to detect a specified departure of the population mean from a reference value with a given probability, or equivalently the power achievable at a fixed sample size. The one-sample t-test arises in many practical contexts: validating a quality-control measurement against a target, comparing a single-arm pilot study against a historical benchmark, testing whether a

within-subject change differs from zero. In every case, the planning question is the same — given the smallest effect of scientific interest and the assumed within-population standard deviation, how many observations are needed for an adequately-powered test? Pre-specifying the calculation in the protocol is now standard for any prospective study.

5.26.2 Prerequisites

A working understanding of the one-sample t-test, the concept of statistical power and type-II error, and Cohen's d as the standardised effect-size measure for one-mean tests.

5.26.3 Theory

Under the alternative hypothesis $H_1 : \mu = \mu_1 \neq \mu_0$, the t-statistic follows a non-central t-distribution with $n - 1$ degrees of freedom and non-centrality parameter

$$\lambda = \frac{\mu_1 - \mu_0}{\sigma / \sqrt{n}} = d\sqrt{n},$$

where $d = (\mu_1 - \mu_0)/\sigma$ is Cohen's standardised effect. Power is the probability $P(|T| > t_{1-\alpha/2, n-1})$ computed under the non-central t-distribution. Closed-form sample-size solutions invert this expression to produce the smallest n achieving target power against a specified d, α, and one-sided / two-sided test direction.

5.26.4 Assumptions

The population is approximately Normal or the sample size is large enough for the central limit theorem to apply, the within-population standard deviation σ is reasonably known (or has been estimated from pilot data with adequate precision), and the effect size d is pre-specified — usually as the smallest clinically or scientifically meaningful difference rather than the expected value.

5.26.5 R Implementation

```r
library(pwr)

pwr.t.test(d = 0.4, sig.level = 0.05, power = 0.80, type = "one.sample")

pwr.t.test(n = 25, sig.level = 0.05, power = 0.80, type = "one.sample")

n_grid <- seq(10, 100, by = 5)
pw <- sapply(n_grid, function(n)
  pwr.t.test(n = n, d = 0.4, type = "one.sample")$power)
plot(n_grid, pw, type = "l", lwd = 2, col = "#2A9D8F",
     xlab = "n", ylab = "Power", main = "One-sample t-test, d = 0.4")
abline(h = 0.80, lty = 2)
```

5.26.6 Output & Results

`pwr.t.test()` returns the sample size required to achieve target power against a specified d, the achievable power for a given n, or the minimum detectable effect size for a given n and target power. Plotting the power curve against n for the chosen d visualises the trade-off and is a useful supplement to the single-point calculation.

5.26.7 Interpretation

A reporting sentence: "A sample of 51 participants is required to detect a standardised effect of $d = 0.4$ with 80 % power at $\alpha = 0.05$ (two-sided) using a one-sample t-test against the reference value 10. With the planned sample of 25 participants, the minimum detectable effect at 80 % power would be $d = 0.58$, which exceeds the smallest clinically meaningful effect of 0.4 and would leave the study underpowered. The protocol therefore stipulates 51 participants, with 60 enrolled to allow for 15 % attrition." Always state d, α, power, and the attrition margin.

5.26.8 Practical Tips

- The one-sample t-test power formula applies equally to the paired t-test if the effect size d is computed using the within-subject difference SD rather than the between-subject SD; this is the most common practical use of the calculation.
- For very small effect sizes, sample sizes grow rapidly: halving d quadruples the required n because power scales with $d\sqrt{n}$. A clear-headed choice of the smallest clinically meaningful d is therefore essential, and pilot-data d estimates should be treated with appropriate uncertainty.
- One-sided power at α approximately equals two-sided power at 2α when the true effect is in the hypothesised direction; the rule of thumb is useful for quick comparisons but exact non-central t computation is preferred for the protocol calculation.
- Choose d on the basis of the minimum clinically important difference (MCID) or smallest scientifically meaningful effect, not the expected effect from pilot data; powering for the expected effect leads to under-powered trials when the truth is more modest.
- Inflate the calculated sample size for expected attrition, dropout, or non-evaluable observations; a 15–20 % inflation is typical for clinical studies, with field-specific norms varying.
- Distinguish achievable power (post-hoc, computed from observed effect size) from planned power (a priori); only planned power is methodologically valid, and "post-hoc power" calculations are widely discouraged because they are mathematically determined by the observed p-value.

5.26.9 R Packages Used

`pwr::pwr.t.test()` for the canonical one-sample and paired t-test power calculation; `WebPower::wp.t()` for an alternative interface mirroring popular online calculators; `Superpower` for ANOVA and factorial-design power simulation; `simr` for power analysis

via simulation when classical formulas do not apply; `MESS::power_t_test()` as a fast alternative.

5.26.10 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.

