

Numerical Measures of Central Tendency

- Often, it is useful to have special numbers which summarize characteristics of a data set
- These numbers are called descriptive statistics or summary statistics.
- A measure of central tendency is a number that indicates the “center” of a data set, or a “typical” value.

Sample mean \bar{X} : For n observations,

$$\bar{X} = \sum X_i / n = \frac{X_1 + X_2 + \dots + X_n}{n}$$

- The sample mean is often used to estimate the population mean μ . (Typically we can't calculate the population mean.)

Alternative: Sample median M : the “middle value” of the data set. (At most 50% of data is greater than M and at most 50% of data is less than M .)

Steps to calculate M :

- (1) Order the n data values from smallest to largest.
- (2) Observation in position $(n+1)/2$ in the ordered list is the median M .
- (3) If $(n+1)/2$ is not a whole number, the median will be the average of the middle two observations.

For large data sets, typically use computer to calculate M .

Example: Per capita CO₂ emissions for 25 European countries (2006): Ordered Data: 3.3 4.2 5.6 5.6 5.7 5.7 6.2 6.3 7.0 7.6 8.0 8.1 8.3 8.6 8.7 9.4 9.7 9.9 10.3 10.3 10.4 11.3 12.7 13.1 24.5

$$\bar{X} = \frac{3.3 + 4.2 + \dots + 24.5}{25} = \frac{220.5}{25} = 8.82$$

$$n = 25 \Rightarrow \frac{n+1}{2} = \frac{26}{2} = 13$$

$$\text{So } M = 8.3$$

Luxembourg with 24.5 metric tons per capita is an outlier (unusual value).

What if we delete this country?

$$\bar{X} = \frac{196}{24} = \underline{8.17}$$

$$n = 24 \Rightarrow \frac{n+1}{2} = \frac{25}{2} = 12.5$$

$$\text{So } M = \frac{8.1 + 8.3}{2} = \underline{8.2}$$

Which measure was more affected by the outlier?

\bar{X} is more affected. The median is more resistant to extreme values (less sensitive).

Shapes of Distributions

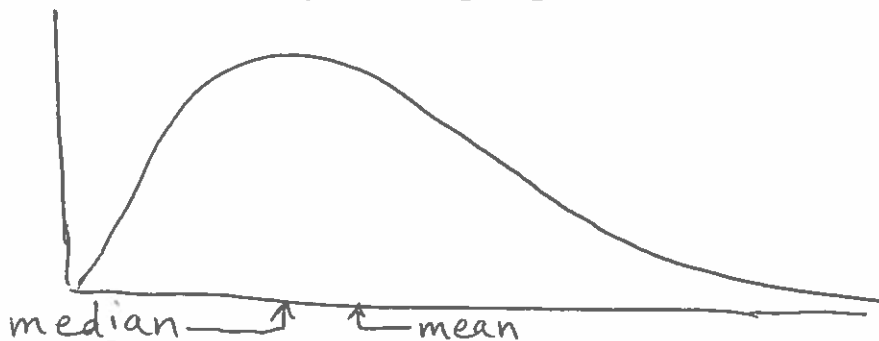
- When the pattern of data to the left of the center value looks the same as the pattern to the right of the center, we say the data have a symmetric distribution.

Picture:

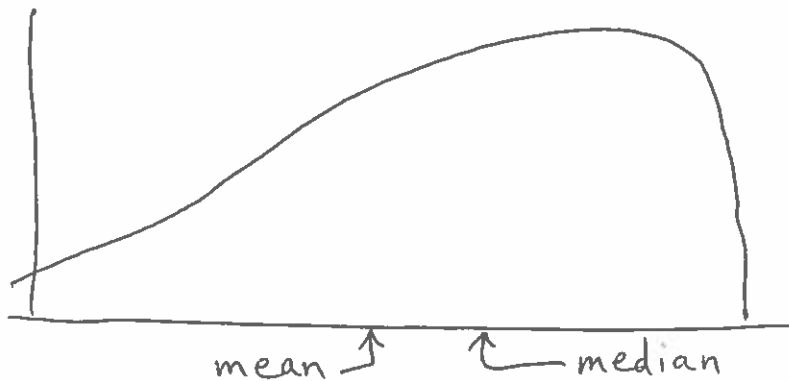


If the distribution (pattern) of data is imbalanced to one side, we say the distribution is skewed.

