Applied Statistics for Public Health Professionals

Dr. Inayat Ullah

Assistant Professor
Dept. of Govt & Public Policy, NUST Islamabad

September 19, 2025

Rates: Incidence and Prevalence

Why Do We Need Rates?

- Incidence and prevalence are core measures in public health research.
- They provide clear definitions of disease frequency in populations.
- Simply reporting the number of events (e.g., deaths, hospital admissions) can be misleading.
- Rates allow us to make fair comparisons across time, groups, and populations.

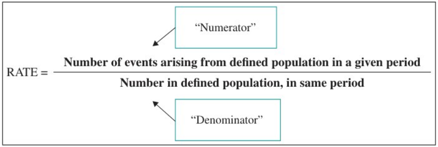
Self-Assessment Exercise 1.3.1

Case: Holy Family Hospital Rawalpindi reported a 30% rise in acute medical admissions for people over 65 in one year, compared to a steady 5% annual increase in the past 5 years.

Questions:

1. What are the possible reasons for the sudden 30% increase?
2. What additional information would help interpret this change?

In these questions we can see the importance of interpreting changes in numbers of events in the light of knowledge about the population from which those events arose. A **Rate** has a **numerator** and a **denominator** and must be determined over a specified period of time. It can be defined as follows:



"Numerator"

$$\text{RATE} = \frac{\textbf{Number of events arising from defined population in a given period}}{\textbf{Number in defined population, in same period}}$$

"Denominator"

Prevalence Rate

Definition:

- Number of existing cases in a population at a given time.
- Expressed as % or per 1,000 / 10,000 population.

$$\text{Prevalence} = \frac{\text{Number of cases at a given time}}{\text{Population at that time}}$$

Examples:

- 100,000 smokers in 400,000 population ⇒ 25% (250 per 1,000).
- 5,000 schizophrenia cases in 400,000 population ⇒ 1.25% (12.5 per 1,000).

Types:

- *Point Prevalence:* Snapshot at one time point.
- *Period Prevalence:* Cases existing or arising over a specified time period.

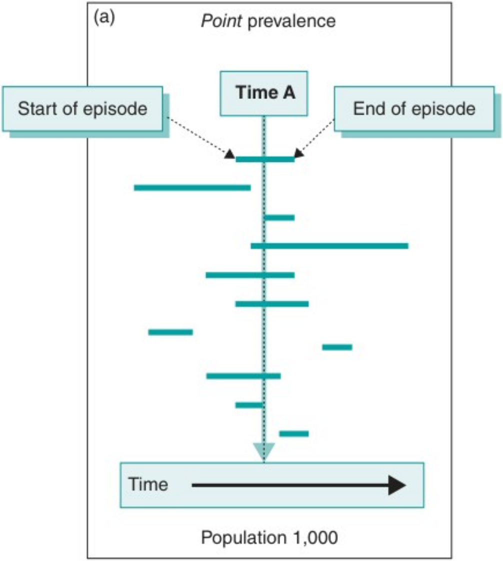


Figure: Period and point prevalence

Exercise 1.3.2: Point vs. Period Prevalence

1. **Calculation**

- *Point prevalence:* 7 active cases at time A in a population of 1,000

$$\frac{7}{1,000} = 0.007 = 0.7\% = 7 \text{ per 1,000}$$

- *Period prevalence:* 10 active cases during period B in the same population

$$\frac{10}{1,000} = 0.01 = 1\% = 10 \text{ per 1,000}$$

2. **Why Different?**

- Period prevalence counts both:
  - Cases that had not yet recovered.
  - New cases appearing during the period.
- Point prevalence is a snapshot at one moment in time.

Incidence Rate

Definition:

- Measures the rate of **new cases** appearing in a population over a specified time period.
- Time period must always be stated.

$$\text{Incidence Rate} = \frac{\text{New cases in a defined period}}{\text{At-risk population during the same period}}$$

Example:

- 20 new cases in an at-risk population of 2,500 over 1 year:

$$\frac{20}{2,500} \times 1,000 = 8 \text{ per 1,000 per year}$$

Key Point:

- Denominator must include only those at risk of becoming a case.
- e.g., For hysterectomy rates, exclude women who have already had the procedure.

Distribution of Data

What is a distribution?

- The distribution describes how values of a variable are spread across possible values.
- Common shapes: *symmetric (normal)*, *right-skewed*, *left-skewed*,
- Visual tools: **histogram**, **boxplot**, **density plot**.

Why it matters for public health:

- Choice of summary statistics and tests depends on distribution (e.g. mean/SD).
- Skewed distributions are common for biomarkers, costs, and event counts.

Frequency & Relative Frequency

Definitions:

- **Frequency (count)**: how many observations fall in each category/bin.
- **Relative frequency**: frequency divided by total ⇒ proportion or percentage.

Example: data $= \{1, 2, 2, 3, 4, 4, 4, 5\}$ (total $n = 8$)

Value	Frequency	Relative frequency
1	1	$1/8 = 0.125$ (12.5%)
2	2	$2/8 = 0.25$ (25%)
3	1	$1/8 = 0.125$ (12.5%)
4	3	$3/8 = 0.375$ (37.5%)
5	1	$1/8 = 0.125$ (12.5%)

Central Tendency: Mean, Median, Mode

Mean (average):

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Example: data $= \{1, 2, 2, 3, 4, 4, 4, 5\}$.
$\bar{x} = \frac{1+2+2+3+4+4+4+5}{8} = \frac{25}{8} = 3.125.$

Median: the middle value (or average of two middle values) when data are ordered.
Example (ordered): $1, 2, 2, 3, 4, 4, 4, 5$ — median = average of 4th and 5th values $= (3+4)/2 = 3.5$.

Mode: most frequent value. Example: mode = 4 (appears 3 times).

Dispersion: Range, Variance, SD, IQR

Range: max − min.
For $\{1, 2, 2, 3, 4, 4, 4, 5\}$: range $= 5 - 1 = 4$.

Variance (sample):

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

Standard deviation: $s = \sqrt{s^2}$.

IQR (interquartile range): $Q_3 - Q_1$ (middle 50% spread). Use with median for skewed data.

Short example (using previous data):

- $\bar{x} = 3.125$. Compute squared deviations, sum, divide by $n - 1 = 7$ to get s^2 (exercise for students), then $s = \sqrt{s^2}$.
- Q_1 and Q_3 can be read from ordered data (here roughly $Q_1 = 2$, $Q_3 = 4$ so IQR ≈ 2).

Covariance & Correlation: Concepts and Formulas

Covariance (sample):

$$\mathrm{Cov}(X, Y) = s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Interpretation: sign indicates direction (positive → variables move together); magnitude depends on units.

Pearson correlation (sample):

$$r = \frac{s_{XY}}{s_X s_Y}$$

where s_X and s_Y are sample standard deviations.
$r \in [-1, 1]$: magnitude shows strength, sign shows direction.

Why both? Covariance has units of $X \times Y$; correlation is unitless and comparable across datasets.

Worked example: Covariance & Correlation

Data (paired):

$$X = \{2, 4, 6, 8\}, \qquad Y = \{1, 4, 5, 9\}$$

Step 1: means

$$\bar{X} = 5.00, \qquad \bar{Y} = 4.75$$

Step 2: deviations and products

i	$x_i - \bar{X}$	$y_i - \bar{Y}$	$(x_i - \bar{X})(y_i - \bar{Y})$
1	-3.00	-3.75	11.25
2	-1.00	-0.75	0.75
3	1.00	0.25	0.25
4	3.00	4.25	12.75
	sum of products		25.00

Step 3: sample covariance

$$s_{XY} = \frac{25}{n-1} = \frac{25}{3} = 8.333\ldots$$

Cont

Step 4: sample SDs

$$s_X = \sqrt{\frac{\sum(x_i - \bar{X})^2}{n-1}} \approx 2.582, \quad s_Y \approx 3.304$$

Step 5: Pearson correlation

$$r = \frac{s_{XY}}{s_X s_Y} \approx \frac{8.3333}{2.58199 \times 3.30404} \approx 0.977$$

Interpretation: Strong positive linear association ($r \approx 0.98$).