Applied Statistics for Public Health Professionals

Dr. Inayat Ullah

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

October 2, 2025

Key Terms

- **Population:** The total set of items we are concerned about.
  Example: All residents of Islamabad Capital Territory.
- **Parameter:** A measure summarizing a population.
  Example: The mean education level in District Attock is 7 years (if every individual is counted).
- **Sample:** A subset of a population.
  In this course, we assume all samples are selected randomly.
- **Random Sample:** Every member of the population has an equal chance of being included.
  If the sample is not random, statistical inference may not hold.
- **Statistic:** A measure summarizing a sample.
  Example: The mean, standard deviation, and median of a sample.

Frequency & Relative Frequency

Definitions:

- **Frequency (count)**: how many observations fall in each category/bin.
- **Relative frequency**: frequency divided by total $\Rightarrow$ 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, 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 and Correlation

Covariance:

$$\text{Cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Correlation:

$$r_{XY} = \frac{\text{Cov}(X, Y)}{s_X s_Y}, \quad -1 \leq r_{XY} \leq 1$$

Policy Example: Explore correlation between literacy rates (Y) and per-capita income (X) across districts.

Stata Example:

```
use PSLM.dta, clear
bys district_name : correlate literacy monthly_income
```

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 ($r \approx 0.98$).
Another Example from PSLM Dataset

```
twoway (scatter monthly_income highest_education_level ///
if highest_education_level < 20 & monthly_income < 500000) ///
 (lfit monthly_income highest_education_level ///
if highest_education_level < 20 & monthly_income < 500000)
```

Policy Example : Variance and Standard Deviation

Compare the variation in literacy rate rural vs. urban.

Stata Example:

```
use PSLM.dta, clear
summarize literacy

by region: summarize literacy
```

Applications of Standard Deviation in Real Life

Why Standard Deviation Matters:
In practice, a mean by itself has limited value because it hides the volatility or variability in the data. Public or nonprofit managers should be cautious in drawing conclusions if only the mean is reported.

Example: A supervisor is told that employees scored an average of 90% on a job skills test. At first, the supervisor is very pleased. But without knowing the *standard deviation*, the interpretation can be misleading.

$$\mu = 90,\ \sigma = 2 \quad \Rightarrow \quad \text{Most employees clustered near 90\%}$$

$$\mu = 90,\ \sigma = 9 \quad \Rightarrow \quad \text{Wide spread; some far below passing level}$$

Policy Insight: For decision-making, reporting both the mean and standard deviation provides a fuller picture of performance.

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 in public policy analysis:

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

Boxplot and Kernel Density Plot

Why Kernel Density Plots? Example 1: Box-and-Whisker Plot

Compare monthly household income across urban and rural households:

```
graph box monthly_income, over(region) ///
    title("Monthly Income: Urban vs Rural") ///
    ytitle("Income (PKR)")
```

Example 2: Kernel Density Plot

Compare the shape of the income distribution:

```
kdensity monthly_income if monthly_income < 200000 & region=
    lcolor(blue) title("Income Distribution: Urban vs Rural'
    addplot(kdensity monthly_income if monthly_income < 200(
    legend(label(2 "Urban") label(1 "Rural"))
```

Frequency Distributions and Histograms

- Frequency distribution: table of counts/percentages.
- Histogram: visualize continuous data distribution.
- Bar chart: categorical comparisons.

Policy Example: Distribution of household income to inspect inequality and detect outliers.

Stata Example:

```
histogram monthly_income , width(5000) frequency
graph bar literacy, over(district_name, label(angle(vertical))
tabulate district_name , summarize( monthly_income )
```

Example of Why Sample Mean is a best estimator of Population Mean

Number of Arrests by Police Officers in Yukon, Oklahoma, 2014

Police Officer	Number of Arrests, 2014
1	14
2	16
3	10
4	18
5	8
6	15
7	17
8	20
9	19
10	13

