Statistical Test Selection · Universal Cheatsheet

1 Identify the design
Outcome type (continuous / ordinal / categorical) · how many groups · independent vs. paired/repeated · one-tailed or two-tailed hypothesis (H_0, H_1, α).

2 Descriptives + plots
n, mean, SD, median, min–max, skew, kurtosis, 95% CI per group. Boxplots for comparison, QQ-plots & histograms for shape.

3 Check assumptions
Normality: Shapiro–Wilk per group + QQ. Equal variances: Levene. Sphericity (repeated only): Mauchly. Independence: from the design.

4 Choose the test path
Assumptions hold → parametric. Variances unequal → Welch. Non-normal / ordinal / small n → nonparametric. (Trees on pages 2–3.)

5 Omnibus → post-hoc
If 3+ groups and omnibus is significant, run pairwise post-hocs with multiplicity correction (Tukey / Games–Howell / Bonferroni / Dunn).

6 Effect size + CI
Significance ≠ importance. Report d / g, η^2 / ω^2, r, or Cramér's V, with confidence intervals for the difference.

7 Report & conclude
APA style: statistic(df), exact p, effect size, CI. State the decision on H_0 and interpret in plain language for the context.

STEP 1 — WHAT TEST FAMILY? (OUTCOME × DESIGN)

OUTCOME VARIABLE	QUESTION / DESIGN	TEST FAMILY → GO TO
Continuous (scores, time, BP…)	Compare means across groups / timepoints	**t-test / ANOVA family** → **Page 2** decision tree
Continuous, non-normal or **Ordinal** (ranks, Likert)	Compare distributions / medians	**Nonparametric branch** of Page 2 (Mann–Whitney, Wilcoxon, Kruskal–Wallis, Friedman)
Categorical (counts, yes/no)	Proportions, association between categories	**χ^2 family** (χ^2, Fisher, McNemar, Cochran's Q) → **Page 3**
Two continuous variables	Strength of relationship	**Correlation** (Pearson / Spearman) → **Page 3**
One outcome + predictors	Predict / adjust for covariates	**Regression** (linear, logistic) / ANCOVA → **Page 3**

The golden rule: the test is chosen by the **design and the data's behavior**, not by which result you'd like. Fix α (usually 0.05) and the tail (one- vs. two-sided) **before** looking at p-values. Use a one-tailed test only when the hypothesis is directional *a priori* (e.g., "music **increases** comprehension"); then $p_1 = p_2/2$ provided the effect is in the hypothesized direction.

Never skip Step 2–3. Descriptives and plots catch outliers, skew, and entry errors that p-values hide. An assumption check is itself a hypothesis test: its p > .05 means "no evidence of violation," which is what lets you proceed parametrically.

STEP 3 — ASSUMPTION CHECKS (THE GATEKEEPERS)

CHECK	TEST / TOOL	RULE OF THUMB
Normality (per group, or of paired differences)	Shapiro–Wilk `W` + QQ-plot, |skew| < 1	**p > .05 → normal OK** **p ≤ .05 → nonparametric or transform**
Equal variances (independent groups)	Levene's test `W`	**p > .05 → pooled test** **p ≤ .05 → Welch version**
Sphericity (3+ repeated measures)	Mauchly's test `W`	**p > .05 → standard RM-ANOVA** **p ≤ .05 → Greenhouse–Geisser ε correction**
Independence	Design, not a test	Same subject measured twice → **paired/repeated** branch, never the independent one
Expected counts (χ^2 tables)	All expected ≥ 5	**Any < 5 → Fisher's exact**

ROBUSTNESS & SAMPLE-SIZE NOTES

- **CLT cushion:** with n ≥ 30 per group, t/ANOVA tolerate mild non-normality; with small n, Shapiro has low power — lean on QQ-plots and prefer nonparametric if shape looks bad.
- **Welch by default is safe:** when in doubt about equal variances (esp. unequal n), Welch's t / Welch ANOVA lose little power and protect Type I error.
- **Sensitivity analysis:** a strong result should survive the robust alternative (e.g., report Welch ANOVA next to classic ANOVA; run Mann–Whitney next to the t-test). Agreement = confidence.
- **Outliers:** investigate, don't auto-delete. If influential, report results with and without.

Comparing a Continuous Outcome · Decision Tree

Design question · Assumption check · Parametric test · Nonparametric alternative · Post-hoc / follow-up

How many groups / conditions am I comparing?

1 group vs. known value μ_0

- **Shapiro–Wilk on sample** + QQ-plot
  - normal → **One-sample t-test** $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$, df = n–1
  - not normal → **Wilcoxon signed-rank** (or sign test)
- **Effect size: d = $(\bar{x}–\mu_0)/s$** + 95% CI for the mean

2 groups

Independent samples or paired?

- independent → **Shapiro–Wilk per group**
  - both normal → **Levene's test**
    - p>.05 → **Student t** pooled SD
    - p≤.05 → **Welch's t** adj. df
  - non-normal → **Mann–Whitney U** ranks; effect size r
- paired (same subjects) → **Shapiro–Wilk on the differences**
  - normal → **Paired t-test** df = n–1
  - not → **Wilcoxon signed-rank**

Cohen's d (or Hedges' g) + CI for mean difference
decide one- vs. two-tailed from H_1, set in advance

3+ groups

Independent groups or repeated measures?

