Two gaps, two answers

A 5.5-point gap with an interval from -0.4 to $+11.3$, and a 19.4-point gap with an interval from -26.1 to -12.8. Same test, same code, opposite conclusions — and module 4 left both of them open on purpose.

Pierre Robentz Cassion

Lesson 3 of 8

Applied Statistics for Programmes — Comparing two groups

What this lesson covers

- The two questions module 4 deferred
- Gap one: the one that is not there
- Gap two: the one that is
- The two results side by side
- Choosing the test
- Check the assumptions, and say you did
- Report the difference, not the two numbers
- What comes next

The two questions module 4 deferred — In Python

```
import pandas as pd
import numpy as np
from statsmodels.stats.proportion import proportions_ztest, confint_proportions_2indep

enrolment = pd.read_csv("school-enrolment-2024.v1.csv")
clean = enrolment[(enrolment["age_years"] <= 20)
                  & (enrolment["school_year"] == 2024)
                  & enrolment["grade"].between(1, 6)]
clean = clean.assign(over_age=clean["age_years"] > clean["grade"] + 5)

boys = clean[clean["sex"] == "m"]["over_age"]
girls = clean[clean["sex"] == "f"]["over_age"]
print(f"boys  {boys.sum()}/{len(boys)} = {boys.mean():.1%}")
print(f"girls {girls.sum()}/{len(girls)} = {girls.mean():.1%}")
```

The two questions module 4 deferred — In R

```r
library(dplyr)

enrolment |>
  filter(age_years <= 20, school_year == 2024, between(grade, 1, 6)) |>
  mutate(over_age = age_years > grade + 5) |>
  summarise(k = sum(over_age), n = n(), .by = sex)
```

Gap one: the one that is not there

Group	Over-age	n	95% interval
Boys	39.2%	510	35.1–43.5%
Girls	33.8%	542	29.9–37.8%

Gap one: the one that is not there — In Python

```
count = np.array([boys.sum(), girls.sum()])
nobs = np.array([len(boys), len(girls)])

stat, pvalue = proportions_ztest(count, nobs)
low, high = confint_proportions_2indep(count[0], nobs[0], count[1], nobs[1])
print(f"difference {boys.mean() - girls.mean():+.1%}")
print(f"95% CI [{low:+.1%}, {high:+.1%}]")
print(f"z = {stat:.2f}, p = {pvalue:.3f}")
```

Gap one: the one that is not there — In R

```
prop.test(c(200, 183), c(510, 542))
```

Gap one: the one that is not there

- Difference +5.5 points, 95% CI −0.4 to +11.3, p = 0.066

Gap two: the one that is — In Python

```
protection = pd.read_csv("protection-referrals-2024.v1.csv")
consenting = protection[protection["consent_to_refer"]]
reached = (consenting["referral_accepted"]
           & consenting["days_to_first_service"].notna())

disability = consenting["disability_reported"].isin([True, "true", "Yes"])
k = np.array([reached[disability].sum(), reached[~disability].sum()])
n = np.array([disability.sum(), (~disability).sum()])

stat, pvalue = proportions_ztest(k, n)
low, high = confint_proportions_2indep(k[0], n[0], k[1], n[1])
print(f"{k[0]}/{n[0]} = {k[0]/n[0]:.1%}   {k[1]}/{n[1]} = {k[1]/n[1]:.1%}")
print(f"difference {k[0]/n[0] - k[1]/n[1]:+.1%}, CI [{low:+.1%}, {high:+.1%}]")
print(f"z = {stat:.2f}, p = {pvalue:.2e}")
```

Gap two: the one that is — In R

```
prop.test(c(54, 663), c(202, 1436))
```

Gap two: the one that is

Group	Completion	n
Disability reported	26.7%	202
Not reported	46.2%	1,436

Gap two: the one that is

- Difference -19.4 points, 95% CI -26.1 to -12.8, $p < 0.001$
- Every value in that interval is a substantial gap — The smallest gap consistent with the data is 12.8 points, which is. . .

The two results side by side — In Python

```
results = pd.DataFrame([
    {"comparison": "over-age, boys vs girls", "diff": 5.5,
     "ci": "-0.4 to +11.3", "n": "510 vs 542", "p": 0.066},
    {"comparison": "completion, disability", "diff": -19.4,
     "ci": "-26.1 to -12.8", "n": "202 vs 1436", "p": 0.0000002},
])
print(results)
```

The two results side by side — In R

```
# Two rows. The conclusion column is the analyst's, not the test's.
```

The two results side by side

- Notice that the significant result has the smaller sample in one arm — 202 cases produced a decisive answer and 510...

Choosing the test

Comparing	Test	In this course
Two proportions	Two-sample z-test, or `prop.test`	Both gaps above
Two means	Welch's t-test	Attendance by feeding, lesson 6
A categorical against a categorical	Chi-square	Closure reason by district

Choosing the test — In Python

```
from scipy import stats

points = pd.read_csv("water-point-monitoring-2024.v1.csv")
table = pd.crosstab(points["admin2"], points["functional_status"])
chi2, p, dof, expected = stats.chi2_contingency(table)
print(f"chi-square {chi2:.1f}, dof {dof}, p = {p:.4f}")
print(f"smallest expected cell: {expected.min():.1f}")
```

Choosing the test — In R

```
chisq.test(table(points$admin2, points$functional_status))
```

Check the assumptions, and say you did

- Independence — Every test above assumes each row is an independent observation
- Expected cell counts for chi-square — The test is unreliable when an expected cell falls below about five
- Equal variances for the t-test — Do not check it — use Welch's t-test always, which does not assume it

Check the assumptions, and say you did — In Python

```
girls_rate, boys_rate = girls.mean(), boys.mean()
print("assumptions checked:")
print(f"  independence: one row per student, no student in two rows -- verified")
print(f"  sample sizes: {len(boys)} and {len(girls)}, both well above 30")
print(f"  expected counts: smallest is {min(len(boys), len(girls)) * min(girls_rate, 1 -
    boys_rate):.0f}")
```

Check the assumptions, and say you did — In R

```
# Write the check into the script, not into your memory of having done it.
```

Report the difference, not the two numbers — Example (cont.)

```
Over-age enrolment by sex, primary, 2024

  Boys    39.2%   n = 510
  Girls   33.8%   n = 542
  Difference +5.5 points, 95% CI -0.4 to +11.3, p = 0.066

  The interval includes zero. This survey cannot establish whether over-age
  enrolment differs by sex; if it does, the gap is between 0 and 11 points
  and boys are the disadvantaged group. Not reported as a finding.

Referral completion by disability status, 1,638 consenting cases

  Disability reported     26.7%   n = 202
  Not reported            46.2%   n = 1,436
  Difference -19.4 points, 95% CI -26.1 to -12.8, p < 0.001
```

Report the difference, not the two numbers — Example (cont.)

```
The smallest gap consistent with the data is 12.8 points. Reported as a
finding, and the pathway lesson locates it at referral-making and
acceptance rather than at consent.
```

Report the difference, not the two numbers

- Both blocks report the difference and its interval as the headline — with the two group figures beneath

What comes next

- One of these gaps is statistically significant.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/applied-statistics-for-programmes/stat-03-two-gaps-two-answers