Skewed to the Right (long right “tail”). Picture:



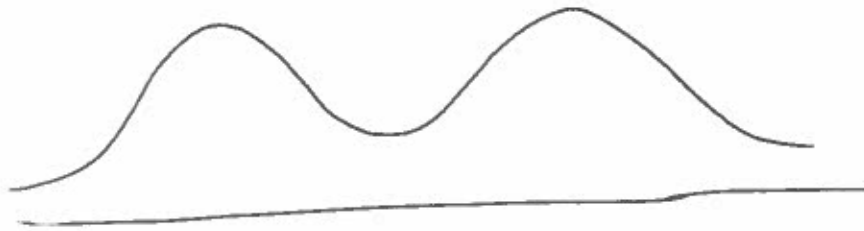
Skewed to the Left (long left “tail”). Picture:



Comparing the mean and the median can indicate the skewness of a data set.

Other measures of central tendency

- **Mode**: Value that occurs most frequently in a data set.
- In a histogram, the **modal class** is the class with the most observations in it.
- A **bimodal** distribution has two separated peaks:



The most appropriate measure of central tendency depends on the data set:

Skewed? Median

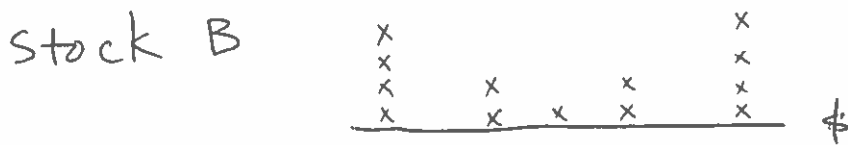
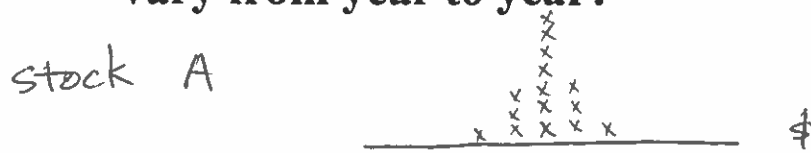
Symmetric? Mean

Categorical? Mode

Numerical Measures of Variability

- Knowing the center of a data set is only part of the information about a variable.
- Also want to know how “spread out” the data are.

Example: You want to invest in a stock for a year. Two stocks have the same average annual return over the past 30 years. But how much does the annual return vary from year to year?



Question: How much is a data set typically spread out around its mean?

Deviation from Mean: For each x -value, its deviation from the mean is: $X_i - \bar{X}$

Example (Heights of sample of plants):

Data: 1, 1, 1, 4, 7, 7, 7. $\Rightarrow \bar{X} = 4$

Deviations: -3, -3, -3, 0, 3, 3, 3

Squared Deviations: 9, 9, 9, 0, 9, 9, 9

- A common measure of spread is based on the squared deviations.
- **Sample variance:** The “average” squared deviation (using $n-1$ as the divisor)

sample variance

Definitional Formula:

$$s^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

Previous example: $s^2 = \frac{54}{6} = 9$

Shortcut formula: $s^2 =$

$$\frac{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}{n-1}$$

Another common measure of spread:

Sample standard deviation = positive square root of sample variance.

Previous example: **Standard deviation: $s = \sqrt{9} = 3$**

Note: s is measured in same units as the original data.

Why divide by $n-1$ instead of n ? Dividing by $n-1$ makes the sample variance a more accurate estimate of the population variance, σ^2 .

The larger the standard deviation or the variance is, the more spread/variability in the data set.

Usually use computers/calculators to calculate s^2 and s .

Rules to Interpret Standard Deviations

- **Think about the shape of a histogram for a data set as an indication of the shape of the distribution of that variable.**

Example: “Mound-shaped” distributions:



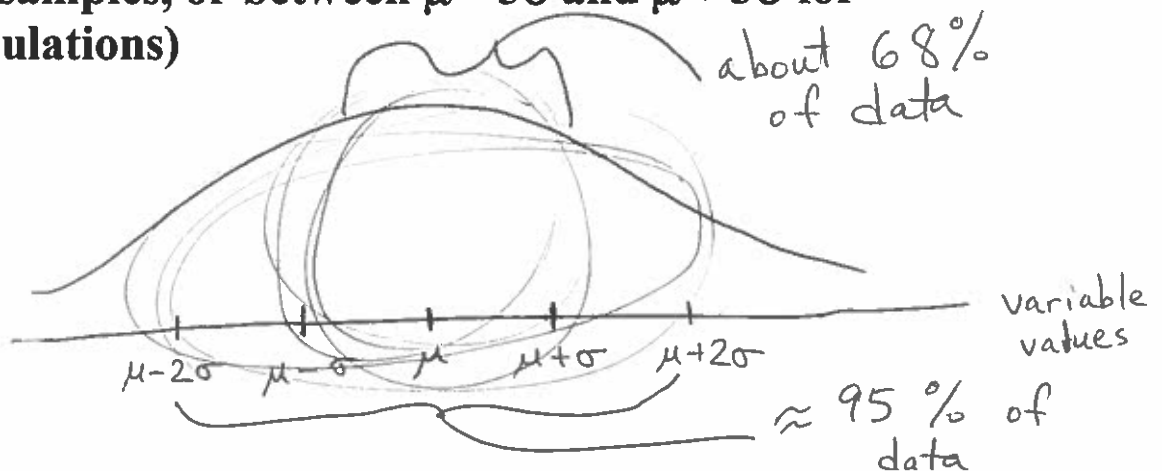
(roughly symmetric, peak in middle)

