Data Analysis: Quantitative Data

Dr. Sahar Hassan

Why do I need an analysis plan?

- To make sure the questions & your data collection instrument will get the information you want
- To align your desired “report” with the results of analysis and interpretation
- To improve reliability consistent measures over time

Key components of a data analysis plan

- Purpose of the evaluation
- Questions
- What you hope to learn from the question
- Analysis technique
- How data will be presented

Getting your data ready

- Assign a unique identifier
- Organize & keep all forms (questionnaires, interviews, testimonials)
- Check for completeness & accuracy
- Remove those that are incomplete or do not make sense

Dig deeper

- Did different groups show different results?
- Were there findings that surprised you?
- Are there things you don’t understand very well – further study needed?

Analyzing and Interpreting Quantitative Data

- Quantitative Data is
  - ✓Presented in a numerical format
  - ✓Collected in a standardized manner

  e.g. surveys, closed-ended interviews, tests
  - ✓Analyzed using statistical techniques

Analyzing Survey Data

Do you want to report…

- ✓ how many people answered a, b, c, d?
- ✓ the average number or score?
- ✓ a change in score between two points in time?
- ✓ how people compared?
- ✓ how many people reached a certain level?

Types of Measurements

- Ordinal – rank order, (1st,2nd,3rd,etc.)
- Nominal – categorized or labeled data
- (red, green, blue, male, female)
- Ratio (Interval) – indicates order as well as magnitude. An interval scale does not include Zero i.e. grade, income

Types of Variables

- Independent Variable – controlled or manipulated by the researcher; causes a change in the dependent variable (x-axis)
- Dependent Variable – the variable being measured (y-axis)
- Discreet Variable – has a fixed value i.e. # cars in a parking lot
- Continuous Variable - can assume any value

Statistical Analysis

Descriptive statistics

- Used to describe and synthesize data

Inferential statistics

- Used to make inferences about the population based on sample data

Frequency Distributions

- A systematic arrangement of numeric values on a variable from lowest to highest, and a count of the number of times (and/or percentage) each value was obtained
- Frequency distributions can be described in terms of:

✓ Shape

✓ Central tendency

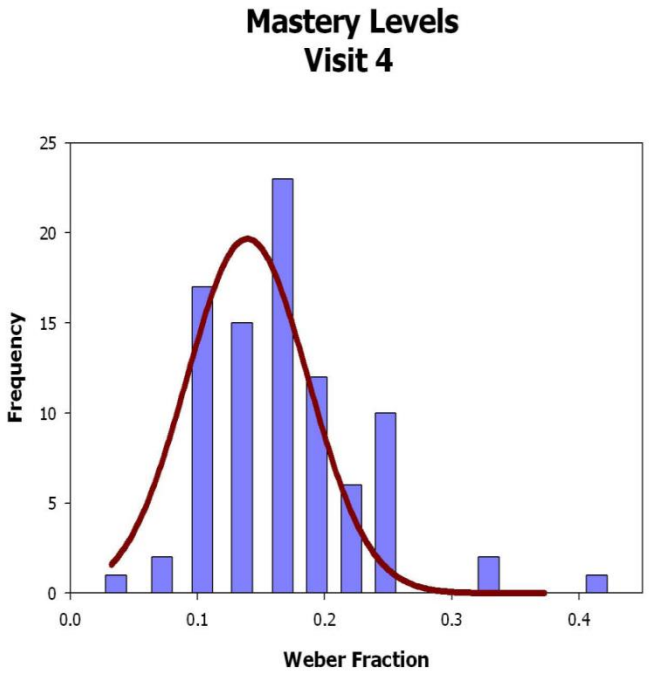
✓ Variability

- Can be presented in a table (*N*s and percentages) or graphically (e.g., **frequency polygons**)

Normal Distribution (normal curve)

Data can usually be characterized by a normal distribution

- Central tendency is represented by the peak of the distribution
- Dispersion is represented by the width of the distribution