5.26.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.27 Power for Paired t-Test

5.27.1 Introduction

Paired designs achieve their statistical power gain by cancelling between-subject variability from the treatment contrast. Because each subject acts as their own control, the between-subject component of variance — typically the largest source of noise in clinical-pharmacology, behavioural, and clinical-laboratory studies — drops out of the within-subject difference. The result is that paired designs are substantially more powerful than independent-groups designs of equal total sample size, with the benefit growing as the within-subject correlation increases. Power analysis for the paired t-test must therefore explicitly account for this correlation; ignoring it under-uses the design's main analytic advantage and inflates the planned sample size unnecessarily.

5.27.2 Prerequisites

A working understanding of the paired t-test, the relationship between within-subject correlation ρ and difference standard deviation σ_D, and the standardised effect-size measures Cohen's d (between-subject) and d_z (within-subject paired).

5.27.3 Theory

For paired differences $D_i = X_i - Y_i$ with within-subject correlation ρ, the difference standard deviation is

$$\sigma_D = \sigma\sqrt{2(1-\rho)},$$

and the appropriate paired-design effect size is

$$d_z = \frac{\mu_D}{\sigma_D}.$$

Higher within-subject correlation reduces σ_D, inflating d_z for a fixed raw effect, and dramatically reducing the required sample size. With $\rho = 0$ the paired design reduces to the independent-groups case (at half the per-condition sample size); with $\rho \to 1$ the paired design approaches infinite efficiency.

5.27.4 Assumptions

The design genuinely produces paired observations (each subject contributes one value per condition), the differences are approximately Normal (or the sample is large enough for the central limit theorem), and the within-subject correlation is reasonably known from pilot data or literature.

5.27.5 R Implementation

```r
library(pwr)

pwr.t.test(d = 0.4, sig.level = 0.05, power = 0.80, type = "paired")

sigma_diff ← function(sigma, rho) sigma * sqrt(2 * (1 - rho))
d_z_from_d ← function(d, rho) d / sqrt(2 * (1 - rho))

d_z_from_d(d = 0.5, rho = 0.7)

pwr.t.test(d = 0.5, power = 0.80, type = "two.sample")$n
pwr.t.test(d = d_z_from_d(0.5, 0.7), power = 0.80, type = "paired")$n
```

5.27.6 Output & Results

The script reports the required number of pairs at $d_z = 0.4$, the conversion from raw d to paired d_z at correlation 0.7, and the side-by-side sample-size comparison between independent-groups and paired designs for the same raw effect. The paired design needs roughly 21 pairs (42 observations) versus 128 observations for the independent comparison — a six-fold efficiency gain at $\rho = 0.7$.

5.27.7 Interpretation

A reporting sentence: “Assuming a within-subject correlation of 0.7 (estimated from pilot data) and a raw mean difference of 0.5 SD, the paired design requires 21 paired observations for 80 % power at $\alpha = 0.05$ (two-sided), corresponding to a paired effect size $d_z = 0.65$. The same raw effect would require 128 observations (64 per arm) in an independent-groups design — a 6-fold efficiency gain. The protocol therefore enrols 25 subjects to accommodate up to 15 % attrition.” Always report the assumed ρ and the efficiency gain.

5.27.8 Practical Tips

- Pre-specify the expected within-subject correlation ρ from pilot data or published literature, and conduct a sensitivity analysis over a plausible range — typically $\rho \in [0.3, 0.8]$. The sample-size calculation is sensitive to this assumption, and protocols should defend the chosen value.
- When $\rho = 0$ the paired design reduces to the independent-groups case at half the per-condition sample size; the paired design is genuinely advantageous only when within-subject correlation is meaningfully positive, which is almost always the case in repeated-measures contexts.
- Report the effect size in paired d_z form (where the denominator is the difference SD) rather than two-sample d form (where the denominator is the within-group SD); the two are commonly confused and lead to apparent disagreements in power calculations across software packages.
- Missing one member of a pair drops the entire pair from the analysis; this attrition pattern is more wasteful than in independent-groups designs and should be reflected in the sample-size inflation factor.
- For more than two repeated measures (three time points, multiple conditions per subject), use repeated-measures ANOVA or mixed-effects power calculations rather than treating the design as a series of pairwise paired tests; the multi-condition analysis is more efficient and avoids multiplicity adjustments.
- Pilot estimates of ρ have substantial uncertainty in small samples; using the lower end of a CI on ρ for power planning is a conservative approach that protects against an over-optimistic point estimate.

5.27.9 R Packages Used

`pwr::pwr.t.test(type = "paired")` for the canonical paired-design power calculation in d_z units; `WebPower::wp.t()` with paired specification for an alternative interface; `Superpower` for paired-design ANOVA and factorial extensions; `simr` for simulation-based power analysis when the design extends to mixed-effects models with random slopes; `BUCSS` for bias- and uncertainty-corrected sample-size estimation when pilot data are limited.

5.27.10 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.

5.27.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.28 Power for Repeated-Measures ANOVA

5.28.1 Introduction

Repeated-measures designs achieve substantial statistical-power gains over independent-groups designs by exploiting the within-subject correlation across the multiple measurements per subject. Each subject contributes information at every time point or condition, so the within-subject component of variance — typically the smaller of the two variance components — drives the sensitivity of the test. Power calculations for repeated-measures ANOVA must therefore explicitly account for the within-subject correlation; ignoring it leads to substantially over-estimated sample-size requirements and wastes resources, while assuming an unrealistically high correlation under-powers the trial when reality intervenes.