Special rule that applies to data having a mound-shaped distribution:

Empirical Rule: For data having a mound-shaped distribution,

- About 68% of the data fall within 1 standard deviation of the mean (between $\bar{x} - s$ and $\bar{x} + s$ for samples, or between $\mu - \sigma$ and $\mu + \sigma$ for populations)
 population mean \uparrow μ \leftarrow *population std. deviation.*
- About 95% of the data fall within 2 standard deviations of the mean (between $\bar{x} - 2s$ and $\bar{x} + 2s$ for samples, or between $\mu - 2\sigma$ and $\mu + 2\sigma$ for populations)
- About 99.7% of the data fall within 3 standard deviations of the mean (between $\bar{x} - 3s$ and $\bar{x} + 3s$ for samples, or between $\mu - 3\sigma$ and $\mu + 3\sigma$ for populations)

Picture:



Example: Suppose IQ scores have mean 100 and standard deviation 15, and their distribution is mound-shaped.

About 68% of scores fall between 85 and 115.
About 95% of scores " " 70 and 130.
About 99.7% " " " " 55 and 145.

Example: The rainfall data have a mean of 34.9 inches and a standard deviation of 13.7 inches.

Empirical Rule: If rainfall distribution is mound-shaped,
 $\approx 68\%$ of data between 21.2 and 48.6 inches
 $\approx 95\%$ of data between 7.5 and 62.3 inches
 $\approx 99.7\%$ " " " " -6.2 and 76 inches.
 $\uparrow 34.9 - 2(13.7)$ $\uparrow 34.9 + 2(13.7)$

What if the data may not have a mound-shaped distribution?

Chebyshev's Rule: For any type of data, the proportion of data which are within k standard deviations of the mean is at least:

$$1 - \frac{1}{k^2} \quad (\text{where } k > 1)$$

In the general case, at least what proportion of the data lie within 2 standard deviations of the mean?

at least $1 - \frac{1}{2^2} = \frac{3}{4} = 0.75$ (75%)

What proportion would this be if the data were known to have a mound-shaped distribution?

$$\approx 95\% \quad \text{or} \quad 0.95$$

Rainfall example revisited: $\bar{x} = 34.9$, $s = 13.7$

At least 75% of data are between $34.9 - 2(13.7)$ and $34.9 + 2(13.7)$ i.e., between 7.6 and 62.3 inches.

$$k=3 \rightarrow 1 - \frac{1}{3^2} = \frac{8}{9} \approx \underline{\underline{89\%}} \quad (\text{at least})$$

Numerical Measures of Relative Standing

- These tell us how a value compares relative to the rest of the population or sample.
- Percentiles are numbers that divide the ordered data into 100 equal parts. The p -th percentile is a number such that at most $p\%$ of the data are less than that number and at most $(100 - p)\%$ of the data are greater than that number.

Well-known Percentiles: Median is the 50th percentile. Lower Quartile (Q_L) is the 25th percentile: At most 25% of the data are less than Q_L ; at most 75% of the data are greater than Q_L .

Upper Quartile (Q_U) is the 75th percentile: At most 75% of the data are less than Q_U ; at most 25% of the data are greater than Q_U .

The 5-number summary is a useful overall description of a data set: (Minimum, Q_L , Median, Q_U , Maximum).

Example (Rainfall data):

(7, 29.1, 36.6, 42.8, 67)

Z-scores

-- These allow us to compare data values from different samples or populations.

-- The z-score of any observation is found by subtracting the mean, and then dividing by the standard deviation.

For any measurement x ,

Sample z-score:
$$z = \frac{x - \bar{x}}{s}$$

Population z-score:
$$z = \frac{x - \mu}{\sigma}$$

The z-score tells us how many standard deviations above or below the mean that an observation is.

Example: You get a 72 on a calculus test, and an 84 on a Spanish test.

Test data for calculus class: mean = 62, $s = 4$.

Test data for Spanish class: mean = 76, $s = 5$.

Calculus z-score:
$$\frac{72 - 62}{4} = \frac{10}{4} = 2.5$$

Spanish z-score:
$$\frac{84 - 76}{5} = \frac{8}{5} = 1.6$$

The calculus score was 2½ standard deviations above the mean.

