

STAT 515 --- STATISTICAL METHODS

Statistics: The science of using data to make decisions and draw conclusions

Two branches:

Descriptive Statistics:

- The collection and presentation (through graphical and numerical methods) of data
- Tries to look for patterns in data, summarize information

Inferential Statistics:

- Drawing conclusions about a large set of individuals based on data gathered on a smaller set
- Example:

Some Definitions

Population: The complete collection of units (individuals or objects) of interest in a study

Variable: A characteristic of an individual that we can measure or observe

Examples:

Sample: A (smaller) subset of individuals chosen from the population

- In statistical inference, we use *sample data* (the values of the variables for the sampled individuals) to make some conclusion (e.g., an estimate or prediction) about the population.

Example:

**How reliable is this generalization to the population?
For inference to be useful, we need some genuine
measure of its reliability.**

Types of Data:

Quantitative (Numerical) Data: Measurements recorded on a natural numerical scale (can perform mathematical operations on data).

Qualitative (Categorical) Data: Measurements classified into one of several categories.

Examples:

Sources of Data:

- **Published Source:** Many government, business, financial, and sports statistics are collected and archived in publications or online.
- **Designed Experiment:** Researcher imposes a treatment on individuals, then observes responses. (Researcher maintains strict control --- often in lab setting.)

Example:

- **Surveys:** Researcher selects sample of individuals and records their responses to questions

Example:

- **Observational Study:** Researcher observes individuals and measures variables, but has no control over process being observed.

Example:

Typically, an experiment is better for establishing cause/effect between two variables, but it's not always practical or possible.

Regardless of the type of study, we must ensure that we have a representative sample, one that has similar characteristics to the population.

The best kind of sample is a *simple random sample* (every subset has an equal chance of being selected)

Standard statistical methods assume the data are a random sample from the population.

Methods for Describing Data Sets

Important Principle in Statistics: Data Reduction

Example:

- **List of 100 numbers would be confusing, not informative**
- **Need to reduce data to a reasonable summary of information.**

Two ways:

- **Graphs and Plots**
- **Numerical Statistics**

Describing Qualitative Data

- **Data are categorized into classes**
- **The number of observations (data values) in a class is the class frequency**
- **Class Relative Frequency = class frequency / n**
- **The CRF's of all classes add up to 1**

Example:

Graphical Displays:

Bar graph: Height of bars indicates frequencies for each category (see p. 32)

Pie chart: Area of “pie slices” indicates relative frequency for each class (see p. 32)

Describing Quantitative Data

To detect and summarize patterns in a set of numerical data, we use:

- **Dot Plots:** these represent each data value with a dot along a numerical scale. When data values repeat, the dots pile up vertically at that value.
- **Stem-and-leaf Display** (good for small data sets): Separate each number in a data set into a stem and a leaf (usually the last digit). There is a column of all the stems in the data set, and at each stem, the corresponding leaf digits line up to the right.

Example:

See p. 42 for another example.

- **Histogram**: Numerical data values are grouped into measurement classes, defined by equal-length intervals along the numerical scale.

Like a bar graph, a histogram is a plot with the measurement classes on the horizontal axis and the class frequencies (or relative frequencies) on the vertical axis.

For each measurement class, the height of the bar gives the frequency (or RF) of that class in the data.

Example:

Guidelines for Selecting Measurement Class Intervals:

- Use intervals of equal width
- Each data value must belong to exactly one class
- Commonly, between 5 and 12 classes are used

Note: Different choices of Class Intervals (in position and number) may produce different-looking histograms.

Most often, we use software to help choose intervals and create histograms.

Histograms don't show individual measurement values (stem-and-leaf displays and dot plots do).

But for large data sets, histograms give a cleaner, simpler picture of the data.

Summation Notation

In statistics, we customarily denote our data values as x_1, x_2, \dots, x_n . (n is the total number of observations.)

The sum of a set of numbers is denoted with Σ .

That is, $x_1 + x_2 + \dots + x_n =$

Or, we might sum the squared observations:

$$x_1^2 + x_2^2 + \dots + x_n^2 =$$

Example:

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,

$$\bar{X} = \Sigma X_i / n =$$

- 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: Mark McGwire's Home Run totals (1987-1998). Ordered Data: 9, 9, 22, 32, 33, 39, 39, 42, 49, 52, 58, 70.

Two years with 9 home runs are outliers (unusual values) due to injury and a players' strike.

What if we delete these years?

Which measure was more affected by the outliers?

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?

Symmetric?

Categorical?

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:

Example (Heights of sample of plants):

Data: 1, 1, 1, 4, 7, 7, 7.

Deviations:

Squared Deviations:

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

Definitional Formula:

$$s^2 =$$

Previous example: $s^2 =$

Shortcut formula: $s^2 =$

Another common measure of spread:

Sample standard deviation = positive square root of sample variance.

Previous example: Standard deviation: $s =$

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)**
- **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.

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

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:

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

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

Rainfall example revisited:

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):

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:

Population z-score:

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:

Spanish z-score:

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

Your friend got a 66 on the Spanish test:

z-score:

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.

Picture:

- 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.
- $IQR = Q_U - Q_L$
- The “whiskers” extend to the smallest and largest data values, except for outliers.

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.
- We generally use software to create boxplots.

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. 261-263)**
- **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.**