Key Principles Underlying Statistical Inference: Probability and the Normal Distribution

CHAPTER 3

OBJECTIVES

After studying this chapter, you should be able to:

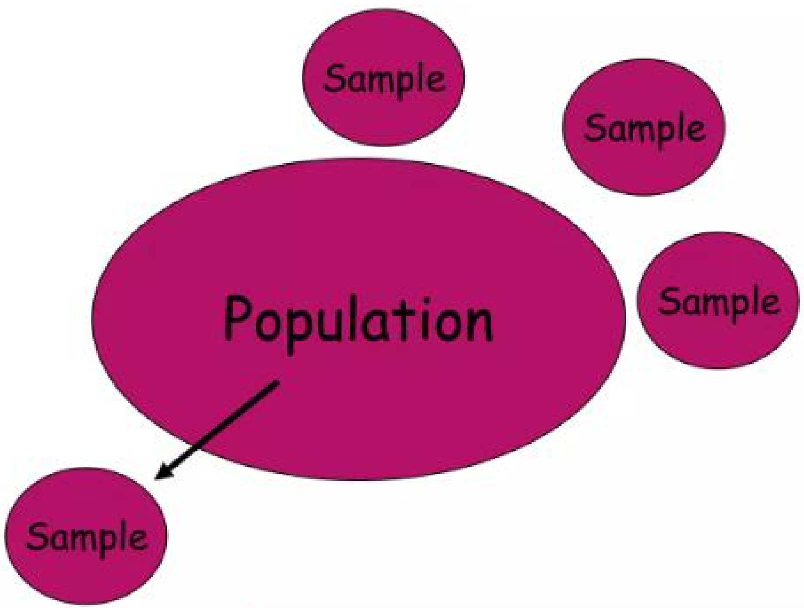
1. Explain the importance of probability theory for statistical inference.
2. Define the characteristics of a probability measure, and explain the difference between a theoretical probability distribution and an empirical probability distribution (*a priori* vs. *a posteriori*).
3. Compute marginal, joint, and conditional probabilities from a cross-tabulation table and correctly interpret their meaning.
4. Define and derive sensitivity, specificity, predictive value, and efficiency from a cross-tabulation table.
5. Identify and describe the characteristics of a normal distribution.
6. Use a standard normal distribution to obtain z-scores and percentiles.
7. Explain the importance of the central limit theorem.

FUNDAMENTAL CONCEPTS IN RESEARCH

- *One of the main objectives of research is to draw meaningful conclusions about a population, based on data collected from a sample of that population*

- researchers use statistical inference as their tool for obtaining information from a sample of data about the population from which the sample is drawn.

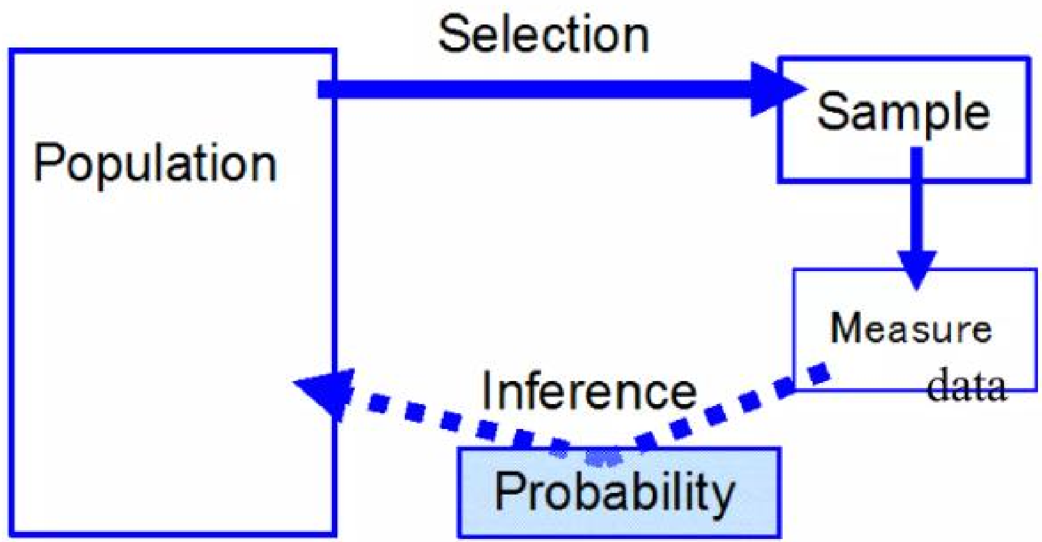
Inferential Statistics



Draw inferences about the larger group

- Statistical inference uses probability to help the researchers to assess the meaning of their findings.
- A particularly important probability distribution for statistical inference in health-related research is the normal (or Gaussian) distribution.

