A proportion of zero with an interval up to 24%

Twelve children, none of them over-age. The textbook interval says 0% to 0%. The right one says 0% to 24.3%, and the difference between those two answers is whether you would act on the cell.

Pierre Robentz Cassion

Lesson 2 of 8

Applied Statistics for Programmes — Look at it first

What this lesson covers

- Why a point estimate is not a result
- Wilson, not Wald
- Read the width, not just the bounds
- Writing it into a sentence
- When the interval changes the decision
- Three intervals this course will not compute this way
- Report it as a range
- What comes next

Why a point estimate is not a result — In Python

```
import pandas as pd
import numpy as np

wash = pd.read_csv("wash-household-survey-2024.v1.csv")
open_defecation = wash["sanitation_facility"].eq("open-defecation")
k, n = open_defecation.sum(), len(wash)
print(f"{k}/{n} = {k / n:.1%}")
```

Why a point estimate is not a result — In R

```
library(dplyr)
wash |> summarise(k = sum(sanitation_facility == "open-defecation"), n = n())
```

Why a point estimate is not a result

- 319 of 2,403 — 13.3% — The interval on that is narrow, because the sample is large

Wilson, not Wald — In Python

```
def wald(k, n, z=1.96):
    p = k / n
    se = np.sqrt(p * (1 - p) / n)
    return p - z * se, p + z * se

def wilson(k, n, z=1.96):
    p = k / n
    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 centre - half, centre + half

for k, n in [(0, 12), (1, 20), (3, 11)]:
    print(f"{k}/{n} = {k/n:.1%}")
    print(f"   Wald   [{wald(k, n)[0]:.1%}, {wald(k, n)[1]:.1%}]")
    print(f"   Wilson [{wilson(k, n)[0]:.1%}, {wilson(k, n)[1]:.1%}]")
```

Wilson, not Wald — In R

```r
# R has this built in and it is the Wilson interval by default.
prop.test(0, 12)$conf.int
binom.test(1, 20)$conf.int
```

Wilson, not Wald

Cell	Wald	Wilson
0 of 12	**[0.0%, 0.0%]**	[0.0%, 24.3%]
1 of 20	**[−4.6%, 14.6%]**	[0.9%, 23.6%]
3 of 11	[1.0%, 53.6%]	[9.7%, 56.6%]

Wilson, not Wald

- Wald says a cell with no events has zero uncertainty, and it produces negative probabilities — Both are absurd and both...
- Use Wilson — or `prop.test` in R and `statsmodels.stats.proportion.proportion_confint(method="wilson")` in Python

Read the width, not just the bounds — In Python

```
cases = [
    ("Open defecation", 319, 2403),
    ("Cholera case fatality", 38, 974),
    ("Cholera CFR, Nord district", 24, 381),
    ("Referral completion, disability reported", 54, 202),
    ("Over-age, grade 6", 67, 132),
]
for name, k, n in cases:
    lo, hi = wilson(k, n)
    print(f"{name:42} {k/n:6.1%}  [{lo:.1%}, {hi:.1%}]  width {hi-lo:.1%}")
```

Read the width, not just the bounds — In R

```
# Same five, and the width is the column to read.
```

Read the width, not just the bounds

Estimate	Value	95% interval	Width
Open defecation	13.3%	12.0–14.7%	**2.7 pts**
Cholera case fatality	3.9%	2.9–5.3%	2.5 pts
Cholera CFR, Nord	6.3%	4.3–9.2%	4.9 pts
Referral completion, disability	26.7%	21.1–33.2%	12.1 pts
Over-age, grade 6	**50.8%**	**42.3–59.1%**	**16.8 pts**

Read the width, not just the bounds

- Grade 6 over-age is 50.8% and could be anywhere from 42% to 59% — The education course reported that number to one…
- Two things drive the width and only one is under your control — Sample size, which the design decided; and how close…

Writing it into a sentence

- Not this — "50.8% of grade 6 students are over-age."
- Nor this — "50.8% (95% CI 42.3–59.1) of grade 6 students are over-age." Correct, and a reader skips the parenthesis
- This — "Between four and six students in ten in grade 6 are over-age (50.8%, 95% CI 42.3–59.1, n=132)."
- Lead with the interval in words, then give the numbers — The words are what a non-analyst reads and they carry the...

Writing it into a sentence — In Python

```
def sentence(label, k, n):
    lo, hi = wilson(k, n)
    return (f"{label}: {k/n:.1%} (95% CI {lo:.1%} to {hi:.1%}, n={n})")

print(sentence("Grade 6 over-age", 67, 132))
```

Writing it into a sentence — In R

```
# Generate the string in code so the number and its interval cannot drift apart.
```

When the interval changes the decision — In Python

```
lo, hi = wilson(38, 974)
threshold = 0.01                        # Sphere: cholera CFR below 1%
print(f"CFR {38/974:.1%}, interval [{lo:.1%}, {hi:.1%}]")
print(f"entire interval above the {threshold:.0%} threshold: {lo > threshold}")
```

When the interval changes the decision — In R

```
prop.test(38, 974)$conf.int
```

When the interval changes the decision

- Cholera case fatality is 3.9% with an interval of 2.9% to 5.3%, and the Sphere threshold is 1% — Every value in the...
- That is when you can act on a point estimate: when the whole interval says the same thing — Where it straddles the...

Three intervals this course will not compute this way

- A proportion from a clustered sample — The survey course established that a cluster design widens the interval by the...
- A median or a skewed mean — The intervals here are for proportions
- A count with no denominator — "142 cases were reported" has no interval, because it is not an estimate of anything

Report it as a range — Example

Over-age enrolment, grade 6

50.8% 95% CI 42.3 to 59.1, n = 132

Between four and six students in ten. The interval is wide because grade 6 holds 132 students; a difference of five points against another grade would not be distinguishable at this sample size.

Report it as a range

- The last sentence is the useful one — It tells the reader in advance which comparisons this number can support, which...

What comes next

- An interval tells you how uncertain one number is.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/applied-statistics-for-programmes/stat-02-the-interval