Module 4: Descriptive statistics and epidemiological measures

Data analysis with Python for health specialists

Yaé Ulrich Gaba

2026

AIRINA Labs

Mean vs. median: why it matters

```
import pandas as pd
glucose = pd.Series([92, 98, 104, 95, 88, 110, 97, 340, 101, 93])

print(f"Mean:   {glucose.mean():.1f} mg/dL")   # 121.8
print(f"Median: {glucose.median():.1f} mg/dL")  # 97.5
```

One patient with DKA (glucose = 340) pulls the mean up by 25 mg/dL. The median is unaffected.

Rule of thumb
Use **median** for skewed data (lab values, costs, length of stay). Use **mean** for symmetric data (height, diastolic BP in healthy adults).

Spread: SD, IQR, and coefficient of variation

```
# Two clinics with similar means but different variability
clinic_a = pd.Series([120, 122, 118, 125, 121, 119, 123, 120])
clinic_b = pd.Series([98, 145, 110, 155, 102, 160, 115, 135])

print(f"Clinic A: mean={clinic_a.mean():.1f}, "
      f"SD={clinic_a.std():.1f}")
print(f"Clinic B: mean={clinic_b.mean():.1f}, "
      f"SD={clinic_b.std():.1f}")
```

Both means $\approx$121, but Clinic B has patients with dangerously high and unusually low blood pressure.

IQR = Q75 − Q25. Robust to outliers, like the median.

CV = SD/mean $\times$ 100%. Compares variability across different scales.

Frequency tables and cross-tabulations

```
# One-way frequency table
freq = recent["le_cat"].value_counts().sort_index()

# Two-way cross-tabulation (the "Table 1" of papers)
ct = pd.crosstab(recent["continent"], recent["le_cat"],
                 margins=True)

# With row percentages
ct_pct = pd.crosstab(recent["continent"], recent["le_cat"],
                     normalize="index").round(3) * 100
```

Clinical context

Every “Table 1” in a clinical paper is essentially a cross-tabulation. In pandas, `pd.crosstab()` builds them in one line.

Prevalence and incidence

Prevalence (burden):

$$\frac{\text{Existing cases}}{\text{Total population}} \times 100$$

```
prev = 625 / 5000 * 100
# 12.5%
```

Incidence rate (risk):

$$\frac{\text{New cases}}{\text{Person-time at risk}}$$

```
ir = 18 / 850  # per
↪  person-year
# = 21.2 per 1,000 PY
```

Key difference

Low incidence + high prevalence = patients survive a long time (diabetes). High incidence + low prevalence = disease kills quickly (Ebola).

Risk ratio (relative risk)

Compares probability of disease between exposed and unexposed:

$$\text{RR} = \frac{a/(a+b)}{c/(c+d)}$$

```
# Smoking and lung cancer
a, b, c, d = 80, 920, 10, 990
risk_exposed = a / (a + b)      # 8.0%
risk_unexposed = c / (c + d)    # 1.0%
rr = risk_exposed / risk_unexposed  # 7.92

# 95% CI via log(RR)
import math
se = math.sqrt(1/a - 1/(a+b) + 1/c - 1/(c+d))
ci = (math.exp(math.log(rr) - 1.96*se),
      math.exp(math.log(rr) + 1.96*se))
```

Smokers have 8× the risk of lung cancer compared to non-smokers.

Odds ratio

Used in **case-control** studies where you cannot compute risk directly:

$$\text{OR} = \frac{a \times d}{b \times c}$$

```
# Obesity and knee osteoarthritis
a, b, c, d = 120, 60, 80, 140
odds_ratio = (a * d) / (b * c)  # 3.50

se = math.sqrt(1/a + 1/b + 1/c + 1/d)
ci = (math.exp(math.log(odds_ratio) - 1.96*se),
      math.exp(math.log(odds_ratio) + 1.96*se))
```

When OR $\approx$ RR
The OR approximates the RR when the disease is **rare** (<10% prevalence). For common outcomes, OR exaggerates the association.

Age-standardized rates

Crude rates mislead when populations have different age structures.

```
# Direct standardization
std_weight = [0.26, 0.42, 0.22, 0.10]  # WHO World Std

# Age-specific rates x standard weights
asr_a = (data_a["rate"] * std_weight).sum()
asr_b = (data_b["rate"] * std_weight).sum()
```

Always report your standard
Different standards (WHO, European, US 2000) give different results. You must use the **same** standard to compare with published rates.

What you can do after this module

1. Choose the right measure of center (mean vs. median) for health data
2. Compute prevalence and incidence with confidence intervals
3. Build 2×2 tables and calculate risk ratios and odds ratios
4. Construct cross-tabulations for “Table 1” style reporting
5. Perform age standardization to fairly compare populations

Next: Module 5 — Visualization for health data

Exercises (Hour 3)

1. **Summary stats:** Compute mean, median, SD, IQR for life expectancy by continent (2007)
2. **Odds ratio:** From a 2×2 table (200 cases, 200 controls), compute OR with 95% CI
3. **Age standardization:** Compare crude vs. standardized mortality for two populations
4. **Mini-project:** Build an epidemiological profile using Gapminder + WHO data