Sampling Distribution

Calculating the Average Sample Mean from Samples of Five

Officers in Sample	Number of Arrests	Sample Mean	Average of Means
1, 3, 2, 8, 4	14, 10, 16, 20, 18	15.6	15.6
1, 6, 8, 3, 4	14, 15, 20, 10, 18	15.4	15.5
7, 4, 1, 5, 9	17, 18, 14, 8, 19	15.2	15.4
2, 10, 7, 4, 6	16, 13, 17, 18, 15	15.8	15.5
7, 10, 3, 6, 5	17, 13, 10, 15, 8	12.6	14.9
10, 7, 2, 1, 4	13, 17, 16, 14, 18	15.6	15.0
10, 3, 7, 4, 2	13, 10, 17, 18, 16	14.8	15.0
10, 3, 8, 9, 4	13, 10, 20, 19, 18	16.0	15.1
6, 4, 8, 9, 10	15, 18, 20, 19, 13	17.0	15.3
2, 8, 4, 1, 5	16, 20, 18, 14, 8	15.2	15.3
8, 3, 9, 10, 5	20, 10, 19, 13, 8	14.0	15.2
2, 3, 7, 5, 1	16, 10, 17, 8, 14	13.0	15.0
10, 8, 5, 6, 4	13, 20, 8, 15, 18	14.8	15.0
9, 3, 6, 2, 7	19, 10, 15, 16, 17	15.4	15.0

Estimating a Population Standard Deviation

Calculating s for a Sample of Five

Officer	Arrests	Arrests − Mean	Squared
1	14	−1.6	2.56
3	10	−5.6	31.36
2	16	.4	.16
8	20	4.4	19.36
4	18	2.4	5.76
	$s = \sqrt{59.2 \div 4} = 3.85$		59.20 sum of squares

Our estimate of the population standard deviation is 3.85, which is close to the population value of 3.7. Note that had we divided by n rather than by $n - 1$, the standard deviation estimate would have been 3.4. Dividing by n gives us a consistently low estimate of the population standard deviation. For this reason, the estimate *always* is made with a denominator of $n - 1$.

Chebyshev's Theorem

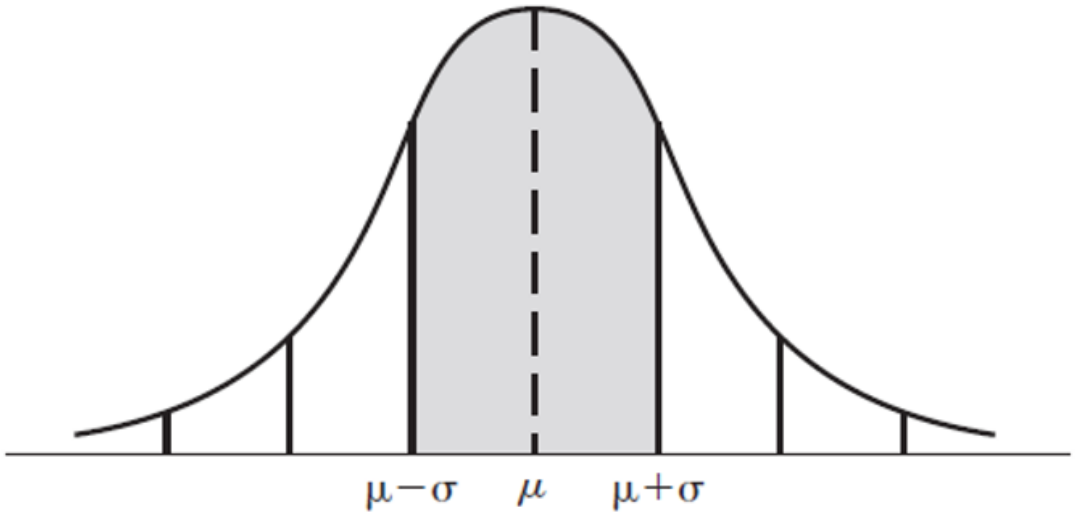
Chebyshev's Theorem:

- $\approx 68.26\%$ of values fall within $\pm 1\sigma$
- $\approx 95.44\%$ of values fall within $\pm 2\sigma$
- $\approx 99.72\%$ of values fall within $\pm 3\sigma$

Teaching Note: The Empirical Rule is a special case of Chebyshev's Theorem when the distribution is normal, giving us much tighter bounds.