5.28.2 Prerequisites

A working understanding of repeated-measures ANOVA, the sphericity assumption and its diagnostics, Cohen's f as the standardised effect-size measure for ANOVA, and the role of within-subject correlation in design efficiency.

5.28.3 Theory

For a within-subjects design with k measurements per subject, effect size f, and average within-subject correlation ρ, the non-centrality parameter of the omnibus F-test is approximately

$$\lambda = n f^2 \frac{k}{1-\rho},$$

where n is the number of subjects. Power is the tail probability of the non-central F distribution evaluated at the conventional critical value. Higher ρ inflates λ and reduces required sample size; near-zero ρ gives little efficiency advantage over an equivalent independent-groups ANOVA. The Cohen-conventional benchmarks for f are 0.10 (small), 0.25 (medium), and 0.40 (large).

5.28.4 Assumptions

Sphericity holds (equal pairwise variances of differences across time points), residuals are approximately Normal, and the within-subject correlation is reasonably known from pilot data or literature. Sphericity violations require Greenhouse-Geisser or Huynh-Feldt corrections, which slightly inflate the required sample size.

5.28.5 R Implementation

```r
library(WebPower)

wp.rmanova(f = 0.25, ng = 1, nm = 4, nscor = 1,
           alpha = 0.05, power = 0.80, type = 1)

rho_grid ← seq(0.2, 0.9, by = 0.1)
ns ← sapply(rho_grid, function(rho) {
  wp.rmanova(f = 0.25, ng = 1, nm = 4, nscor = rho,
             alpha = 0.05, power = 0.80, type = 1)$n
})
data.frame(rho = rho_grid, n = round(ns))
```

5.28.6 Output & Results

`wp.rmanova()` returns the required sample size at specified f, k, ρ, and target power. The sensitivity table shows required n falling sharply as within-subject correlation increases — from 54 subjects at $\rho = 0.2$ to only 8 at $\rho = 0.9$. This dramatic dependence makes pre-specification of ρ (and a defensible sensitivity range) essential to honest sample-size planning.

5.28.7 Interpretation

A reporting sentence: "For a within-subjects design with four time points, a medium effect ($f = 0.25$), and assumed within-subject correlation 0.5 from pilot data, $n = 34$ subjects provide 80 % power at $\alpha = 0.05$ for the omnibus F-test under the sphericity assumption. Sensitivity analyses across $\rho \in [0.30, 0.70]$ yield required n from 47 to 23 respectively; the protocol enrols 50 to accommodate the most conservative scenario plus 15 % attrition." Always state ρ and report sensitivity.

5.28.8 Practical Tips

- If sphericity is suspected to fail, inflate the calculated sample size by 10–30 % to compensate for the Greenhouse-Geisser or Huynh-Feldt corrections that will be applied at analysis time; the corrections reduce effective degrees of freedom and therefore reduce achievable power.
- For mixed designs combining between-subjects and within-subjects factors, use `wp.rmanova(type = 2)` or `type = 3` for the appropriate stratum, or use simulation for non-standard designs; classical formulas can be brittle in mixed designs with complex interaction structures.
- Higher numbers of repeated measurements per subject inflate power dramatically because each subject contributes more information; planners often face a trade-off between recruiting more subjects (expensive, slow) and measuring each subject more times (cheap, fast).
- When the primary scientific question is a specific contrast (a linear trend over time, a difference between two conditions), a contrast-specific power calculation typically yields more power than the omnibus F-test and should be used in preference; the omnibus test diffuses power across all interactions.

- For non-Normal outcomes, unbalanced data, or designs with informative dropout, simulation via a linear mixed-effects model (using `simr::powerSim()`) is the gold-standard approach; classical formulas assume a balanced complete-data design and Normality.
- Pilot estimates of ρ have substantial uncertainty in small samples; use the lower end of a reasonable CI on ρ for protective sample-size planning rather than the point estimate.

5.28.9 R Packages Used

`WebPower::wp.rmanova()` for canonical repeated-measures-ANOVA power across one-way and mixed designs; `pwr` family of functions for related ANOVA-power tools; `Superpower::ANOVA_power()` for ANOVA-design power simulation including repeated-measures structures; `simr::powerSim()` for simulation-based power on linear mixed-effects models with arbitrary covariance structure; `afex::aov_car()` for fitting the underlying ANOVA model with sphericity correction.

5.28.10 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.

5.28.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.29 Power for Stepped-Wedge Trials

5.29.1 Introduction

Stepped-wedge cluster-randomised trials (SW-CRT) progressively introduce an intervention across clusters over time until every cluster receives it. The design is efficient when the intervention is expected to help and must be rolled out anyway. Power calculations use the Hussey-Hughes (2007) formula.

5.29.2 Prerequisites

Cluster-RCT, ICC, time effects.