Shapes of Distributions

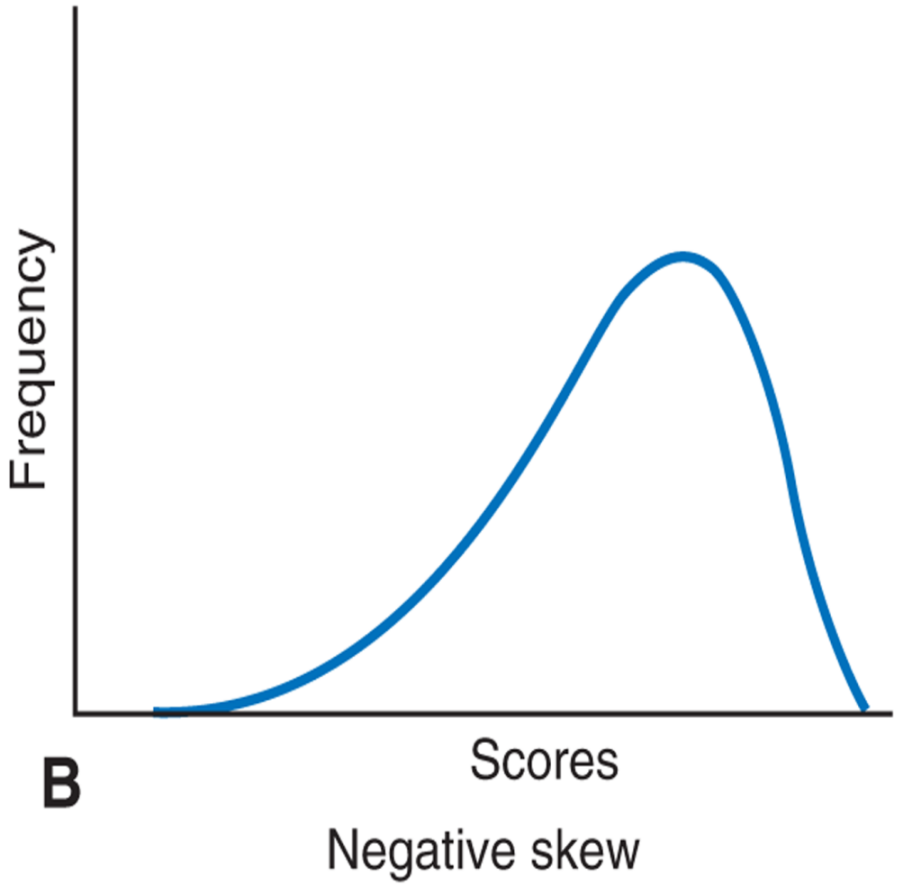
Symmetry

- Symmetric
- Skewed (asymmetric)

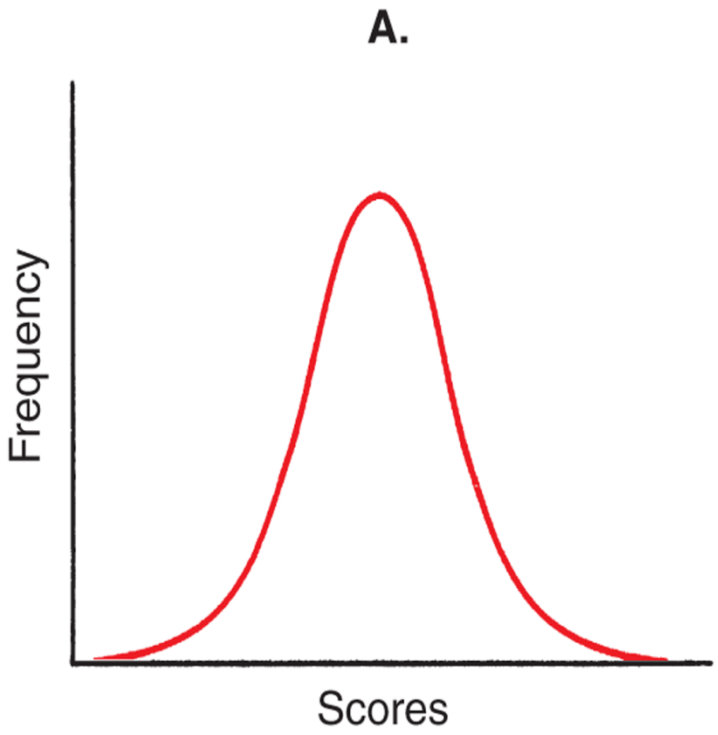
✓ Positive skew (long tail points to the right)

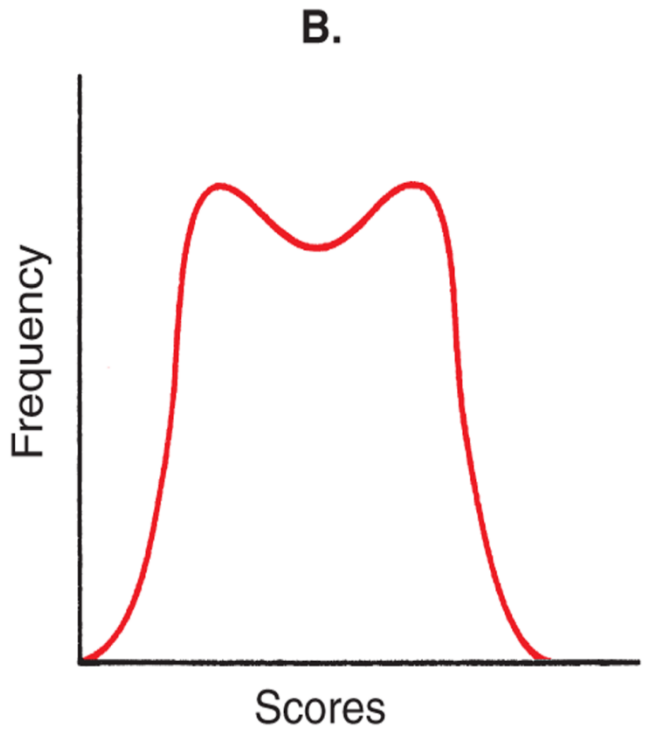
✓ Negative skew (long tail points to the left)

Examples of Skewed Distribution



Examples of Symmetric Distributions





Common descriptive statistics

- Count (frequencies)
- Percentage

Measures of Central Tendency

- Mean (average)
- Mode (most frequent)
- Median (middle)
- Range

Common descriptive statistics

Measures of Dispersion

- Standard deviation: a measure that is used to quantify the amount of variation or dispersion of a set of data values

Measures of Association

- Correlation

Which calculation do I use?
It depends on what you want to know..

Do you want to know how many individuals checked each answer?	Frequency
Do you want the proportion of people who answered in a certain way?	Percentage
Do you want the average number or average score?	Mean
Do you want the middle value in a range of values or scores?	Median
Do you want to show the range in answers or scores?	Range
Do you want to compare one group to another?	Cross tab
Do you want to report changes from pre to post?	Change score
Do you want to show the degree to which a response varies from the mean?	Standard deviation

