STOR 155.010 Lecture 04
Numerical Data

Aniruddhan (Ani) Ganesaraman

Fall 2026
UNC Chapel Hill

27 August 2026

Overview

1. Understanding Numerical Data

2. Visualizing Numerical Data

Centrality and Dispersion Measures

- **Centrality measures** provide insight into the “central” or typical values of numerical variables.
- Common centrality measures: **mean**, **median**, **mode**
- **Measures of dispersion** help gauge the spread of data (how stretched or squeezed) about a central value.
- Common measures of dispersion: **range**, **variance**, **standard deviation**

Mean

- The **mean** (or **average**) is calculated by summing all values of a numerical variable and dividing by the total number of observations.

- Highly sensitive to outliers.

- An appropriate measure of central tendency when the data is roughly symmetrically distributed.

Sample Mean

- The sample mean $\bar{x}$ is calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

  where $x_1, x_2, \cdots, x_n$ represent the n observed values.

- The population mean μ is also computed the same way. It is often not possible to calculate μ since population data are rarely available.

- The sample mean is a sample statistic, and serves as a point estimate of the population mean.

- Sample mean is a good estimate if the sample is good.

Median

- The **median** is the value that separates the higher half from the lower half of the values in a variable.
- If the dataset has an odd number of observations, the median is the middle number. If it's even, the median is the average of the two middle numbers.
- Unlike the mean, the median is not affected by extreme values (outliers), i.e., it is a **robust** statistic.

Computing the median

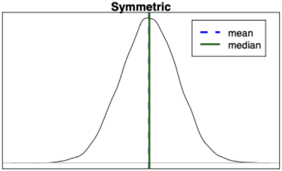
1. Order the numbers in increasing order.
2. If the multitude of numbers is odd, then the median is the number in the middle.
3. If the multitude of numbers is even, then the median is the average of the two middle numbers.

Example: Compute the mean and median of

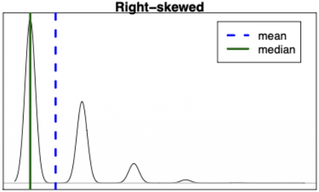
$$2, 5, 3, 8, 1, 6, 3, 2.$$

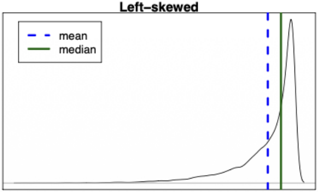
Mean vs. median

- If the distribution is **symmetric**, the center is mean $\approx$ median.



- If the distribution is **skewed** or has extreme outliers, center is often defined as the median
  - Right-skewed: mean $>$ median
  - Left-skewed: mean $<$ median





Mean vs. Median: A Comparison

- **Sensitivity to Outliers**:
  - Mean: Highly sensitive
  - Median: Not sensitive (Robust)
- **Applicability**:
  - Mean: Best for symmetric distributions
  - Median: Useful for skewed distributions
- **Calculation Complexity**:
  - Mean: Requires all data points
  - Median: If the data is ordered, not all points are required

Q1 and Q3

- The 25^{th} percentile is also called the first quartile, Q1. It is computed as the median of the ordered data to the left of the median.

- The 50^{th} percentile is also called the median.

- The 75^{th} percentile is also called the third quartile, Q3. It is computed as the median of the ordered data to the right of the median.

Example: Odd sample size

Data: $2, 4, 5, 7, 10$

Example: Odd sample size

Data: $2, 4, 5, 7, 10$

- Sample size $= 5$ (odd)
- Median (Q2) $= 5$
- Lower half: $2, 4$
- Upper half: $7, 10$
- Q1 $=$ Median of Lower half $= \frac{2+4}{2} = 3$
- Q3 $=$ Median of Upper half $= \frac{7+10}{2} = 8.5$

Example: Even sample size

Data: $1, 2, 4, 5, 7, 10$

Example: Even sample size

Data: $1, 2, 4, 5, 7, 10$

- Sample size $=$ 6 (even)
- Median (Q2) $= 4.5$
- Lower half: $1, 2, 4$
- Upper half: $5, 7, 10$
- Q1 $=$ Median of Lower half $= 2$
- Q3 $=$ Median of Upper half $= 7$

Mode

- The **mode** is the value that appears most frequently in a dataset.
- A dataset may have one mode (unimodal), more than one mode (multimodal), or no mode at all.
- The mode can be used for both numerical and categorical data.
- *Example:* The mode of $\{4, 2, 2, 3, 5, 4, 6, 2\}$ is 2, because it occurs three times, which is more than any other number in the dataset.