Chain of Reasoning for Inferential Statistics



Are our inferences valid?…Best we can do is to calculate probabi
about inferences

Example of statistical inference

- A recent population-based study of an urban area using the New York City Health and Nutrition Examination Survey (NYCHANES).

- In this study, the authors found that the prevalence of diabetes (diagnosed and undiagnosed combined) among adults aged 20 and above was 12.5%. (Thorpe et al., 2009)

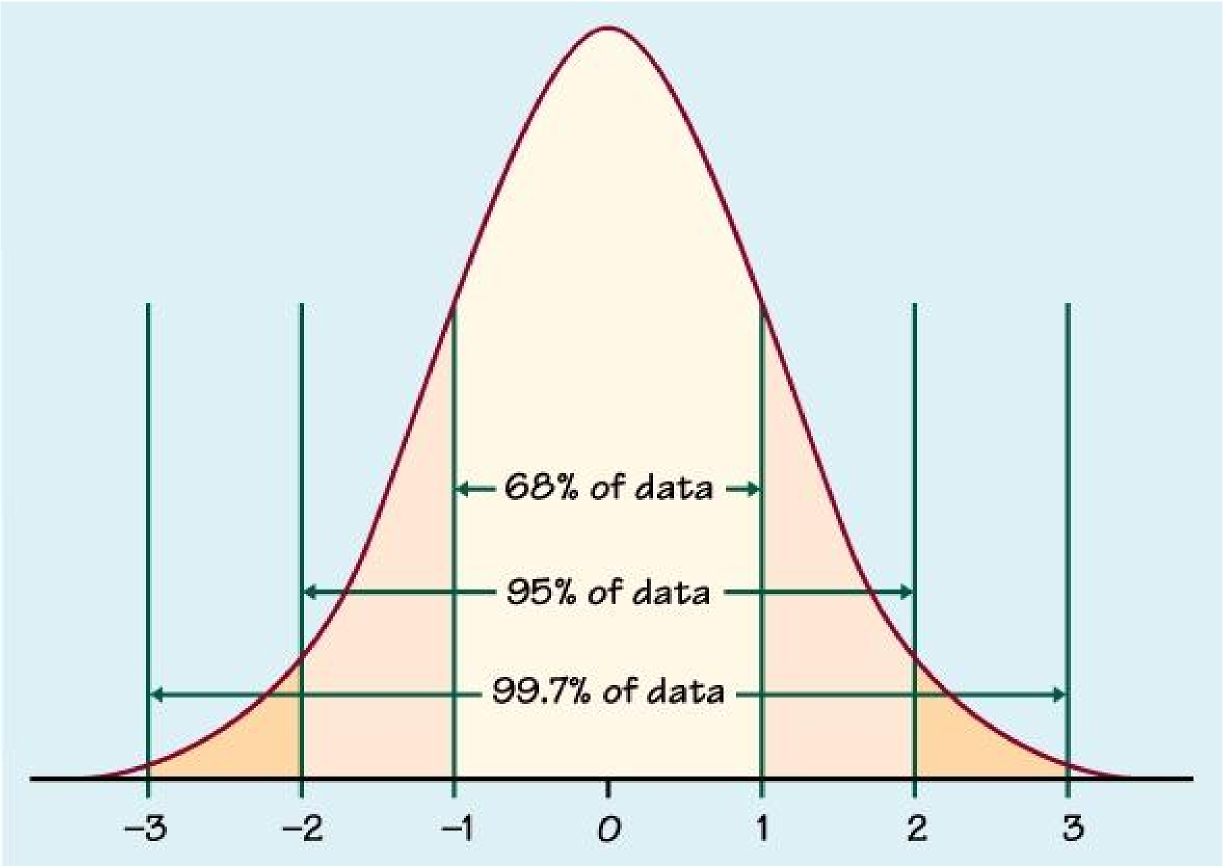
Statistical Inference

- Involves obtaining information from sample of data about population from which sample was drawn & setting up a model to describe this population
- When random sample is drawn from population, every member of population has equal chance of being selected in the sample

Types of Statistical Inference

- Parameter Estimation takes two forms
  - **Point Estimation**: when estimate of population parameter is single number
    - Ex. Mean, median, variance & SD
- **Hypothesis-Testing**:
  - More common type

Normal Curve



Normal Curve

- 68% of cases fall within ± 1 SD of the mean
- 96% of cases fall within ± 2 SD of the mean
- 100% of cases fall within ± 3 SD of the mean.