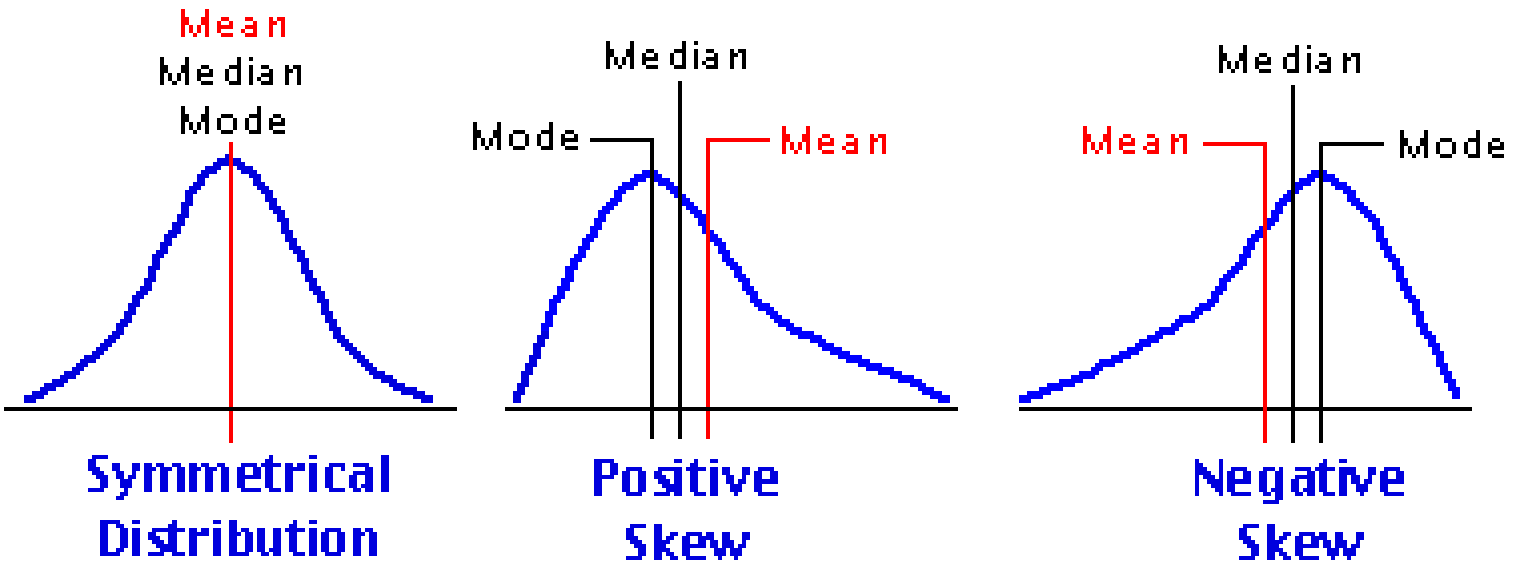
Comparison of Measures of Central Tendency

- Mode: useful mainly as gross descriptor, especially of nominal measures
- Median: useful mainly as descriptor of typical value when distribution is skewed (e.g., household income)
- Mean: most stable and widely used indicator of central tendency

Skewness of distributions

- Measures look at how unbalanced distributions are: how far from the ideal of the normal curve they are
- When the median and the mean are different, the distribution is skewed.
- The greater the difference, the greater the skew.

Different Shapes of Distributions

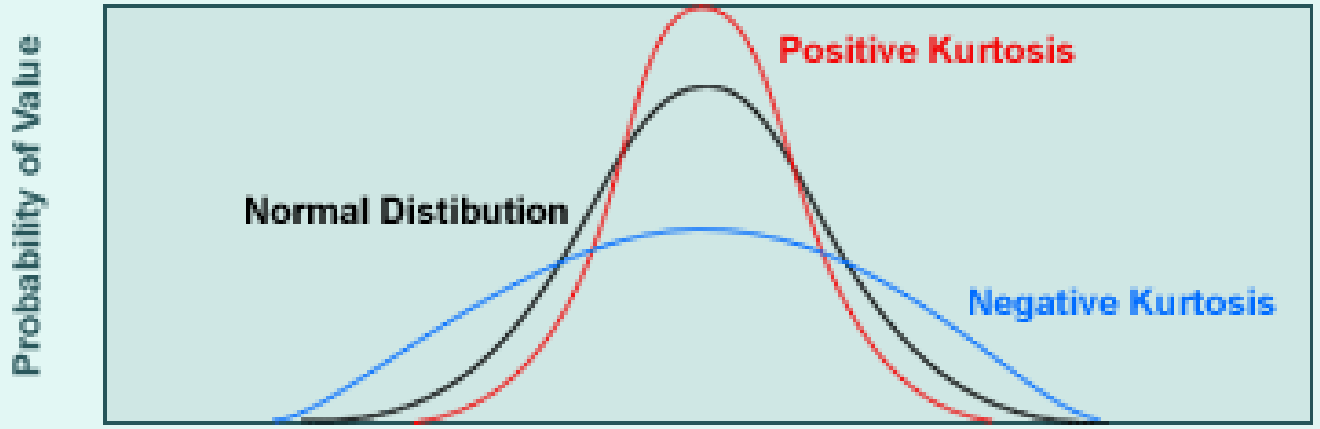


Source: http://faculty.vassar.edu/lowry/f0204.gif

Kurtosis

- Measures of kurtosis look at how sharply the distribution rises to a peak and then drops away