Range

- The range is the difference between the largest and smallest values in the dataset. It provides a measure of total spread.
- Mathematically, the range of n numbers $x_1, \ldots, x_n$ is defined to be the statistic

$$R = \max\{x_1, \ldots, x_n\} - \min\{x_1, \ldots, x_n\}.$$

- *Example:* The range of the 8 numbers $\{2, 5, 3, 8, 1, 6, 3, 2\}$ is _____.
- Between Q1 and Q3 is the middle 50% of the data. The range these data span is called the interquartile range, or the IQR.

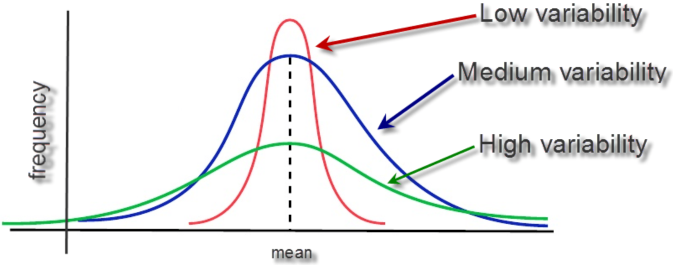
$$IQR = Q3 - Q1$$

Variance

Variance is roughly the average squared deviation from the mean.

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

Example: Compute the variance of $2, 5, 3, 8, 7$.



Different Idea

- Why use squares and not just take the average of the differences from the mean?

- In that case, we would always get

$$\begin{aligned}\frac{(x_1 - \bar{x}) + \cdots + (x_n - \bar{x})}{n} &= \frac{(x_1 + \cdots + x_n) - n\bar{x}}{n} \\ &= \frac{x_1 + \cdots + x_n}{n} - \bar{x} = 0,\end{aligned}$$

  which is not informative.

Standard deviation

The standard deviation is the square root of the variance, and has the same units as the data:

$$s = \sqrt{s^2}$$

Example: Compute the standard deviation of 2, 5, 3, 8, 7.

Overview

1. Understanding Numerical Data

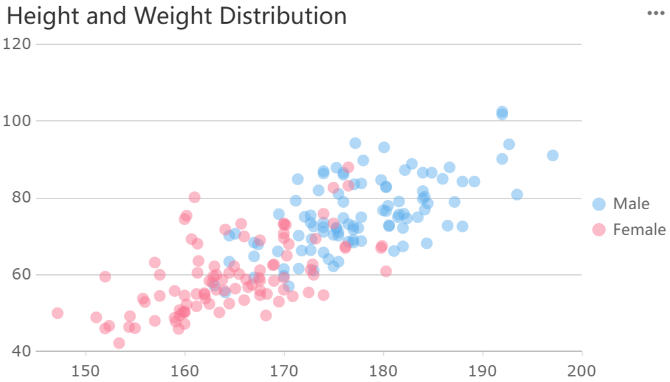
2. Visualizing Numerical Data

Visualization of Numerical Data

- Scatterplot
- Dot plot
- Histogram
- Boxplot

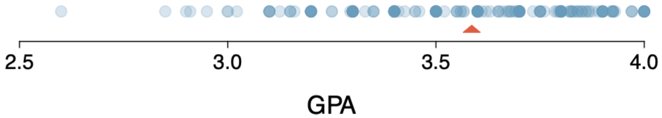
Scatterplot

- Useful for visualizing the relationship between **two** numerical variables.
- They provide information on positive/negative association.



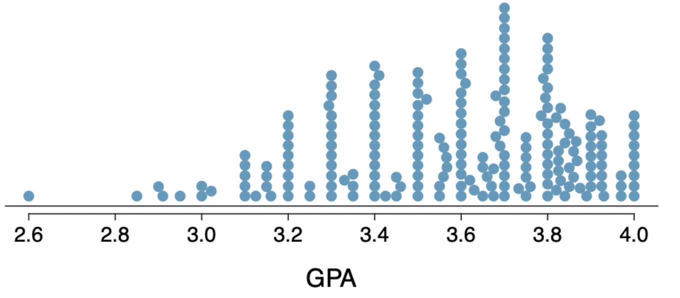
Dot plots

- Useful for visualizing **one** numerical variable.
- Darker colors represent areas where there are more observations.
- Indicate the mean of the data in the dot plot to give an idea about the center of the distribution.



Stacked Dot Plot

- Higher bars represent areas where there are more observations.
- Makes it a little easier to judge the center and the shape of the distribution.