Normal curve

- In practice, this distribution is among the most important distributions in statistics for three reasons (Vaughan, 1998).

- First, although most distributions are not exactly normal, many biological and population-level variables (such as height and weight) tend to have approximately normal distributions.
- Second, many inferential statistics assume that the populations are distributed normally.

Normal curve

- In practice, this distribution is among the most important distributions in statistics for three reasons (Vaughan, 1998).
- Third, the normal curve is a probability distribution and is used to answer questions about the likelihood of getting various particular outcomes when sampling from a population

Normal curve

- For example, when we discuss hypothesis testing, we will talk about the probability(or the likelihood) that a given difference or relationship could have occurred by chance alone.

Normal curve

- Useful Characteristics of the Normal Distribution.

- it is bell shaped
- the mean, median, and mode are equal
- it is symmetrical about the mean
- the total area under the curve above the x-axis is equal to 1.

Normal curve

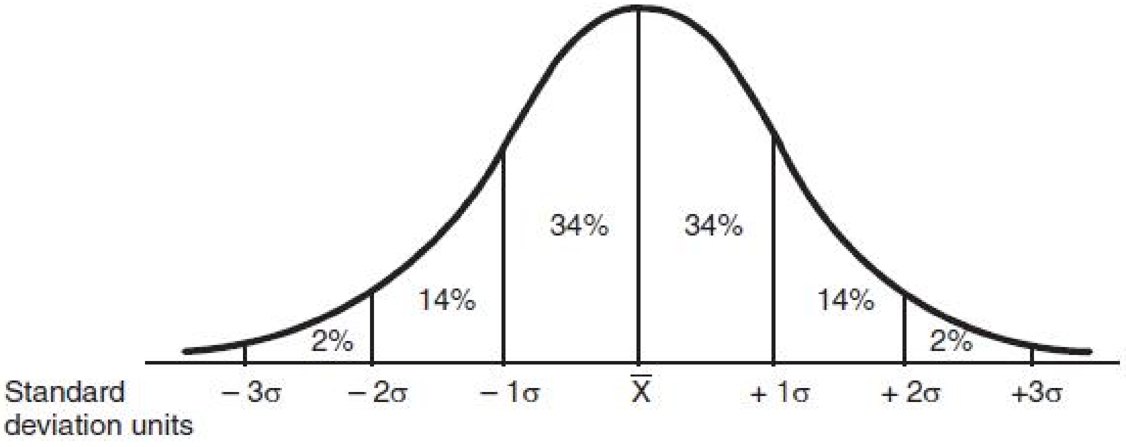


FIGURE 3-1 Normal distribution with Standard deviation units.

Normal Distribution & Z score: Understanding and Using z-Scores

- A z-score measures the number of SDs that an actual value lies from the mean; it can be positive or negative.
- Knowing the probability that someone will score below, at, or above the mean on a test can be very useful.

$$z = \frac{x - \mu}{\sigma}$$

- In health care, the criteria that typically define laboratory tests (e.g., glucose, thyroid, and electrolytes) as abnormal are based on the standard normal distribution, with scores that occur less than 95% of the time
- In particular, those with a z-score of ±2 or greater (representing very large and very small values) are defined as abnormal (Seaborg, 2007).

Normal Distribution & Z score: Understanding and Using z-Scores

- CENTRAL LIMIT THEOREM
- Two Important Z scores:
  - $\pm$ 1.96z = 95% confidence interval
  - $\pm$ 2.58z = 99% confidence interval

Confidence Interval (CI)

- Degree of confidence, expressed as a percent, that the interval contains the population mean (or proportion), & for which we have an estimate calculated from sample data
  - 95% CI = X $\pm$ 1.96 (standard error)
  - 99% CI = X $\pm$ 2.58 (standard error)

Confidence Interval (CI)

- *In general, the central limit theorem states that when a number of different samples are drawn from the same population, the distribution of the sample means tends to be normally distributed.*
- If you draw a sample from a population and calculate its mean, how close have you come to knowing the mean of the population?
- Statisticians have provided us with formulas that allow us to determine just how close the mean of our sample is to the mean of the population.