Which score was better relative to the class's performance? Calculus

Your friend got a 66 on the Spanish test:

z-score: $\frac{66-76}{5} = \frac{-10}{5} = -2 \rightarrow$ 2 std. deviations below the mean.

Boxplots, Outliers, and Normal Q-Q plots

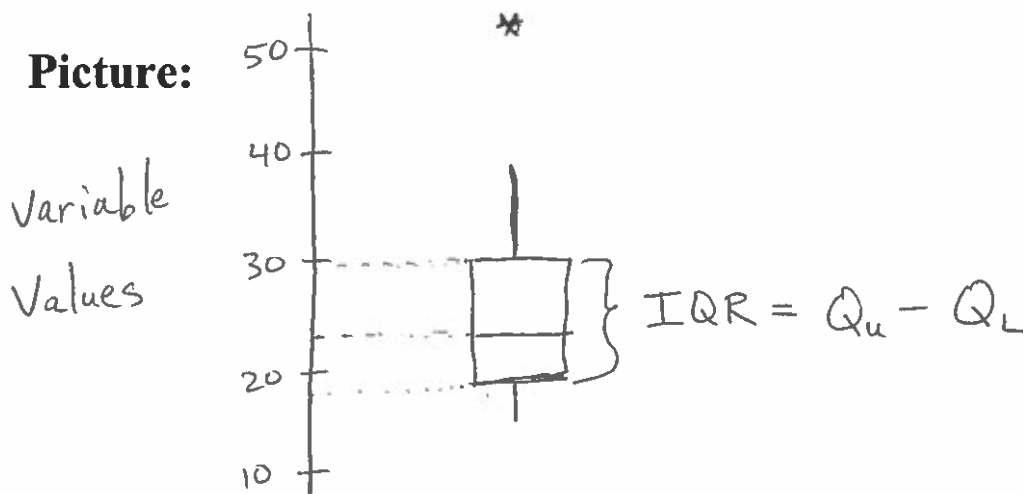
Outliers are observations whose values are unusually large or small relative to the whole data set.

Causes for Outliers:

- (1) Mistake in recording the measurement
- (2) Measurement comes from some different population
- (3) Simply represents an unusually rare outcome

Detecting Outliers

Boxplots: A boxplot is a graph that depicts elements in the 5-number summary.



- The “box” extends from the lower quartile Q_L to the upper quartile Q_U .
- The length of this box is called the Interquartile Range (IQR) of the data.
- $IQR = Q_U - Q_L$ ← A robust measure of variability (resistant to outliers)
- The “whiskers” extend to the smallest and largest data values, except for outliers.
- We generally use software to create boxplots.

Defining an outlier:

- If a data value is less than $Q_L - 1.5(IQR)$ or greater than $Q_U + 1.5(IQR)$, then it is considered an outlier and given a separate mark on the boxplot.
- A different rule of thumb is to consider a data value an outlier if its z-score is greater than 3 or less than -3.

Interpreting boxplots

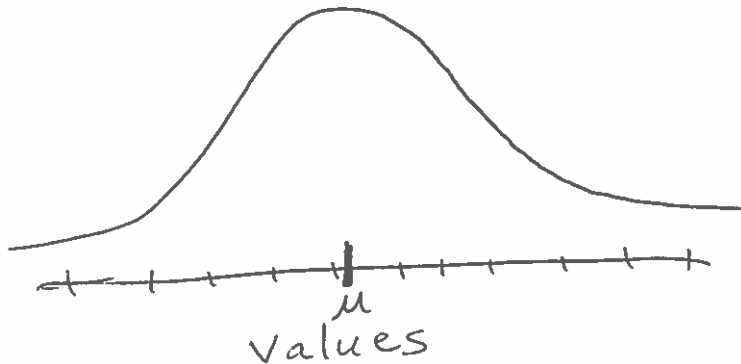
- A long “box” indicates large variability in the data set.
- If one of the whiskers is long, it indicates skewness in that direction.
- A “balanced” boxplot indicates a symmetric distribution.

Outliers should be rechecked to determine their cause. Do not automatically delete outliers from the analysis --- they may indicate something important about the population.

Assessing the Shape of a Distribution

-- A normal distribution is a special type of symmetric distribution characterized by its “bell” shape.

Picture:



- How do we determine if a data set might have a normal distribution?
- Check the histogram: Is it bell-shaped?
- More precise: Normal Q-Q plot (a.k.a. Normal probability plot). (see p. ~~250-251~~ 254-256)
- Plots the ordered data against the z-scores we would expect to get if the population were really normal.
- If the Q-Q plot resembles a straight line, it's reasonable to assume the data come from a normal distribution.
- If the Q-Q plot is nonlinear, data are probably not normal.