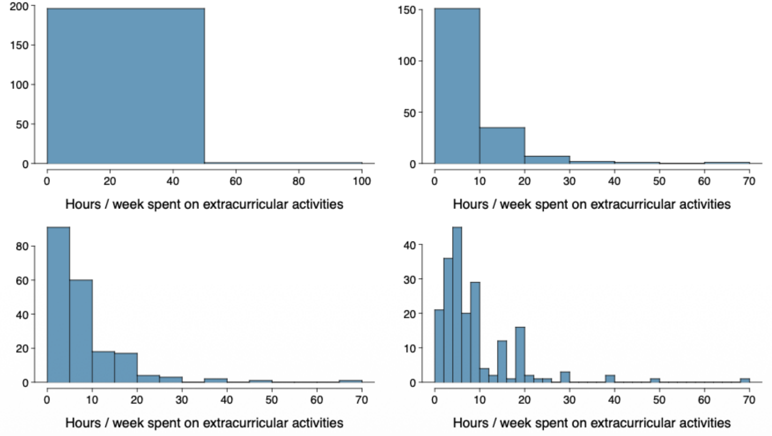


Histograms

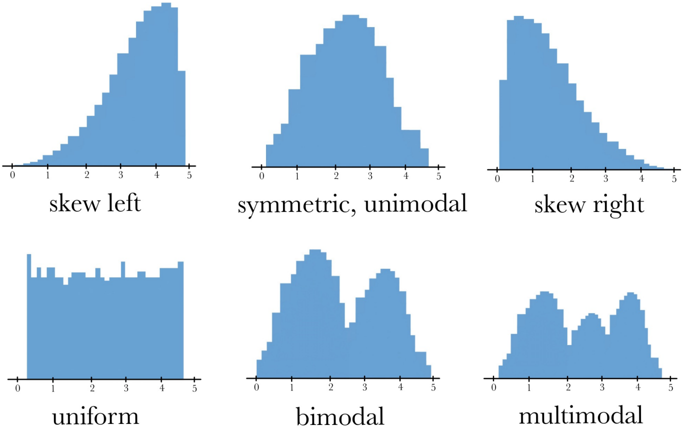
- Used to visualize **one** numerical variable.
- Rectangles whose **area** represents frequency; width represents a consecutive, non-overlapping interval.
- X-axis: possible values of the variable. Y-axis: frequency of each value.
- Horizontal range split into regular sub-ranges called **bins**.
- Height of each rectangle = number of data points in that bin.
- Great for spotting **skewness** and **kurtosis** in a large dataset.

Bin width

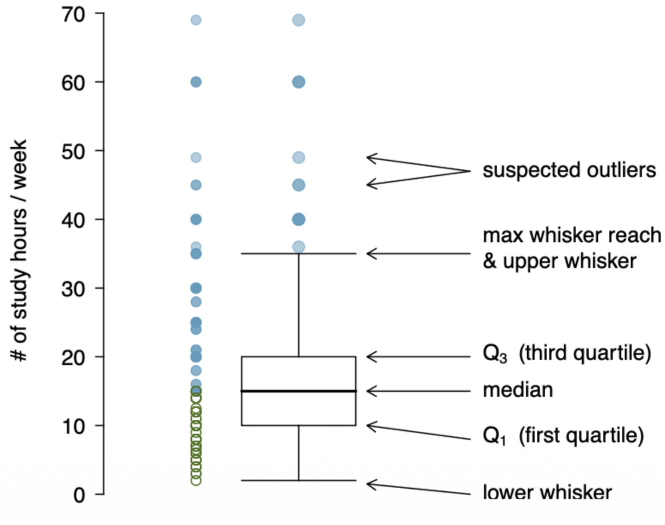
- The same dataset can lead to very different histograms based on the chosen bin width.
- Preferred: neither too many nor too few bins.



Skewness and Modality



Box plot


70
60
50
40
30
20
10
0
# of study hours / week
suspected outliers
max whisker reach
& upper whisker
Q3 (third quartile)
median
Q1 (first quartile)
lower whisker

Whiskers and outliers

- Define the **interquartile range** to be $IQR = Q_3 - Q_1$.
- **Whiskers** of a box plot can extend up to $1.5 \times IQR$ away from the quartiles.

$$\text{max upper whisker reach} = Q3 + 1.5 \times IQR$$
$$\text{max lower whisker reach} = Q1 - 1.5 \times IQR$$

$$\text{IQR} : 20 - 10 = 10$$
$$\text{max upper whisker reach} = 20 + 1.5 \times 10 = 35$$
$$\text{max lower whisker reach} = 10 - 1.5 \times 10 = -5$$

- A potential **outlier** is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Robust statistics

Median and IQR are more **robust** to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread