What the last digit tells you

Digit preference, heaping and the calibration that separates a finding from noise. One team's heights end in .0 or .5 sixty-nine percent of the time; the district's most suspicious facility turns out to be chance.

Pierre Robentz Cassion

Lesson 6 of 8

Data Quality Assessment — Finding the sites worth visiting

What this lesson covers

- Measurements have a texture
- The check
- Age heaping, and why it is usually nobody's fault
- Calibrate against what chance produces
- Why the null case matters more
- What you may write
- What comes next

Measurements have a texture

- A number produced by reading an instrument has a particular texture: the last digit is close to uniform, because the true value is equally likely to fall anywhere within the smallest division of the scale.

The check — In Python

```
import pandas as pd

smart = pd.read_csv("smart-nutrition-survey-2024.v1.csv")

smart["terminal"] = (smart["height_cm"] * 10).round() % 10
by_team = (
    smart.assign(rounded=smart["terminal"].isin([0, 5]))
    .groupby("team")
    .agg(rounded_share=("rounded", "mean"), n=("rounded", "size"))
)
print((by_team["rounded_share"] * 100).round(0))
```

The check — In R

```r
library(dplyr)

smart |>
  mutate(terminal = round(height_cm * 10) %% 10,
         rounded = terminal %in% c(0, 5)) |>
  summarise(rounded_share = mean(rounded), n = n(), .by = team)
```

The check

Team	Heights ending .0 or .5	Children measured
1	18%	217
2	**69%**	**248**
3	18%	248
4	23%	217

The check

- That is not a marginal result and it does not need a test — Nothing about a population makes one team's children's...

Age heaping, and why it is usually nobody's fault — In Python

```
ages = smart["age_months"].dropna()
whole_years = (ages % 12 == 0).mean()

print(f"{whole_years:.0%} of ages fall on an exact multiple of 12 months")
print(ages.value_counts().reindex([22, 23, 24, 25, 26]))
```

Age heaping, and why it is usually nobody's fault — In R

```
smart |>
  filter(!is.na(age_months)) |>
  summarise(whole_years = mean(age_months %% 12 == 0))

smart |> count(age_months) |> filter(age_months %in% 22:26)
```

Age heaping, and why it is usually nobody's fault

Age in months	**22**	**23**	**24**	**25**	**26**
Children	21	15	**66**	19	17

Calibrate against what chance produces — In Python

```
reported = vax[vax["reported"]].copy()
base_rate = (reported["doses_administered"] % 10 == 0).mean()

by_facility = (
    reported.assign(round_number=reported["doses_administered"] % 10 == 0)
    .groupby("facility_id")
    .agg(rounds=("round_number", "sum"), n=("round_number", "size"))
)
by_facility["share"] = by_facility["rounds"] / by_facility["n"]
print(f"district base rate: {base_rate:.3f}")
print(by_facility.nlargest(3, "share"))
```

Calibrate against what chance produces — In R

```r
base_rate <- mean(vax$doses_administered[vax$report_submitted] %% 10 == 0)

by_facility <- vax |>
  filter(report_submitted) |>
  summarise(rounds = sum(doses_administered %% 10 == 0), n = n(), .by = facility_id) |>
  mutate(share = rounds / n) |>
  arrange(desc(share))
```

Calibrate against what chance produces — In Python

```
from scipy.stats import binomtest

worst = by_facility.nlargest(1, "share").iloc[0]
p = binomtest(int(worst["rounds"]), int(worst["n"]), base_rate,
              alternative="greater").pvalue
print(f"p = {p:.4f}, Bonferroni across 38 facilities: {min(1, p * 38):.3f}")
```

Calibrate against what chance produces — In R

```
worst <- by_facility[1, ]
p <- binom.test(worst$rounds, worst$n, base_rate, alternative = "greater")$p.value
c(p = p, bonferroni = min(1, p * 38))
```

Calibrate against what chance produces

- $p = 0.003$ on its own, and 0.11 after correcting for having looked at thirty-eight facilities — Three facilities fall. . .

Calibrate against what chance produces

The result is: **no evidence of digit preference in the vaccination reporting.** That is a real finding, it took four lines, and it is the one you should hope for.

Why the null case matters more

- **Compute the base rate from the data**, not from theory. Terminal digits are not uniform when counts are small, when a…
- **Correct for the number of units you looked at.** Testing thirty-eight facilities at 5% produces about two positives…
- **Require a mechanism.** Team 2's 69% has one — the team read the measuring board to the nearest half centimetre. If…

What you may write

Do not write	Write
"FAC020's data appear fabricated"	"FAC020 reports round numbers more often than the district average; the difference is not significant after adjusting for the number of facilities examined"
"Team 2's measurements are unreliable"	"Team 2's height readings end in .0 or .5 in 69% of cases against 18–23% for other teams, consistent with reading the board to the nearest half centimetre"
"Ages are inaccurate"	"24% of ages fall on an exact whole year against 9% expected, indicating age was estimated rather than documented for a substantial share of children"

What comes next

- You now have findings — a completeness collapse, a verification factor of 1.42, a team rounding its heights.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/data-quality-assessment/dqa-06-digit-preference