The statistics section a reviewer cannot dismantle

Seven lessons produced eleven numbers. This one puts them into a report — what goes in the body, what goes in the annex, what a reviewer will ask for, and the four sentences that answer every question before it is asked.

Pierre Robentz Cassion

Lesson 8 of 8

Applied Statistics for Programmes — What the report claims

What this lesson covers

- What a reviewer actually does
- The four sentences
- Body, annex, and script
- The annex table
- The limitations paragraph, written from the course
- What to do when the number will not settle
- Report it whole
- Where this goes next

What a reviewer actually does

The reviewer asks	Answered by
"How sure are you of this number?"	The interval — lesson 2
"Is that difference real?"	The test and its assumptions — lesson 3
"Does it matter?"	The effect size in the decision's unit — lesson 4
"How many things did you look at?"	The test count and the correction — lesson 5

What a reviewer actually does

- The order is not the order you computed them in — The interval comes first because it is the only one a reader can. . .

The four sentences

- One: the number and its interval — "Referral completion among cases reporting a disability is 26.7% (95% CI 21.1 to...
- Two: the comparison and its test — "That is 19.4 points below cases with no disability reported (46.2%, n = 1,436; 95%...
- Three: the size, in the unit of the decision — "Applied to the 2024 caseload, the gap is 39 cases that would have...
- Four: what the analysis cannot say — "The comparison is observational

The four sentences — In Python

```
import pandas as pd
import numpy as np

protection = pd.read_csv("protection-referrals-2024.v1.csv")
consenting = protection[protection["consent_to_refer"]]
disability = consenting["disability_reported"].isin([True, "true", "Yes"])
reached = consenting["referral_accepted"] & consenting["days_to_first_service"].notna()

k, n = reached[disability].sum(), disability.sum()
comparator = reached[~disability].mean()
print(f"cases short of the comparator rate: {comparator * n - k:.0f}")
```

The four sentences — In R

```
# Sentence three is arithmetic, and it is the sentence a manager reads.
```

The four sentences

- Sentence three is the one analysts skip and managers need — Percentage points do not commission anything; thirty-nine...

Body, annex, and script

- The body carries the claim — The estimate, its interval, the comparison, the effect in programme units, and the…
- The annex carries the machinery — Test names, statistics, degrees of freedom, assumption checks, the count of…
- The script carries the reproduction — Every number in both, generated by code that runs from the committed file — which…

Body, annex, and script — In Python

```
def claim(label, k, n, digits=1):
    p = k / n
    z = 1.96
    denom = 1 + z**2 / n
    centre = (p + z**2 / (2 * n)) / denom
    half = z * np.sqrt(p * (1 - p) / n + z**2 / (4 * n**2)) / denom
    return (f"{label}: {p:.{digits}%} "
            f"(95% CI {centre - half:.{digits}%} to {centre + half:.{digits}%}, n={n})")

print(claim("Referral completion, disability reported", 54, 202))
```

Body, annex, and script — In R

```
# Generate the sentence, do not type it. A typed interval drifts from its number.
```

Body, annex, and script

- A number typed into a document is a number that will be wrong after the next data correction — Generating the sentence...

The annex table — In Python

```
annex = pd.DataFrame([
    {"comparison": "Completion by disability status", "test": "two-sample z",
     "statistic": "z = -5.21", "p": "<0.001", "n": "202 vs 1,436",
     "effect": "-19.4 pts (CI -26.1 to -12.8)"},
    {"comparison": "Over-age by sex", "test": "two-sample z",
     "statistic": "z = 1.84", "p": "0.066", "n": "510 vs 542",
     "effect": "+5.5 pts (CI -0.4 to +11.3)"},
    {"comparison": "Attendance by sex", "test": "two-sample z",
     "statistic": "z = 2.28", "p": "0.023", "n": "68,267 marks",
     "effect": "+0.56 pts, risk ratio 1.006"},
    {"comparison": "Attendance by school feeding", "test": "Welch t, school level",
     "statistic": "t = 3.11, df 12.7", "p": "0.008", "n": "15 vs 9 schools",
     "effect": "+5.2 pts (CI 1.6 to 8.8)"},
])
print(annex.to_string(index=False))
```

The annex table — In R

```
# One row per test run, including the ones that found nothing.
```

The annex table

Comparison	Test	Statistic	p	Effect
Completion by disability	Two-sample z	$z = -5.21$	<0.001	-19.4 pts
Over-age by sex	Two-sample z	$z = 1.84$	0.066	$+5.5$ pts
Attendance by sex	Two-sample z	$z = 2.28$	0.023	$+0.56$ pts, RR 1.006
Attendance by feeding	Welch t, school	$t = 3.11$, df 12.7	0.008	$+5.2$ pts
Over-age by sex, 24 schools	Two-sample z $\times 24$	2 of 24 at $p<0.05$	—	0 survive Bonferroni

The annex table

- Three of those five rows are in the annex and not in the body — because a non-significant gap, a significant…
- Listing the tests that found nothing is what makes the ones that found something credible — An annex containing only…

The limitations paragraph, written from the course

- Observational comparison — "Groups were not randomised; differences may reflect who enters each group rather than what...
- Clustering — "The programme is assigned by school, so the analysis uses 24 independent units
- Multiple comparisons — "Twenty-four school-level tests were run; 1.2 significant results are expected by chance and 2...
- Range restriction — "Attendance in this cohort runs from 86.7% to 97.8% in the middle half, so this analysis cannot...
- Each is one sentence and each pre-empts a specific question — A limitations paragraph that says "the data have...

What to do when the number will not settle

Result	Why it is unresolved	What to do
Over-age by sex, $p = 0.066$	Underpowered for a 5-point gap	Say so; pool the next round rather than closing the question
Attendance and literacy, $r = 0.04$	Range restriction	Report the null and name the restriction
Feeding and attendance, 24 clusters	Few clusters, observational	Report the interval; do not claim an effect

What to do when the number will not settle

- "The data cannot settle this" is a finding a programme can act on — because the action is to collect differently

Report it whole — Example

```
Statistical methods

  Proportions are reported with Wilson 95% confidence intervals. Group
  comparisons use two-sample z-tests for proportions and Welch's t-test for
  means. Where a programme is assigned above the level of the observation,
  the analysis is conducted at the level of assignment.

  Analyses were specified before the data were examined. Where a family of
  comparisons was run, the number of tests is stated with the result and a
  Bonferroni correction applied.

  All figures are generated from the committed dataset files by the scripts
  in this annex; no number in this report is typed by hand.

  Findings reported in the body: 2. Comparisons run in total: 29.
```

Report it whole

- The last line is the one that changes how the section is read — Two findings from twenty-nine comparisons is a. . .

Where this goes next

- But the discipline is complete here — Every number in a programme report can carry an interval, an effect size, a test...

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/applied-statistics-for-programmes/stat-08-the-statistics-in-the-report