- independent → **Shapiro per group → Levene**
  - all OK → **One-way ANOVA** F = MSB / MSW
    - → **Tukey HSD** + Hedges' g pairwise
  - variances unequal → **Welch ANOVA**
    - → **Games–Howell**
  - non-normal → **Kruskal–Wallis H**
    - → **Dunn (Bonf.)**
- repeated (within-subjects) → **Shapiro per condition → Mauchly sphericity**
  - OK → **RM-ANOVA** partial η^2
  - sphericity fails → **RM-ANOVA + G–G ε** Greenhouse–Geisser df
  - non-normal → **Friedman test**

Paired t-tests (Bonferroni) / Wilcoxon (Bonf.)
+ Cohen's d per pair

Why never multiple t-tests instead of ANOVA? k groups → k(k–1)/2 comparisons; with 5 groups at α=.05 the family-wise error ≈ $1–0.95^{10}$ ≈ 40%. Run the omnibus first; only if it is significant, do corrected post-hocs. **Two factors at once?** (e.g., method × gender) → two-way ANOVA with interaction; **adjusting for a covariate?** → ANCOVA.

KEY STATISTICS AT A GLANCE

TEST	STATISTIC	DF	TEST	STATISTIC	DF
One-sample t	$t = (\bar{x} - \mu_0) / (s/\sqrt{n})$	n – 1	**One-way ANOVA**	$F = MS_{between} / MS_{within}$	k – 1, N – k
Independent t (pooled)	$t = (\bar{x}_1 - \bar{x}_2) / SE_{pooled}$	$n_1 + n_2 - 2$	**RM-ANOVA**	$F = MS_{condition} / MS_{error}$	(k–1), (k–1)(n–1)
Paired t	$t = \bar{d} / (s_d/\sqrt{n})$ (d = differences)	n – 1	**χ^2**	$\chi^2 = \Sigma (O - E)^2 / E$	(r–1)(c–1)
CI for mean diff	$(\bar{x}_1 - \bar{x}_2) \pm t_{crit} \times SE_{diff}$	—	**Cohen's d**	$d = (\bar{x}_1 - \bar{x}_2) / SD_{pooled}$	—

Categorical Outcomes & Relationships

Categorical outcome (counts / proportions)

What is the design?

- 1 variable vs. expected distribution → **χ^2 goodness-of-fit**
- 2 variables, independent samples → **All expected counts ≥ 5?**
  - yes → **χ^2 independence** df = (r−1)(c−1)
  - no → **Fisher's exact test**
- paired binary (before/after) → **McNemar's test** 3+ repeated binary → Cochran's Q

Effect size: φ (2×2) or Cramér's V · report % and counts

2×2 shortcut: comparing two proportions = z-test for proportions ≡ χ^2 on the 2×2 table. Report risk difference / relative risk / odds ratio with CI when clinically meaningful.

Relationship between two variables

Both continuous?

yes → scatterplot first

Linear pattern & bivariate normal? No big outliers?

- yes → **Pearson r**
- monotonic / ordinal / outliers → **Spearman ρ** (small n, many ties → Kendall τ)

Test H_0: ρ=0 · r^2 = variance explained
correlation ≠ causation

|R| INTERPRETATION (ROUGH GUIDE)	
0.1 – 0.3	weak
0.3 – 0.5	moderate
0.5 – 0.7	strong
> 0.7	very strong

Prediction / adjustment (modeling)

What type is the outcome?

- continuous → **Linear regression** simple / multiple
  - **Residuals: normal, homoscedastic, independent; linearity; no heavy multicollinearity**
- binary → **Logistic regression** report odds ratios + CI
- count / time-to-event → **Poisson / Cox** (advanced)

Group comparison + covariate → ANCOVA
ANOVA is regression with categorical predictors

When to model instead of test: you need to adjust for confounders, use multiple predictors, or estimate the size of an effect per unit change — not just detect "a difference exists."

NONPARAMETRIC ↔ PARAMETRIC EQUIVALENTS (ONE-LINE MAP)

DESIGN	PARAMETRIC	NONPARAMETRIC TWIN	DESIGN	PARAMETRIC	NONPARAMETRIC TWIN
1 sample vs. μ_0	One-sample t	Wilcoxon signed-rank	3+ independent groups	One-way ANOVA	Kruskal–Wallis
2 independent groups	Student / Welch t	Mann–Whitney U	3+ repeated measures	RM-ANOVA	Friedman
2 paired measures	Paired t	Wilcoxon signed-rank	Correlation	Pearson r	Spearman ρ / Kendall τ

SCENARIO → TEST (SELF-CHECK DRILL)

SCENARIO	CORRECT CHOICE	SCENARIO	CORRECT CHOICE
Exam scores across 3 different teaching methods, different students in each	**One-way ANOVA** → Tukey	Anxiety measured on the *same* patients under 5 relaxation conditions	**RM-ANOVA** (Mauchly!) → paired t + Bonf.
Reading comprehension: music group vs. silence group, directional H_1	**Independent t**, one-tailed	Blood pressure before vs. after one treatment, same patients	**Paired t** (Shapiro on differences)
Pain rated 1–10 (ordinal) in two independent groups	**Mann–Whitney U**	Smoker (yes/no) vs. disease (yes/no), one cell expected = 3	**Fisher's exact**

Post-hoc, Effect Sizes & Reporting

WHICH POST-HOC AFTER WHICH OMNIBUS?