5.29.3 Theory

For I clusters and J time periods (with one cluster per step crossing over), the Hussey-Hughes variance of the treatment effect is complex and depends on ICC, cluster-period correlation, cluster size, and total steps. A key insight: adding time periods increases efficiency because each cluster contributes both control and intervention observations.

Software (`swCRTdesign`, `SWSamp`) computes the needed number of clusters given ICC, cluster-autocorrelation, cluster size, and effect.

5.29.4 Assumptions

- Fixed time periods common to all clusters.
- No lag from randomisation to intervention effect (or explicit lag period).
- Underlying secular time trend appropriately modelled.
- ICC and cluster-autocorrelation pre-specified.

5.29.5 R Implementation

```r
library(swCRTdesign)

# Design: 8 clusters, 6 time periods, cluster size 20, ICC = 0.05
swDsn ← swDsn(clusters = c(rep(1, 7), 2))
swDsn

# Power for effect size 0.3, sigma = 1
swPwr(design = swDsn,
      distn = "gaussian",
      n = 20,
      mu0 = 0, mu1 = 0.3,
      tau = 0.05, sigma = 1,
      alpha = 0.05,
      retDATA = FALSE)
```

5.29.6 Output & Results

Power around 0.85 for 8 clusters x 6 periods x 20 subjects/cluster at effect 0.3, ICC 0.05. The exact number depends heavily on the assumed autocorrelation.

5.29.7 Interpretation

“With 8 clusters, 6 time periods, cluster size 20, ICC 0.05, and an assumed intervention effect of 0.3 SD, the stepped-wedge trial achieves 85 % power at $\alpha = 0.05$.”

5.29.8 Practical Tips

- Stepped-wedge designs trade statistical efficiency for practical feasibility; a parallel CRT is usually more powerful at equal total observations.
- The cluster-period autocorrelation is a second nuisance parameter often under-specified; include sensitivity analysis.
- Longer follow-up and more time periods dramatically improve efficiency.

- Analyse with mixed models that include cluster, time-period, and (usually) cluster-by-period random effects.
- Published tools (`swCRTdesign`, `SWSamp`, online calculators from Hemming and Girling) differ slightly; cross-check.

5.29.9 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.

5.29.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.30 Power for Two-Proportion Test

5.30.1 Introduction

Comparing proportions between two independent groups is one of the most common designs in clinical research, public-health epidemiology, and quality-control. Whenever the primary outcome is binary — event vs no event, response vs non-response, success vs failure — the trial's sample-size calculation reduces to a two-proportion power analysis. Several closely-related approaches are widely used: Cohen's arcsine-transformed effect size h that maps cleanly to a Normal-approximation formula, direct sample-size formulas in the original-proportion scale (e.g., Kelsey's formula with continuity correction), and exact Fisher-based methods for rare events. The choice depends on the expected event rates, the desired allocation ratio, and the regulatory or methodological context.

5.30.2 Prerequisites

A working understanding of the two-proportion test, the arcsine variance-stabilising transformation, and Cohen's h as the standardised effect-size measure for proportion-difference designs.

5.30.3 Theory

Cohen's h is the arcsine-transformed proportion difference

$$h = 2\arcsin\sqrt{p_1} - 2\arcsin\sqrt{p_2}.$$

For balanced two-arm designs, the sample size per group at two-sided α and power $1-\beta$ is approximately

$$n \approx 2\left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{h}\right)^2 \text{ per group.}$$

For unequal allocation with ratio $k = n_2/n_1$, the total sample size grows by a factor $(1 + 1/k)(1 + k)/4$ relative to the balanced minimum. Cohen's benchmarks for h are 0.20 (small), 0.50 (medium), and 0.80 (large).

5.30.4 Assumptions

The two groups are independent, the standard Normal-approximation conditions $np \geq 10$ and $n(1-p) \geq 10$ are satisfied for each group, and the proportions under the alternative hypothesis are pre-specified.

5.30.5 R Implementation

```
library(pwr); library(Hmisc)

pwr.2p.test(h = ES.h(0.20, 0.10), sig.level = 0.05, power = 0.80)

bsamsize(p1 = 0.20, p2 = 0.10, alpha = 0.05, power = 0.80)

pwr.2p2n.test(h = ES.h(0.20, 0.10), n1 = 100, sig.level = 0.05)
n1_fix ← 100
n2_vals ← seq(40, 300, 10)
pw ← sapply(n2_vals, function(n2)
  pwr.2p2n.test(h = ES.h(0.20, 0.10), n1 = n1_fix, n2 = n2)$power)
n2_vals[which(pw ⩾ 0.80)[1]]
```

5.30.6 Output & Results

`pwr.2p.test()` returns the per-group sample size at specified h, α, and power; `Hmisc::bsamsize()` provides the same calculation directly in the original-proportion scale. For $p_1 = 0.20$ vs $p_2 = 0.10$ at 80 % power, $n \approx 199$ per group (398 total). The unequal-allocation analysis shows that fixing $n_1 = 100$ requires roughly $n_2 = 280$ to maintain 80 % power, illustrating the inefficiency of imbalanced designs.