Confidence Interval (CI)

- When many samples are drawn from a population, the means of these samples tend to be normally distributed; that is, when they are charted along a baseline, they tend to form the normal curve.
- The larger the number of samples, the more the distribution approximates the normal curve.
- Also, if the average of the means of the samples is calculated (the mean of the means), this average (or mean) is very close to the actual mean of the population
- Again, the larger the number of samples, the closer this overall mean is to the population mean.

Confidence Interval (CI)

- When many samples are drawn from a population, the means of these samples tend to be normally distributed; that is, when they are charted along a baseline, they tend to form the normal curve.
- The larger the number of samples, the more the distribution approximates the normal curve.
- Also, if the average of the means of the samples is calculated (the mean of the means), this average (or mean) is very close to the actual mean of the population
- Again, the larger the number of samples, the closer this overall mean is to the population mean.

Confidence Interval (CI)

- It also states that the SD of the distribution of the sample means (i.e., the standard error of the mean used in constructing confidence intervals) can be computed using this equation:

$$se_{\overline{X}} = \frac{s}{\sqrt{n}}$$

- The larger the sample size, the smaller the standard error (and thus the more accurate the measure)
- In general, the approximation to normality of the sampling distribution of the mean becomes better as the sample size increases.
- A sample size of 30 or greater has been found to be sufficient for the central limit theorem to apply (Vaughan, 1998).

A z-score can give information about
a. the mean of a distribution.
b. the standard deviation (SD) of a distribution.
c. the percentile rank of a data point.
d. none of the above.

A z-score of 0 corresponds to the
a. mean.
b. SD.
c. interquartile range.
d. 75th percentile.

A sample population curve is more likely to look like the population curve when

a. the bell shape is wide.
b. the sample size is small.
c. the sample size is greater than 30.
d. none of the above.

Percentiles

- Tells the relative position of a given score
- Allows us to compare scores on tests with different means & SDs.
- Calculated as

- $\frac{\text{(\# of scores less than given score)}}{\text{total \# of scores}} \times 100$

Percentile

- 25th percentile = 1st quartile
- 50th percentile = 2nd quartile
  Also the median
- 75th percentile = 3rd quartile

Standard Scores

- Way of expressing a score in terms of its relative distance from the mean
  - z-score is example of standard score
- Standard scores are used more often than percentiles
- Transformed standard scores often called T-scores
  - Usually has M = 50 & SD = 10

Standard Error of Mean (SE)

- Is standard deviation of the population
- Constant relationship between SD of a distribution of sample means (SE), the SD of population from which samples were drawn & size of samples
- As size of sample increases, size of error decreases
- The greater the variability, the greater the error

Probability Axioms

- Fall between 0% & 100%
- No negative probabilities
- Probability of an event is 100% less the probability of the opposite event

Definitions of Probability

- Frequency Probability based on number of times an event occurred in a given sample (n)

$$\frac{\text{\# of times event occurred}}{\text{total \# of people in n}} \times 100$$

- *P* .... Probability value that observed data are consistent with null hypothesis

Definitions of Probability

- Subjective Probability:percentage expressing personal, subjective belief that event will occur
- p values of .05, often used as a probability cutoff in hypothesis-testing to indicate something unusual happening in the distribution

Definitions of Probability

- p values of .05, often used as a probability cutoff in hypothesis-testing to indicate something unusual happening in the distribution

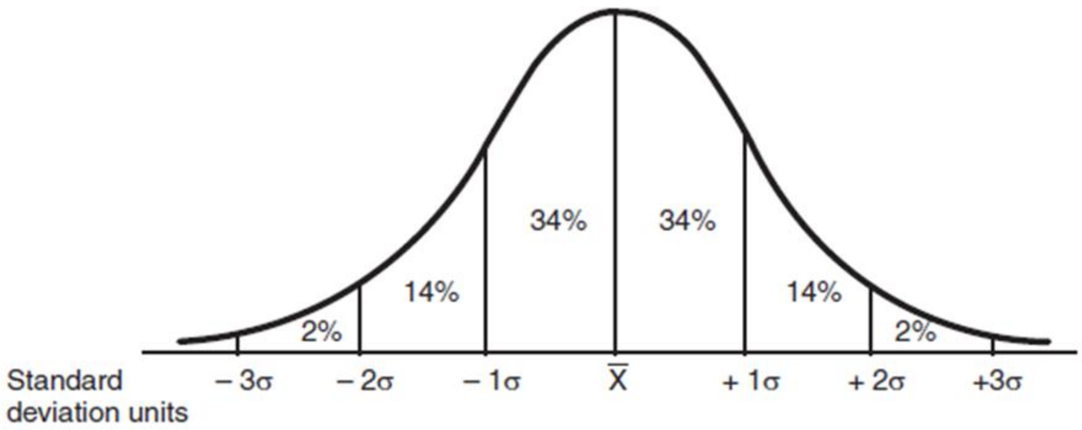


FIGURE 3-1 Normal distribution with Standard deviation units.

Hypothesis

- Tentative prediction of the relationship between two or more variables.
- A statement of predicted relationship between two or more variabales

Purposes of Hypotheses

- Objectivity of scientific investigation by pinpointing specific part of a theory to be tested
- Theoretical propositions tested in *reality*, and then advance knowledge by supporting or failing to support theory would be gained.
- Even when research hypothesis not supported, knowledge gained and would be a guide for further revision of theory
- Hypothesis provide reader with an understanding of researcher's expectations about the study before data collected.

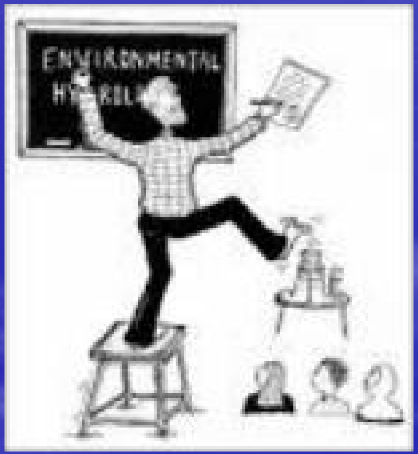
Sources of study hypotheses

- Personal experience
- Previous research studies
- Theories

  - propositions

  - empirical generalization

  - CF



Classification of hypotheses

- *<u>Simple</u>*: concerns the R/S b/w one independent and one dependent variables
- *<u>Complex</u>*: concerns the R/S b/w 2 or more independents and 2 or more dependent variables.

Simple or complex H_o

- Subjects receiving antiemetic therapy by a patient-controlled pump will report

  less nausea than subjects receiving the therapy by nurse administration.

- The greater the degree of sleep deprivation, the higher the anxiety level of ICU patients

- The quality of life of a family during the survivor phase after cancer diagnosis is affected by family resources and illness survival stressors such as fear of recurrence and the family meaning of the illness

- Among breast cancer survivors, emotional wellbeing is influenced by the women’s self-esteem in terms of their resourcefulness and their degree of social support.

Null and Research Hypotheses

- Null form: H_0
  - predicts that no relationship exists between variables
  - indicated for statistical logic

- Research Hypotheses
  H_1 (statistical-scientific-substantive-theoretical-alternative)

Null or Research Hypotheses

- There is no relationship between smoking and lung cancer among Jordanian male nurses.
- There is no difference in student-faculty interaction between students who have good GPA scores compared to students who have low GPA scores.
- Subjects receiving antiemetic therapy by a patient-controlled pump will vomit less than subjects receiving the therapy by nurse administration.

Directional and Non directional hypotheses

- Directional: the researcher predicts the type of relationship. It makes clear expectation and allow more precise testing of theoretical propositions.

- Non directional: the researcher merely predicts that a relationship exists.

Directional or Non directional hypotheses

- *Patients receiving a warmed solution for body cavity irrigation during surgical procedures will maintain a higher core body temperature than patients receiving a room temperature solution.*
- *There is a difference in survival rate among ICU patients who have received CPR by nurses attended ACLS course compared to those patients who received CPR by nurses who were not attended ACLS course*
- *The implementation of an evidence-based protocol for urinary incontinence will result in decreased frequency of urinary incontinence episodes*

- ***Sociodemographic vulnerability (less than high school education, etc.) will be associated with emotional distress in both men and women.***

- ***Pregnant women who receive prenatal instruction regarding postpartum experiences are not likely to experience postpartum depression.***

- ***There is a correlation between smoking, weight, gender and occurrence of coronary artery syndrome among Jordanian male clients***

Hypothesis-Testing

- Prominent feature of quantitative research
- Hypotheses originate from theory that underpins research
- Two types of hypotheses:
  - Null H_o
  - Alternative

Null Hypothesis - H_o

- H_o proposes no difference or relationship exists between the variables of interest
- Foundation of the statistical test
- When you statistically test an hypothesis, you assume that H_o correctly describes the state of affairs between the variables of interest

Null Hypothesis - H_o

- If a statistically significant relationship is found ($p \leq .05$), H_o is rejected

- If no statistically significant relationship is found ($p. \geq .05$), H_o is accepted

Alternative Hypothesis - H_a

- A hypothesis that contradicts H_o
- Can indicate the direction of the difference or relationship expected
- Often called the research hypothesis & represented by H_r

Sampling Error

- Inferences from samples to populations are always probabilistic, meaning we can never be certain our inference was correct
- Drawing the wrong conclusion is called an **error of inference,** defined in terms of H_o as Type I and Type II

Types of Errors

- We summarize these in a 2x2 box:

Decision	H_0 True	H_0 False
Accept H_0	**Right decision** α = ***significance***	**Wrong decision** $1-\beta$ = **type II error**
Reject H_0	**Wrong decision** **$1-\alpha$ = type I error**	**Right decision** B = ***power***

Significance Level

- States risk of rejecting H_o when it is true
- Commonly called p value
  - Ranges from 0.00 - 1.00
  - Summarizes the evidence in the data about H_o
  - Small p value of .001 provides strong evidence against H_o, indicating that getting such a result might occur 1 out of 1,000 times

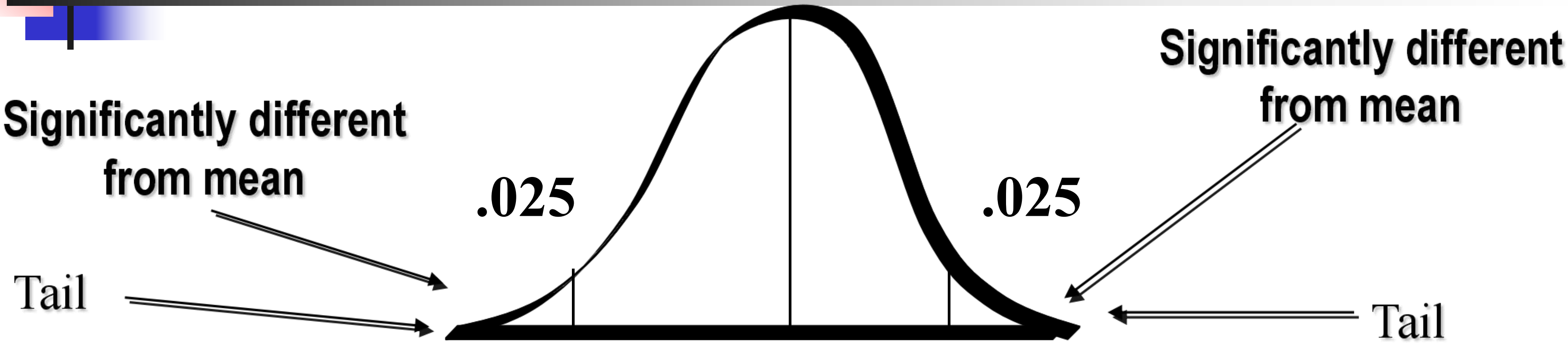
Testing a Statistical Hypothesis

- State H_o
- Choose appropriate statistic to test H_o
- Define degree of risk of incorrectly concluding H_o is false when it is true
- Calculate statistic from a set of randomly selected observations
- Decide whether to accept or reject H_o based on sample statistic

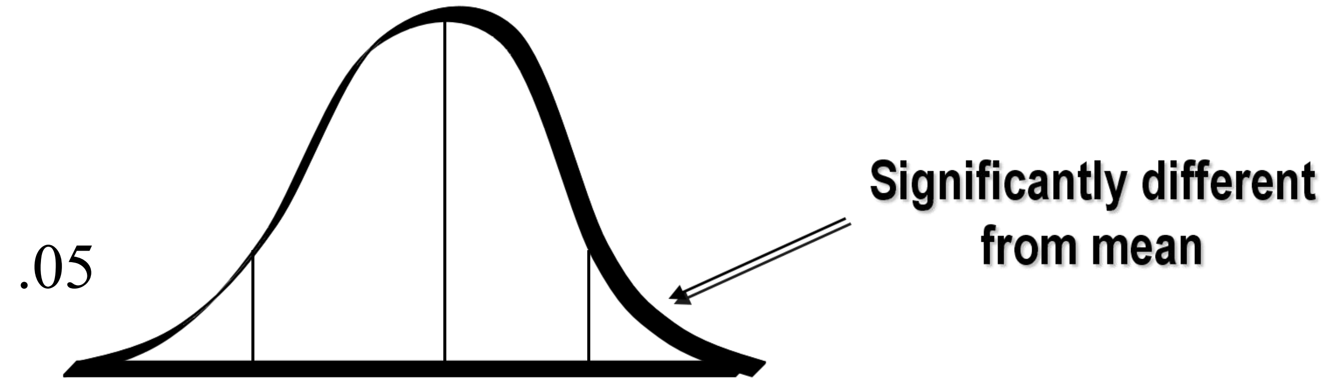
One-Tailed & Two-Tailed Tests

- Tails refer to ends of normal curve
- When we hypothesize the direction of the difference or relationship, we state in what tail of the distribution we expect to find the difference or relationship
- One-tailed test is more powerful & is used when we have a directional hypothesis

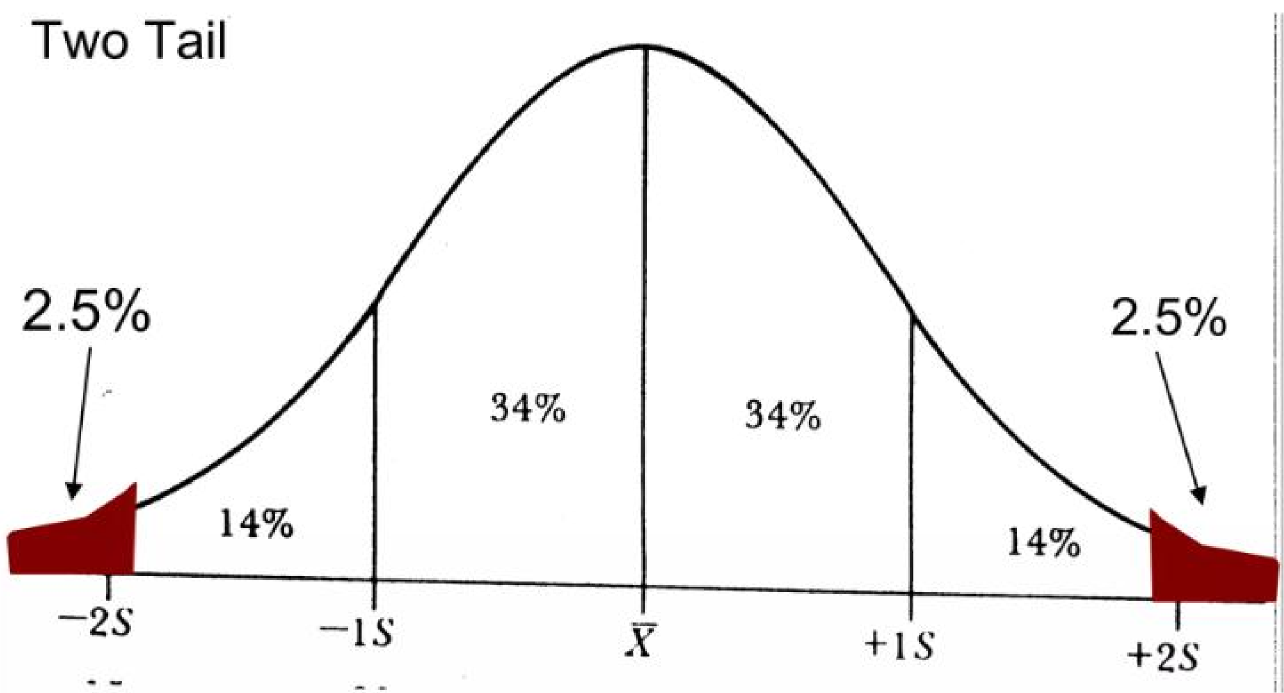
Tailedness



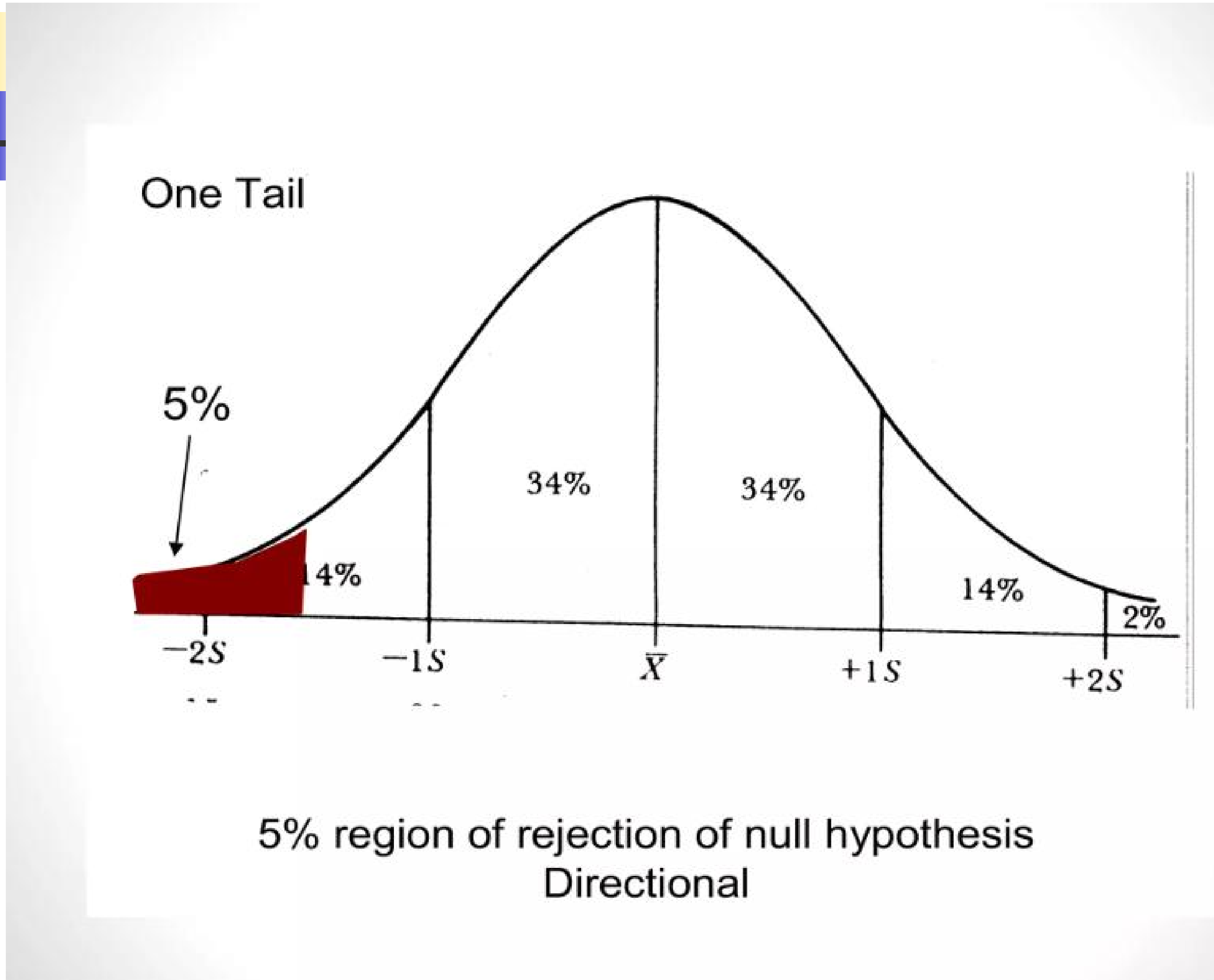
Two-Tailed Test- .05 Level of Significance



One-Tailed Test- .05 Level of Significance



5% region of rejection of null hypothesis
Non directional


One Tail
5%
34%
34%
14%
14%
2%
−2S
−1S
X̄
+1S
+2S
5% region of rejection of null hypothesis
Directional

Degrees of Freedom (df)

- The freedom of a score's value to vary given what is known about other & the sum of the scores
  - Ex. Given three scores, we have 3 df, one for each independent item. Once you know mean, we lose one df
  - df = n - 1, the number of items in set less 1
- Df (degrees of freedom): the extent to which values are free to vary in a given specific number of subjects and a total score

Relationship Between Confidence Interval & Significance Levels

- 95% CI contains all the (H_o) values for which $p \geq .05$
- Makes it possible to uncover inconsistencies in research reports
- A value for H_o within the 95% CI should have a p value > .05, & one outside of the 95% CI should have a p value less than .05

Statistical Significance VS Meaningful Significance

- Common mistake is to confuse statistical significance with substantive meaningfulness
- Statistically significant result simply means that if H_o were true, the observed results would be very unusual
- With $N \geq 100$, even tiny relationships/differences are statistically significant

Statistical Significance VS Meaningful Significance

- Statistically significant results say nothing about clinical importance or meaningful significance of results
- Researcher must always determine if statistically significant results are substantively meaningful.
- Refrain from statistical “sanctification” of data

Hypothesis testing procedure

- State statistical Hypothesis to be tested
- Choose an appropriate statistics to test Null Hypothesis
- Define degree of risk of Type I error (α)
- Calculate statistics from randomly sampled observations
- Decide upon P value less or more than α to accept or reject null Hypotehsis


KEEP
CALM
AND
READ
AGAIN