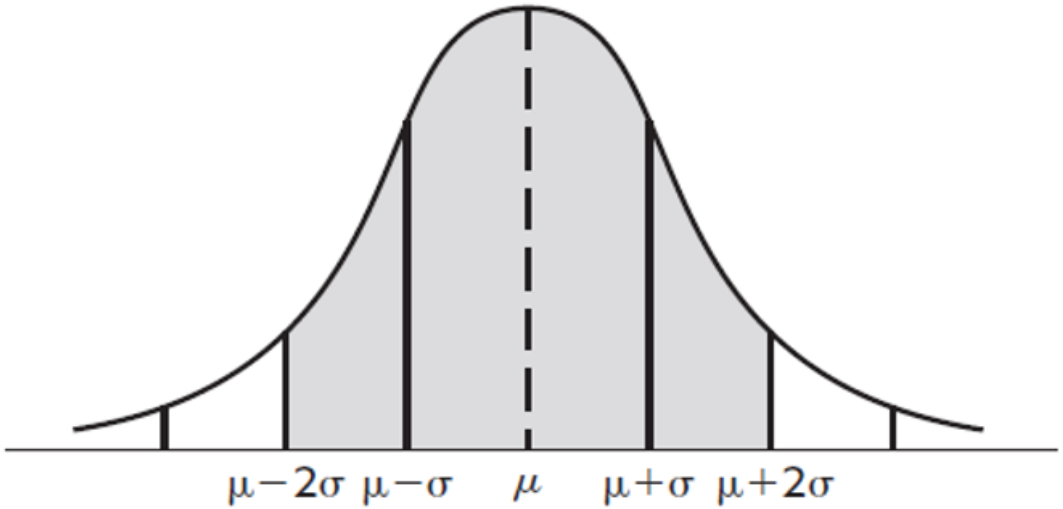
Chebyshev's

The Normal Distribution and the Standard Deviation



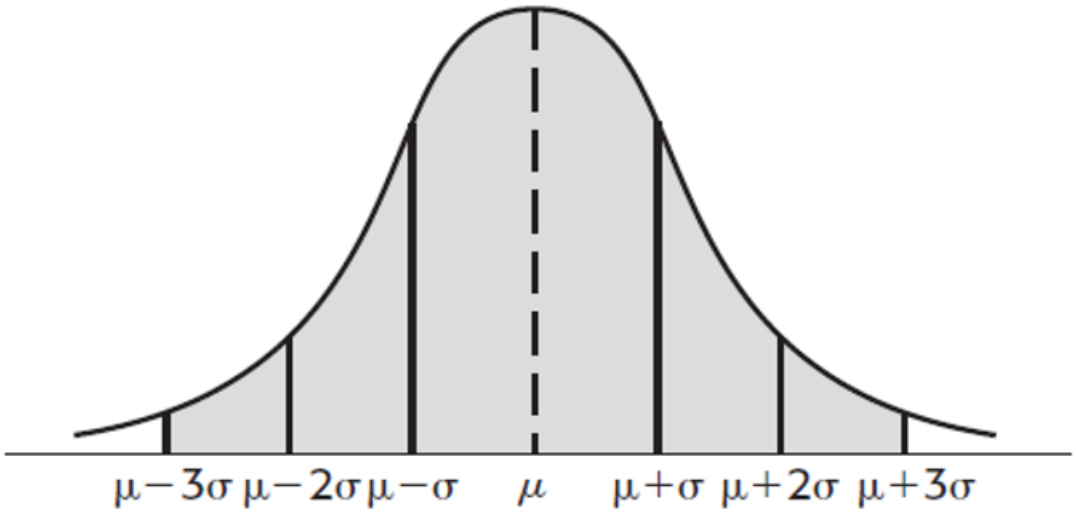
(a) 68.26% of all values lie within one standard deviation of the mean

Chebyshev's



(b) 95.44% of all values lie within two standard deviations of the mean

Chebyshev's



(c) 99.72% of all values lie within three standard deviations of the mean

Communicating Findings Effectively

- Policymakers value clarity over technical detail.
- Use simple graphs (bar charts, histograms) with clear titles and labels.
- Avoid excessive jargon; emphasize implications (e.g., "income inequality is greater in rural districts").
- Highlight actionable insights: What should decision makers focus on?