5.30.7 Interpretation

A reporting sentence: "To detect a difference in event rates from 10 % (historical control) to 20 % (active treatment) at $\alpha = 0.05$ two-sided with 80 % power, 199 patients per arm are required (398 total) under a 1:1 allocation. The protocol enrols 440 patients to allow for 10 % attrition. Sensitivity analyses across $p_2 \in [0.15, 0.25]$ yield required per-arm n from 392 to 105, bracketing the assumed effect; the conservative end of this range was used to justify the operational sample size." Always report the assumed proportions and the allocation ratio.

5.30.8 Practical Tips

- Equal allocation (1:1) is most efficient for detecting a difference between two proportions; unequal allocation (2:1, 3:1) costs 10–30 % more total n for the same

power and is justified only when the cost or feasibility of the two arms differs substantially.
- For rare events ($p < 0.01$ in either arm), the Normal-approximation formulas are unreliable; use exact Fisher's-test-based power calculations or simulation, available in `Exact::power.exact.test()` or via Monte Carlo enumeration.
- Continuity correction (Yates) inflates the required sample size slightly and is the historical default for 2×2 chi-squared tests; modern practice often omits it for likelihood-ratio or score tests, so verify what your software is computing.
- For cluster-randomised designs, inflate the calculated n by the design effect $1 + (\bar{m} - 1)\mathrm{ICC}$, where $\bar{m}$ is the average cluster size and ICC is the intraclass correlation; ignoring clustering grossly under-powers cluster trials.
- Cohen's h benchmarks (0.20 / 0.50 / 0.80) are useful for translating effect-size language into sample-size expectations, but tying the calculation to the actual expected proportions via `ES.h(p1, p2)` is more defensible than invoking the benchmarks directly.
- For non-inferiority or equivalence designs, the calculation differs from a superiority test; use `pwr.2p.test()` with the appropriate margin or specialised tools like `bsamsize()` with non-inferiority arguments.

5.30.9 R Packages Used

`pwr::pwr.2p.test()` and `pwr::pwr.2p2n.test()` for canonical Cohen's h Normal-approximation calculations; `Hmisc::bsamsize()` for direct-proportion-scale calculations with optional continuity correction; `Exact::power.exact.test()` for exact-binomial power across small-sample two-proportion designs; `clusterPower` for cluster-randomised two-proportion power; `Mediana` for trial-design simulation including two-proportion power across complex designs.

5.30.10 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.

5.30.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.31 Sample Size for a Two-Sample t-Test

5.31.1 Introduction

Before running a two-group comparison, the investigator needs to know how many observations per group will be required to detect a difference of scientific importance with acceptable probability. Power analysis for the two-sample t-test answers this question by linking four quantities: the true effect size, the significance level, the power, and the sample size. Given any three, R can solve for the fourth.

5.31.2 Prerequisites

Familiarity with the two-sample t-test and with the concepts of type I and type II error is assumed. The reader should know that "power" means $1 - \beta$, the probability of rejecting H_0 when H_1 is true.

5.31.3 Theory

Under H_0, the t statistic follows a central t distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom (Student's test, equal variances). Under H_1 where $\mu_1 - \mu_2 = \delta$, the statistic follows a *non-central* t distribution with non-centrality parameter

$$\lambda = \frac{\delta}{\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}.$$

Power is the probability that the test statistic exceeds the critical value under this non-central distribution. For Cohen's standardised effect size $d = \delta / \sigma$ and equal group sizes $n_1 = n_2 = n$, this simplifies to $\lambda = d\sqrt{n/2}$.

The four power quantities and their trade-offs:

- **Effect size** d: the minimum standardised difference considered scientifically meaningful. By Cohen's conventions, $d = 0.2$ is small, 0.5 is medium, 0.8 is large – but these conventions are field-dependent and should be used with judgement.
- **Significance level** α: usually 0.05 for two-sided testing.
- **Power**: conventionally 0.80 for exploratory work, 0.90 for confirmatory or registered trials.
- **Sample size per group** n: what we typically solve for.

5.31.4 Assumptions

Power analysis inherits the assumptions of the t-test it plans for: independent groups, approximately normal data (or large enough n for the CLT), and equal variances for Student's test (unequal for Welch's).

5.31.5 R Implementation

```r
library(pwr)

pwr.t.test(d = 0.5, sig.level = 0.05, power = 0.80,
           type = "two.sample", alternative = "two.sided")

pwr.t.test(n = 30, sig.level = 0.05, power = 0.80,
           type = "two.sample", alternative = "two.sided")

pwr.t.test(n = 30, d = 0.5, sig.level = 0.05,
           type = "two.sample", alternative = "two.sided")

library(pwrss)
pwrss.t.2means(mu1 = 110, mu2 = 100, sd1 = 18, sd2 = 15,
               power = 0.80, alpha = 0.05)
```

The first `pwr.t.test()` call asks: how many subjects per group do I need to detect a medium-sized effect ($d = 0.5$) with 80% power? The second asks: given $n = 30$ per group, what is the minimum detectable effect at 80% power? The third asks: given $n = 30$ and $d = 0.5$, what is my power? The `pwrss::pwrss.t.2means()` call works directly in raw-mean terms, which is often more natural when planning a clinical trial.

