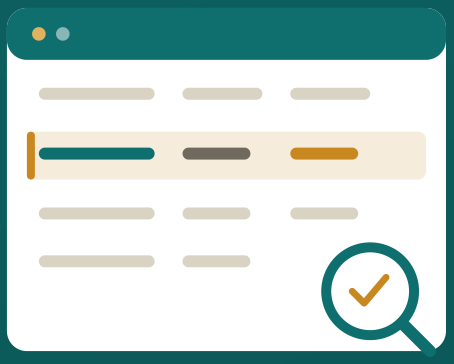
Dissertation Statistics

DISSERTATION STATISTICS FIELD GUIDE

What a P-Value Actually Means

The definition that survives a viva, the inference frameworks you must not blend, how to claim no effect without overreaching, which multiplicity correction to use, and how to show your significant results are not the product of selective testing. From PhD statisticians.



Written by PhD statisticians dissertationstatisticshelp.org

START HERE

A p-value answers one narrow question, and only that one

More results chapters are weakened by a misstated p-value than by any other single error. The number is easy to obtain; what it licenses you to say is where students, and a large share of published papers, go wrong. Said precisely it is defensible; said loosely it invites the exact question every committee knows to ask.

PRINCIPLE

A p-value is the probability of a test statistic at least as extreme as the one observed, computed under the null hypothesis *and* the full model behind it. It conditions on the null and the model; it never measures the probability that the null is true, and a large p-value is consistent with the model being wrong, not just the null being right. The 2016 American Statistical Association statement makes this its first principle for a reason: nearly every misuse starts by forgetting what the number conditions on.

The four misreadings that get flagged

1. **Not the probability the null is true**

   p = 0.03 does not mean a 3 per cent chance the null holds. That posterior requires a prior and Bayes theorem; the p-value supplies neither. Inverting it is the transposed-conditional fallacy, and it is the single error examiners probe most.

2. **Not the probability your result was due to chance**

   The result is fixed. The p-value is the long-run frequency of data this extreme under the null, not a statement about whether chance produced your one finding. The false-positive risk you actually care about depends on power and the prior plausibility of the hypothesis, which the p-value does not contain.

3. **Not a measure of effect magnitude or importance**

   In a large sample a trivial effect attains a tiny p-value; significance reflects detectability times sample size, not size or importance. This is why an effect size and interval are mandatory, not optional.

4. **Not a binary that "absence of evidence" can flip**

   A non-significant result is not evidence for the null. It may simply mean the study lacked power. To argue for no meaningful effect you need an equivalence test or a Bayes factor, not a large p-value.

COHERENCE

Decide whether you are doing Fisher or Neyman-Pearson, and do not blend them

Most muddled significance sections silently mix two incompatible logics. Examiners who know the difference will press exactly here.

TWO FRAMEWORKS, ONE INCOHERENT HYBRID TO AVOID

Framework	What the p-value is for	What you report
Fisherian (evidential)	A continuous measure of evidence against the null	The exact p-value, interpreted on a graded scale, no fixed cutoff
Neyman-Pearson (decision)	A rule that controls long-run error rates	A pre-set alpha, the decision, and the power behind it
The common hybrid (avoid)	A cutoff treated as if exact p also graded evidence	p < .05 stars plus "highly significant" language, with no pre-set alpha or power

WATCH OUT FOR

If you set alpha in advance and frame the study as a decision, then "highly significant" and significance stars are out of place, because the decision is binary at your alpha. If you take the evidential view, then a fixed cutoff is out of place, because evidence is graded. Pick one logic, state it in the methods, and report consistently. A useful intuition for the evidential view is the surprisal, or S-value, equal to minus the base-two logarithm of p: p = 0.05 carries about 4.3 bits of surprise, no more shocking than getting four to five heads in a row.

CLAIMING NO EFFECT

To argue for the null, test for it directly

A non-significant t-test does not support equivalence. If your contribution rests on showing two conditions are practically the same, the burden is on you to test that, against a threshold you defined in advance.

1. **Set a smallest effect size of interest**
   Define the bound below which an effect is practically negligible in your field, before you analyse. This is the equivalence margin, and choosing it after seeing data is not defensible.