OMNIBUS TEST (SIGNIFICANT)	POST-HOC	WHY THIS ONE
One-way ANOVA (equal var.)	**Tukey HSD**	All pairwise, controls family-wise error exactly
Welch ANOVA (unequal var.)	**Games–Howell**	Doesn't assume equal variances / n
RM-ANOVA	**Paired t + Bonferroni** (or Holm)	Pairs are within-subject; Bonferroni: $p_{corr} = p \times m$
Kruskal–Wallis	**Dunn + Bonferroni**	Rank-based pairwise
Friedman	**Wilcoxon + Bonferroni**	Paired rank-based
vs. one control group only	**Dunnett**	Fewer comparisons → more power
Few planned contrasts (a priori)	**Planned contrasts / Holm**	Pre-specified, more powerful than all-pairs

EFFECT SIZES — MAGNITUDE GUIDE

MEASURE (CONTEXT)	SMALL	MEDIUM	LARGE
Cohen's d / Hedges' g (mean difference; g corrects small-n bias)	0.2	0.5	0.8
η^2 / ω^2 (ANOVA variance explained; ω^2 less biased)	.01	.06	.14
r (correlation; also U- and W-based r = Z/√n)	.10	.30	.50
ϕ / Cramér's V (χ^2, df=1)	.10	.30	.50

CI logic: a 95% CI for a difference that **excludes 0** ⇔ $p < .05$ (two-tailed). Wider level (99%) → wider interval. Report the CI of the difference, not just each group's CI — overlapping group CIs do *not* prove "no difference."

APA-STYLE REPORTING TEMPLATES

- **Independent t:** `t(48) = 4.93, p < .001 (one-tailed), d = 1.39, 95% CI of diff [2.49, 5.93]`
- **Paired t:** `t(19) = −4.42, p = .0003, d = −1.21`
- **One-way ANOVA:** `F(2, 57) = 25.23, p < .001, η² = .47 (ω² = .45)` — then Tukey: `PBL > Traditional, Δ = 5.59, p_tukey < .001, g = 2.00`
- **Welch ANOVA:** `F(2, 36.97) = 20.50, p < .001` (note the fractional df)
- **RM-ANOVA:** `Mauchly W = 0.83, p = .95 → sphericity OK; F(4, 76) = 41.2, p < .001, partial η² = .68`
- **Mann–Whitney:** `U = 312, p = .004, r = .41, medians 62.8 vs 50.8`
- **χ^2:** `χ²(1, N = 200) = 9.14, p = .003, φ = .21`
- **Correlation:** `r(58) = .62, p < .001, r² = .38`

Pattern: statistic(df) = value, exact p (or p < .001), effect size, CI — then one plain-language sentence tied to the research question.

COMMON PITFALLS — FINAL CHECKLIST

- ☐ Paired data analyzed with an independent test (or vice versa) — check the design first, always.
- ☐ Multiple uncorrected t-tests instead of omnibus + corrected post-hocs (inflates Type I error).
- ☐ Choosing one-tailed *after* seeing the data — the tail comes from H_1, stated in advance.
- ☐ "p = .06 proves no effect" — absence of evidence ≠ evidence of absence; check power / effect size.
- ☐ Reporting p without effect size and CI — significance says "exists," effect size says "matters."
- ☐ Normality checked on the pooled data instead of per group (or on differences for paired designs).
- ☐ Using χ^2 with expected counts < 5 instead of Fisher's exact.
- ☐ Interpreting correlation as causation, or extrapolating regression beyond the data range.
- ☐ Dropping outliers silently — justify, and show the analysis both ways.
- ☐ Forgetting the decision statement: "Reject / fail to reject H_0 at $\alpha = .05$, therefore …"

POWER & ERRORS — QUICK FACTS

- **Type I (α):** rejecting a true H_0 — controlled by α and multiplicity corrections. **Type II (β):** missing a real effect — controlled by power (1−β, aim ≥ .80).
- Power rises with larger n, larger true effect, lower variability, one-tailed tests, and paired/repeated designs (each subject is their own control).
- Non-significant + small n → likely underpowered, not "proven equal." Report the observed effect size and CI; consider an a-priori power analysis next time.

IF THE DATA ARE SKEWED — THREE HONEST OPTIONS

- **Nonparametric twin** (default choice): Mann–Whitney, Wilcoxon, Kruskal–Wallis, Friedman — tests ranks, no normality needed.
- **Transform** (right-skew: log, √; then re-check Shapiro) — interpret results on the transformed scale.
- **Bootstrap CI** for means/differences — resampling makes no shape assumption; good for reporting effect estimates.