5.31.6 Output & Results

The first call returns $n \approx 64$ per group – 128 subjects in total. The second returns a minimum detectable $d \approx 0.735$, corresponding to a raw difference of about 11 points if $\sigma = 15$. The third returns a power of approximately 0.48, showing that 30 per group is badly underpowered for a medium effect.

```
     Two-sample t test power calculation

              n = 63.76576
              d = 0.5
      sig.level = 0.05
          power = 0.8
    alternative = two.sided
```

5.31.7 Interpretation

The recommended manuscript phrasing is: “Assuming a between-group standardised mean difference of $d = 0.5$ (a medium effect by Cohen’s criteria), a two-sided α of 0.05, and power of 0.80, the required sample size is 64 per group. Allowing for 10% loss to follow-up, we aim to enrol 72 per group (144 total).” The power analysis should be reported in the methods section of the grant or protocol, not buried in an appendix.

5.31.8 Practical Tips

- Base your effect size on published data or a pilot, not on a convention like $d = 0.5$. If the published effect is $d = 0.3$, planning for $d = 0.5$ will leave you underpowered.
- Use Welch’s sample size for unequal variances (`pwr.t2n.test()` or `pwrss.t.2means()` with `sd1 ≠ sd2`). It is slightly more conservative, which is what

you want at the planning stage.
- Plan for loss to follow-up, noncompliance, and data exclusion by inflating the computed n by a realistic percentage (typically 10-25%).
- If the minimum clinically important difference (MCID) is known in raw units, use that directly rather than translating to d. It is more interpretable to clinicians.
- Always produce a sensitivity analysis: a curve of power against n, or a tornado plot of sample size against plausible values of effect size and SD.

5.31.9 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.

5.31.10 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.32 Sample Size Sensitivity Analysis

5.32.1 Introduction

Every power calculation rests on assumptions that the investigator cannot know exactly in advance: the true effect size, the outcome SD, the dropout rate, the ICC in a cluster design. Sensitivity analysis reports sample size across a realistic range of these assumptions, rather than a single point estimate.

5.32.2 Prerequisites

Power analysis.

5.32.3 Theory

A sensitivity analysis is a two-dimensional (or higher) grid of required sample sizes, varying one or two assumptions while holding others fixed. Tabular or graphical (contour / heatmap) presentation lets readers see the robustness of the chosen n.

5.32.4 Assumptions

The chosen assumption grid covers the plausible parameter space. Too-narrow grids miss important regions; too-wide grids are unfocused.

5.32.5 R Implementation

```
library(pwr); library(ggplot2)

# 2D sensitivity: n per group vs d and power
d_grid ← seq(0.2, 0.8, by = 0.05)
power_grid ← c(0.70, 0.80, 0.90)

res ← expand.grid(d = d_grid, power = power_grid)
res$n ← mapply(function(d, p)
  pwr.t.test(d = d, power = p, sig.level = 0.05, type = "two.sample")$n,
  res$d, res$power)

ggplot(res, aes(d, n, colour = factor(power))) +
  geom_line(linewidth = 1) +
  scale_y_log10() +
  scale_colour_manual(values = c("#F4A261", "#6A4C93", "#2A9D8F")) +
  labs(x = "Cohen's d", y = "n per group (log scale)",
       colour = "Power",
       title = "Sample size sensitivity: d x power") +
  theme_minimal()

# Inflation for dropout
n_base ← 64
dropout_rates ← c(0, 0.05, 0.10, 0.15, 0.20)
n_inflated ← round(n_base / (1 - dropout_rates))
data.frame(dropout = dropout_rates, n_enrol = n_inflated)
```

5.32.6 Output & Results

For $d = 0.5$: n ranges from 51 at power 0.70 to 86 at power 0.90. For $d = 0.3$: from 138 to 234. Dropout inflation at 15 % pushes n from 64 to 76 per group.

5.32.7 Interpretation

“Sensitivity analysis shows that for plausible effect sizes (Cohen’s d between 0.35 and 0.55), the required sample size ranges from 72 to 164 per arm at 80 % power. We plan to enrol 180 per arm, accommodating the upper end of this range and allowing for 10 % attrition.”

5.32.8 Practical Tips

- Always report sensitivity analysis alongside the headline sample-size calculation in grant proposals.
- Present the calculation that underpins your chosen n in bold; show neighbouring values for context.
- Two-dimensional grids (effect size vs. power, effect size vs. variance) are typical; higher dimensions confuse readers.
- Include a dropout inflation factor explicitly; reviewers frequently check.
- For adaptive or sequential designs, sensitivity to interim analyses and alpha spending deserves separate reporting.

5.32.9 Reporting

A useful sensitivity table is small enough that a reviewer can read every row and large enough that the chosen design point is clearly bracketed by less and more conservative alternatives. State which inputs were varied and which were held fixed, justify the ranges in terms of pilot data or published evidence, and identify the row that drives the final enrolment target so the audit trail from assumption to commitment is explicit. When dropout inflation is applied, write the formula and the assumed completion proportion in the protocol, because reviewers and ethics committees rely on that line to confirm the analysis sample is consistent with the powered effect.

5.32.10 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.

5.32.11 See also — labs in this chapter

- Sample size, power, and Quarto reporting
- Power: closed-form and simulation

5.33 Week 1, Session 5 — Power: closed-form and simulation

Inference lab using the five-step template: Hypothesis → Visualise → Assumptions → Conduct → Conclude.

5.33.1 Learning objectives

- Run a closed-form power calculation for a two-sample *t*-test with `pwr`.
- Simulate power for a scenario where closed-form solutions do not exist using `simr`.
- Describe the inputs to any power calculation (effect, variance, alpha, design).

5.33.2 Prerequisites

inference and mixed-model basics (we will fit a small `lmer`).

5.33.3 Background

Power is the probability of rejecting the null hypothesis when a specific alternative is true. Four inputs set it: the effect size you care about, the variability of the outcome, the significance level, and the sample size (with the design). Closed-form formulas exist for

standard one- and two-sample tests and for some simple regression settings. For anything more elaborate — clustered data, non-linear models, interim looks, complex survival designs — you simulate.

Simulation-based power is conceptually straightforward: for a given scenario, generate many datasets under the alternative, run the planned analysis on each, and compute the fraction of times you reject the null. The `simr` package does this for models fit with `lme4`, allowing you to vary the sample size or the effect and trace out a power curve.

A useful rule of thumb: if your closed-form calculation is for a model simpler than the one you will actually fit, the closed-form is giving you a lower bound on the sample size you need, not an answer.

5.33.4 Setup

```
library(tidyverse)
library(pwr)
library(lme4)
library(lmerTest)
library(simr)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

5.33.5 1. Hypothesis

Detect a standardised mean difference of $d = 0.4$ between two groups with 80% power at a 5% two-sided significance level. Then power a mixed model with 20 clusters and 10 observations each to detect a fixed effect of 0.3.

5.33.6 2. Visualise

```
ns ← seq(20, 400, by = 20)
pw ← map_dbl(ns, \(n) pwr.t.test(
  n = n / 2, d = 0.4, sig.level = 0.05, type = "two.sample"
)$power)

tibble(n = ns, power = pw) ▷
  ggplot(aes(n, power)) +
  geom_line() +
  geom_hline(yintercept = 0.8, linetype = "dashed", colour = "grey40") +
  labs(x = "Total sample size", y = "Power")
```

5.33.7 3. Assumptions

Closed-form assumes equal variance across arms, independent observations, and the nominal effect size. The simulation below adds random intercepts at the cluster level and therefore relaxes independence.

5.33.8 4. Conduct

```
# Closed-form
res ← pwr.t.test(d = 0.4, power = 0.8, sig.level = 0.05, type = "two.sample")
res
```

```
# Simulation: a small two-arm cluster study
J   ← 20                              # clusters
nj  ← 10                              # obs per cluster
dat ← tibble(
  cluster = factor(rep(seq_len(J), each = nj)),
  arm     = rep(rep(c(0, 1), each = nj), length.out = J * nj),
  y       = rnorm(J * nj)             # placeholder; simr overwrites
)

m0 ← makeLmer(
  y ~ arm + (1 | cluster),
  fixef   = c(0, 0.3),
  VarCorr = 0.5,
  sigma   = 1.0,
  data    = dat
)
m0
```

```
ps ← powerSim(m0, nsim = 50, test = fixed("arm"), progress = FALSE)
ps
```

Fifty simulations is enough for a teaching demo; in practice use at least 1000 for a stable estimate.

5.33.9 5. Concluding statement

> A two-sample *t*-test targeting *d* = 0.4 at 80% power requires about `ceiling(res$n) * 2` participants in total. A mixed model with 20 clusters of size 10 and a fixed effect of 0.3 has estimated power of `round(summary(ps)$mean * 100)` % in this simulation (50 replicates).

5.33.10 Common pitfalls

- Using the power formula for independent observations on a clustered design.
- Quoting an effect size taken from a pilot study without acknowledging its noise.
- Computing power from the smallest clinically important effect and the observed pilot effect interchangeably.
- Simulating with too few replicates.

5.33.11 Further reading

- Cohen J. *Statistical Power Analysis for the Behavioral Sciences*.
- Green P, MacLeod CJ (2016), *simr: an R package for power analysis of generalized linear mixed models*.
- Chow SC et al. *Sample Size Calculations in Clinical Research*.

5.33.12 Session info