Normal Distribution and Distributions Illustrating Kurtosis (±)



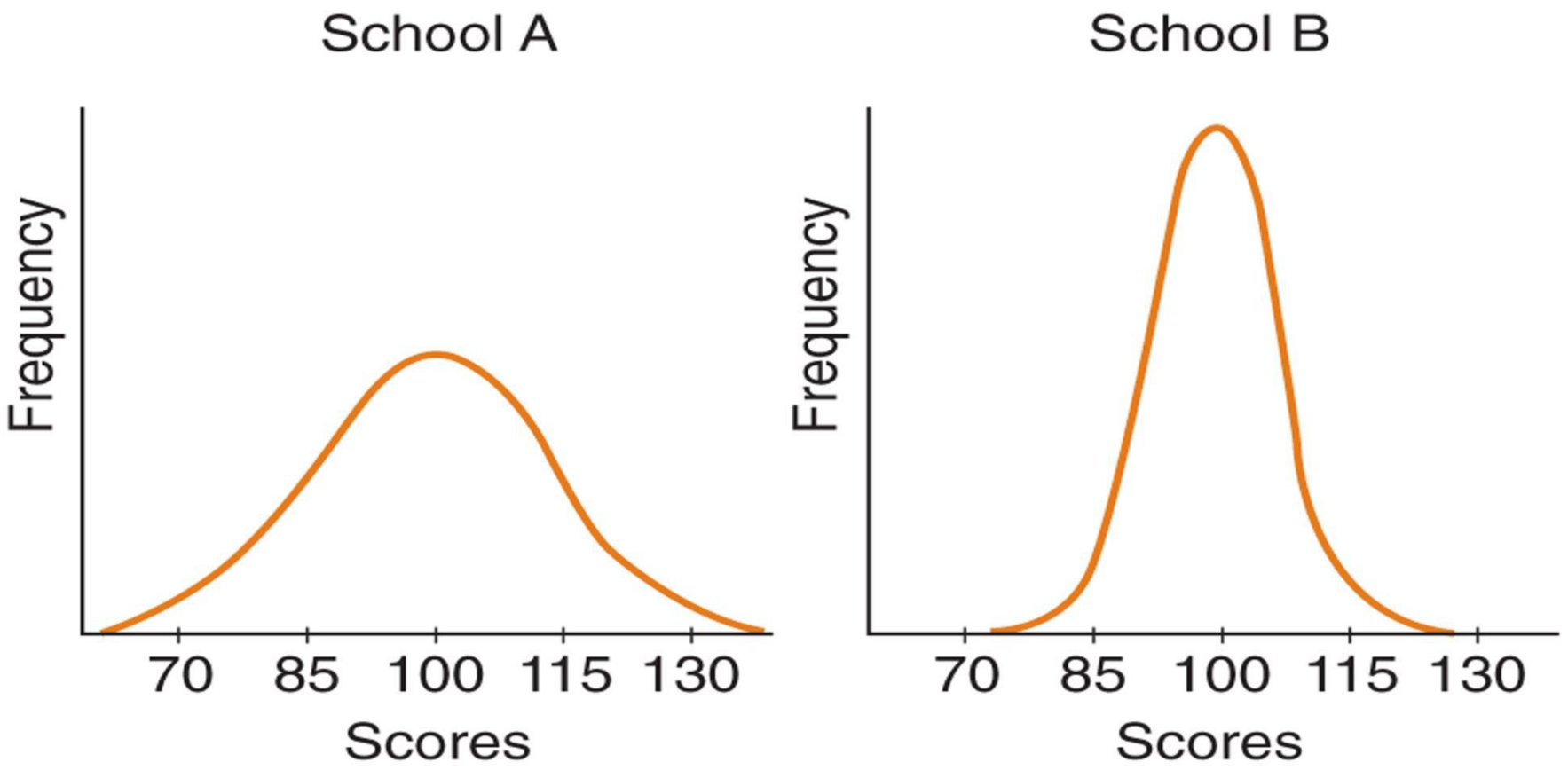
Variability

- The degree to which scores in a distribution are spread out or dispersed (how spread out the data is)

✓ Homogeneity: little variability

✓ Heterogeneity: great variability

Two Distributions of Different Variability



Indexes of Variability

- Range: highest value minus lowest value
- Standard deviation (SD): average deviation of scores in a distribution from the mean

Bivariate Descriptive Statistics

Bivariate Descriptive Statistics

- Used for describing the relationship between two variables
- Two common approaches:

✓ Crosstabs (contingency tables)

✓ Correlation coefficients

Correlation Coefficients (cont.)

- ✓ The correlation coefficient *r* measures the strength & direction of a linear relationship between two variables on a scatterplot
- ✓ It can range from −1.00 to +1.00

Correlation Coefficients

Negative relationship (0.00 to −1.00)

- ✓ One variable increases in value as the other decreases, e.g., amount of exercise and weight.

Positive relationship (0.00 to +1.00)

- ✓ Both variables increase, e.g., calorie consumption and weight

Correlation Coefficients (cont.)

Exactly –1. A perfect downhill (negative) linear relationship

–0.70. A strong downhill (negative) linear relationship

–0.50. A moderate downhill (negative) relationship

–0.30. A weak downhill (negative) linear relationship

0. No linear relationship

+0.30. A weak uphill (positive) linear relationship

+0.50. A moderate uphill (positive) relationship

+0.70. A strong uphill (positive) linear relationship

Exactly +1. A perfect uphill (positive) linear relationship

Correlation Coefficients (cont.)

- ✓ The greater the absolute value of the coefficient, the stronger the relationship:
- With multiple variables, a correlation matrix can be displayed to show all pairs of correlations.

Inferential Statistics

Inferential Statistics

- A means of drawing conclusions about a population given data from a sample
- Based on laws of probability

Statistical Inference—Two Forms

- Estimation of parameters
- Hypothesis testing (more common)

Inferential Statistics

Null Hypothesis

- Statistical hypotheses usually assume no relationship between variables
- E.g. There is no association between eye color & eyesight
- If the result of your statistical test is significant, then the original hypothesis is false and you can say that the variables in your experiment are somehow related

Hypothesis Testing

- Based on rules of negative inference: research hypotheses are supported if null hypotheses can be rejected
- Involves statistical decision making to either:
  - Accept the null hypothesis, or
  - Reject the null hypothesis