2. **Run two one-sided tests (TOST)**
   Equivalence testing rejects the presence of a meaningful effect when the confidence interval for the effect falls entirely inside the equivalence bounds. Report the bound, the interval, and the conclusion in those terms.

3. **Or quantify relative evidence with a Bayes factor**
   A Bayes factor can express how much more the data support the null than a specified alternative. State the prior on the alternative; the answer depends on it, and a reviewer will ask.

PRINCIPLE

"Not significant" and "equivalent" are different claims with different tests. Writing the first and meaning the second is one of the most common over-reaches in a discussion section, and one of the easiest for a committee to catch.

MANY TESTS

Choose the multiplicity correction that matches your goal

Twenty independent tests at alpha 0.05 produce, on average, one false positive. The right adjustment depends on whether you are protecting against any false positive or controlling the share of false discoveries.

WHICH CORRECTION, AND WHEN

Method	Controls	Use it when
Bonferroni	Family-wise error rate	Few tests, each must be near-certain; accept low power
Holm step-down	Family-wise error rate	You want Bonferroni protection with uniformly more power
Benjamini-Hochberg	False discovery rate	Many tests, exploratory; tolerate a known share of false positives
Pre-registration / single primary	The problem itself	You can name one confirmatory test in advance and demote the rest

WATCH OUT FOR

The cleanest defence against multiplicity is structural: pre-specify one primary outcome and a small number of confirmatory tests, and label everything else exploratory. Reporting only the comparisons that reached significance, with no account of how many were run, is the practice that p-curve analysis is designed to expose.

THE OUTPUT

Report significance the way examiners expect, and show it is not p-hacked

A clean report pairs the test, the exact p-value, an effect size, and an interval, then states the conclusion in words the number supports, and makes the analysis decisions auditable.

Reporting template (adapt to your test)

```
Test run:          [the specific test, e.g. independent t-test]
Result:            [statistic, df], p = [exact value to 3 decimals]
Effect size:       [e.g. Hedges' g, value, 95% CI]
Sample / power:    [n per group; a priori power for the target effect]
Adjustment:        [none / Holm / Benjamini-Hochberg, and family size]
In words:          [the groups differed on [outcome]; the effect was
                   [magnitude] and the interval [excluded / included]
                   a practically meaningful value]

Report p = .003, not p < .05, when you have the exact value.
Report p = .07 as not significant, never "trending" or "approaching".
```

PRINCIPLE

The confidence interval and the test are dual: the 95 per cent interval is the set of null values your data would not reject at 0.05, so reporting the interval gives the reader the whole family of tests at once. Where you make a directional claim across a set of studies or outcomes, a right-skewed p-curve is the evidence that your significant findings carry evidential value rather than reflecting selective reporting. Pre-registering the analysis plan is what makes all of this credible rather than asserted.

BEFORE YOU WRITE YOUR RESULTS

The p-value reporting checklist

Every box ticked means your significance claims say exactly what the statistics support, in a single coherent inferential logic, and no more.

- [ ] You can state the definition: probability of data this extreme under the null and the model.
- [ ] You never describe the p-value as the chance the null is true or the chance of a fluke.
- [ ] You chose one framework, Fisherian or Neyman-Pearson, and reported consistently with it.
- [ ] You report an effect size and confidence interval alongside every p-value.
- [ ] Any claim of no effect rests on an equivalence test or Bayes factor, not a large p-value.
- [ ] You set alpha and the primary outcome in advance, ideally pre-registered.
- [ ] You corrected for multiplicity with a method matched to your goal, and stated the family size.
- [ ] You avoid "approaching" or "trending toward" significance, and report exact p-values.

WHEN THE RESULTS HAVE TO HOLD UP

Have your analysis run and reported so it cannot be picked apart

If you want the right test chosen for each question, the framework kept coherent, multiplicity handled, and your p-values reported with effect sizes, intervals, and equivalence tests where they belong, our PhD statisticians do it with you. You stay the author and understand every number you present.

Get a quote for your analysis

dissertationstatisticshelp.org . Written by PhD statisticians . We work in writing, on your schedule.