```
sessionInfo()
```

5.33.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)

5.34 Week 4, Session 5 — Sample size, power, and Quarto reporting

Workflow lab: Goal → Approach → Execution → Check → Report.

5.34.1 Learning objectives

- Compute analytic sample size for the core inferential designs.
- Produce the same answer by simulation and understand when each approach is preferable.
- Assemble a short Quarto report whose numbers, tables, and figures all regenerate from a single render.

5.34.2 Prerequisites

Courses 1 Weeks 1–4 up to this lab; `pwr`, `gtsummary`, and `broom` installed.

5.34.3 Background

Sample size is one of the few places where applied statistics is consulted *before* the data exist. A study that is too small cannot distinguish a real effect from noise; a study that is too large wastes participants and money. The four ingredients of a sample-size calculation — effect size, variability, the significance level, and power — translate directly between a clinical protocol and a funding application, and no competent ethics committee will approve a protocol without them.

There are two complementary approaches. Closed-form formulae, packaged in `pwr` and friends, are fast, transparent, and correct for the textbook designs: one- and two-sample t-tests, one- and two-proportion tests, correlation, and ANOVA. Simulation-based power, by contrast, is the honest answer for anything the textbooks skip — mixed models, skewed outcomes, interim analyses — and costs only a few minutes of compute. The habit to build is to compute analytically first, then verify by simulation; if the two disagree, either the formula does not apply to your design or your simulation is wrong, and you need to know which.

Reporting is the other half of this lab. Every number in a well-written report should trace back to the code that produced it. Quarto, with `gtsummary` for tables and `broom::tidy()` for model output, makes this nearly automatic.

5.34.4 Setup

```
library(tidyverse)
library(pwr)
library(gtsummary)
library(broom)
set.seed(42)
theme_set(theme_minimal(base_size = 12))
```

5.34.5 1. Goal

Plan a hypothetical two-arm trial comparing a new antihypertensive with placebo, choose a sample size with 80% power to detect a 5 mmHg difference at α = 0.05, verify by simulation, and produce a reporting table we can paste into the protocol.

5.34.6 2. Approach

Closed-form first. Cohen's *d* for a 5 mmHg effect on a scale with SD 10 mmHg is *d* = 0.5 — a medium effect by convention.

```
pwr_calc ← pwr.t.test(
  d = 0.5, power = 0.80, sig.level = 0.05,
  type = "two.sample", alternative = "two.sided"
)
pwr_calc
```

The formula calls for about `ceiling(pwr_calc$n)` per arm.

```
ns ← seq(10, 200, by = 5)
powers ← sapply(ns, function(n) {
  pwr.t.test(n = n, d = 0.5, sig.level = 0.05,
             type = "two.sample")$power
})
tibble(n_per_arm = ns, power = powers) ▷
  ggplot(aes(n_per_arm, power)) +
  geom_line(linewidth = 0.8) +
  geom_hline(yintercept = 0.8, linetype = 2, colour = "grey50") +
  labs(x = "N per arm", y = "Power",
       title = "Power curve for d = 0.5, α = 0.05")
```

5.34.7 3. Execution

Verify by simulation. A function that runs the trial once and returns the p-value:

```r
sim_trial <- function(n_per_arm, delta = 5, sd = 10) {
  placebo <- rnorm(n_per_arm, 0, sd)
  drug    <- rnorm(n_per_arm, -delta, sd)
  t.test(drug, placebo, var.equal = FALSE)$p.value
}

sim_power <- function(n_per_arm, reps = 500, ...) {
  mean(replicate(reps, sim_trial(n_per_arm, ...)) < 0.05)
}

n_try  <- ceiling(pwr_calc$n)
sim_power(n_try, reps = 1000)
```

Within Monte-Carlo error of the analytic 0.80, as expected.

5.34.8 4. Check

Now imagine the data have been collected. We simulate one realisation and fit the planned analysis.

```r
n <- ceiling(pwr_calc$n)
trial <- tibble(
  arm = rep(c("placebo", "drug"), each = n),
  sbp_change = c(rnorm(n, 0, 10), rnorm(n, -5, 10))
)

fit <- t.test(sbp_change ~ arm, data = trial, var.equal = FALSE)
broom::tidy(fit)
```

5.34.9 5. Report

```r
trial |>
  tbl_summary(
    by = arm,
    statistic = list(all_continuous() ~ "{mean} ({sd})"),
    digits    = all_continuous() ~ 1,
    label     = list(sbp_change = "Change in SBP (mmHg)")
  ) |>
  add_p() |>
  modify_caption("**Table 1. Change in systolic blood pressure, by arm.**")
```

> In a simulated two-arm trial (n = `n` per arm), mean change in systolic blood pressure was lower in the drug arm than in placebo by `round(-diff(fit$estimate), 1)` mmHg (95% CI: `round(-fit$conf.int[2], 1)` to `round(-fit$conf.int[1], 1)`; p = `signif(fit$p.value, 2)`).

5.34.10 Common pitfalls

- Plugging a standardised effect size into a sample-size calculator without stopping to think whether it is plausible in your field.
- Quoting “power = 0.8” as a universal target when your decision would justify 0.9 or 0.95.
- Reporting a p-value without the effect size and interval that make it interpretable.

- Pasting numbers into the manuscript by hand — anything that breaks the render trail will eventually break the numbers.

5.34.11 Further reading

- Writing a report.
- Champely S. (2024). *pwr: Basic functions for power analysis.*
- Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences.*

5.34.12 Session info

```
sessionInfo()
```

5.34.13 See also — chapter index

- Methods in this chapter
- Find your method (Chapter 0)