One-Tailed and Two-Tailed Tests

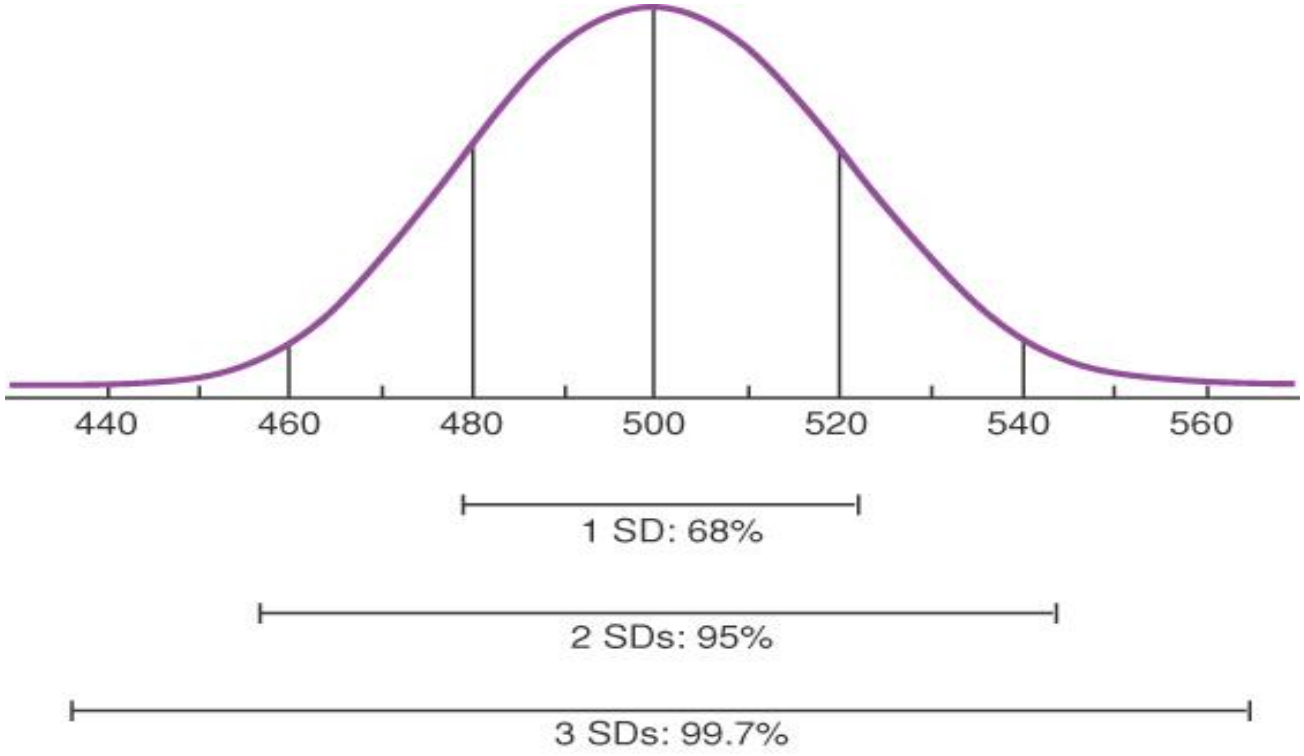
Two-tailed tests

Hypothesis testing in which both ends of the sampling distribution are used to define the region of improbable values

One-tailed tests

Critical region of improbable values is entirely in one tail of the distribution—the tail corresponding to the direction of the hypothesis

Sampling Distribution



Overview of Hypothesis-Testing Procedures

- Select an appropriate test statistic
- Establish the level of significance (e.g., $\alpha = .05$)
- Select a one-tailed or a two-tailed test
- Compute test statistic with actual data

Overview of Hypothesis-Testing Procedures (cont'd)

- Obtain a tabled value for the statistical test
- Compare the test statistic to the tabled value
- Make decision to accept or reject null hypothesis

Commonly Used Bivariate Statistical Tests

1) *t*-Test: Tests the difference between two means

- *t*-Test for independent groups (between subjects)
- *t*-Test for dependent groups (within subjects)

Commonly Used Bivariate Statistical Tests

2) Analysis of Variance (ANOVA)

- Tests the difference between 3+ means
  - One-way ANOVA
  - Multifactor (e.g., two-way) ANOVA
  - Repeated measures ANOVA (within subjects)

Commonly Used Bivariate Statistical Tests

3) Correlation

- Pearson's *r,* a parametric test
- Tests that the relationship between two variables is not zero
- Used when measures are on an interval or ratio scale

Commonly Used Bivariate Statistical Tests

4) Chi-Square Test

- Tests the difference in proportions in categories within a contingency table
- A nonparametric test

Commonly Used Bivariate Statistical Tests

2. Analysis of variance (ANOVA): tests difference btw 3+ means

3. Pearson's r: tests that the R.S. btw 2 variables is not zero (is data is on an interval/ratio scale)

4. Chi-square test: tests difference in proportions in categories within a contingency table

END