A mean of 29.6 and a median of zero

The mean E. coli count in tested households is 29.6 CFU/100 mL. The median is 0. More than half the sample has no detectable contamination, and the mean is describing a tail that reaches 608.

Pierre Robentz Cassion

Lesson 1 of 8

Applied Statistics for Programmes — Look at it first

What this lesson covers

- Compute the summary last
- Look at the shape
- Three shapes and what each demands
- When the skew is the data error
- What to report for each shape
- The histogram you should always draw
- What comes next

Compute the summary last — In Python

```
import pandas as pd

wash = pd.read_csv("wash-household-survey-2024.v1.csv")
ecoli = wash["ecoli_cfu_100ml"].dropna()

print(ecoli.describe().round(1))
```

Compute the summary last — In R

```
library(dplyr)

wash |> filter(!is.na(ecoli_cfu_100ml)) |>
  summarise(n = n(), mean = mean(ecoli_cfu_100ml),
            median = median(ecoli_cfu_100ml), sd = sd(ecoli_cfu_100ml))
```

Compute the summary last

	Value
n	802
Mean	**29.6**
Median	**0.0**
Standard deviation	85.6
Maximum	608

Compute the summary last

- The mean is 29.6 and the median is 0 — Those two numbers describe the same 802 households, and only one of them...

Look at the shape — In Python

```
bins = [-1, 0, 10, 100, 1000]
labels = ["0 (none detected)", "1-10", "11-100", ">100"]
print(pd.cut(ecoli, bins, labels=labels).value_counts().reindex(labels))
print(f"\nskew: {ecoli.skew():.2f}")
```

Look at the shape — In R

```
wash |> filter(!is.na(ecoli_cfu_100ml)) |>
  count(band = cut(ecoli_cfu_100ml, c(-1, 0, 10, 100, Inf)))
```

Look at the shape

Band	Households	Share
0 — none detected	430	**53.6%**
1–10	173	21.6%
11–100	139	17.3%
>100	60	7.5%

Look at the shape

- Skew +4.04 — A symmetric distribution has a skew near zero; anything past about +1 means the mean and the median are...
- This is why the WASH course reported risk classes rather than a mean — The classes are not a presentational choice —...

Three shapes and what each demands — In Python (cont.)

```
import numpy as np

def profile(series, name):
    s = series.dropna()
    return {
        "indicator": name, "n": len(s),
        "mean": round(s.mean(), 1), "median": round(s.median(), 1),
        "skew": round(s.skew(), 2),
    }

points = pd.read_csv("water-point-monitoring-2024.v1.csv")
protection = pd.read_csv("protection-referrals-2024.v1.csv")

print(pd.DataFrame([
    profile(wash["ecoli_cfu_100ml"], "E. coli, CFU/100mL"),
    profile(points["days_since_breakdown"], "days a water point is down"),
```

Three shapes and what each demands — In Python (cont.)

```
    profile(protection["days_to_first_service"], "days to first service"),
    profile(wash["litres_per_person_day"], "litres per person per day"),
]))
```

Three shapes and what each demands — In R

```
# One function, four indicators, four different answers about which summary
# to report.
```

Three shapes and what each demands

Indicator	n	Mean	Median	Skew
E. coli, CFU/100 mL	802	29.6	**0.0**	+4.04
Days a water point is down	709	125.7	56.0	+1.64
Days to first service	728	10.1	9.0	+0.52
Litres per person per day	2,403	24.2	23.7	**+8.54**

Three shapes and what each demands

- Days to first service is nearly symmetric — mean 10.1, median 9.0, skew +0.52
- Days a water point is down is heavily right-skewed — a mean of 125.7 against a median of 56, because a handful of...
- Litres per person per day looks almost symmetric and has a skew of +8.54 — That combination is the interesting one

When the skew is the data error — In Python

```
litres = wash["litres_per_person_day"].dropna()
print(f"median {litres.median():.1f}, p90 {litres.quantile(0.90):.1f}, "
      f"max {litres.max():.1f}")
print(f"above 80: {(litres > 80).sum()} households")
```

When the skew is the data error — In R

```
wash |> summarise(median = median(litres_per_person_day),
                  p90 = quantile(litres_per_person_day, 0.9),
                  max = max(litres_per_person_day))
```

When the skew is the data error — In Python

```
clean = litres[litres <= 80]
print(f"after removing 11 rows: skew {clean.skew():.2f}, "
      f"mean {clean.mean():.1f}, median {clean.median():.1f}")
```

When the skew is the data error — In R

```
# Recompute after the correction and see whether the shape was real.
```

When the skew is the data error

- Eleven rows out of 2,403 and the skew goes from +8.54 to +0.06 — Mean 23.5, median 23.6 — a distribution that was...
- Compute the skew, then decide whether it is a finding or a bug — Both are common and they are told apart by looking at...

What to report for each shape

Shape	Report	Do not report
Roughly symmetric	Mean and standard deviation	—
Right-skewed	Median and interquartile range	A mean without the median beside it
Zero-inflated	**The share at zero**, then the distribution of the rest	A mean at all
Bounded proportion	The proportion and its interval	A standard deviation

What to report for each shape

- E. coli is zero-inflated — which is a distinct case from merely skewed: 53.6% of the sample sits at exactly one value,. . .

What to report for each shape — Example

E. coli at point of collection, 802 households tested

No detectable E. coli	53.6%	430 households
Among the 372 with any detected:		
median	13 CFU/100 mL	
interquartile range	5 to 49	
maximum	608	

Mean over all 802 households is 29.6 CFU/100 mL. It is not reported as a summary because 53.6% of the sample sits at zero and the mean falls in a range occupied by almost no household.

The histogram you should always draw — In Python

```
import matplotlib.pyplot as plt

fig, axes = plt.subplots(1, 2, figsize=(9, 3))
axes[0].hist(ecoli, bins=40)
axes[0].set_title("E. coli, raw")
axes[1].hist(np.log1p(ecoli), bins=40)
axes[1].set_title("log(1 + E. coli)")
plt.tight_layout()
```

The histogram you should always draw — In R

```
hist(wash$ecoli_cfu_100ml, breaks = 40)
hist(log1p(wash$ecoli_cfu_100ml), breaks = 40)
```

The histogram you should always draw

- Draw it before you summarise it, every time, and do not put it in the report — The histogram is an instrument for you,...
- Two minutes of looking would have prevented every error in this lesson — and that is the entire argument for the habit

What comes next

- You now know which single number describes an indicator.

Read the full lesson, with runnable code

https://data-analysis.cassion.dev/courses/applied-statistics-for-programmes/stat-01-the-mean-that-describes-nobody