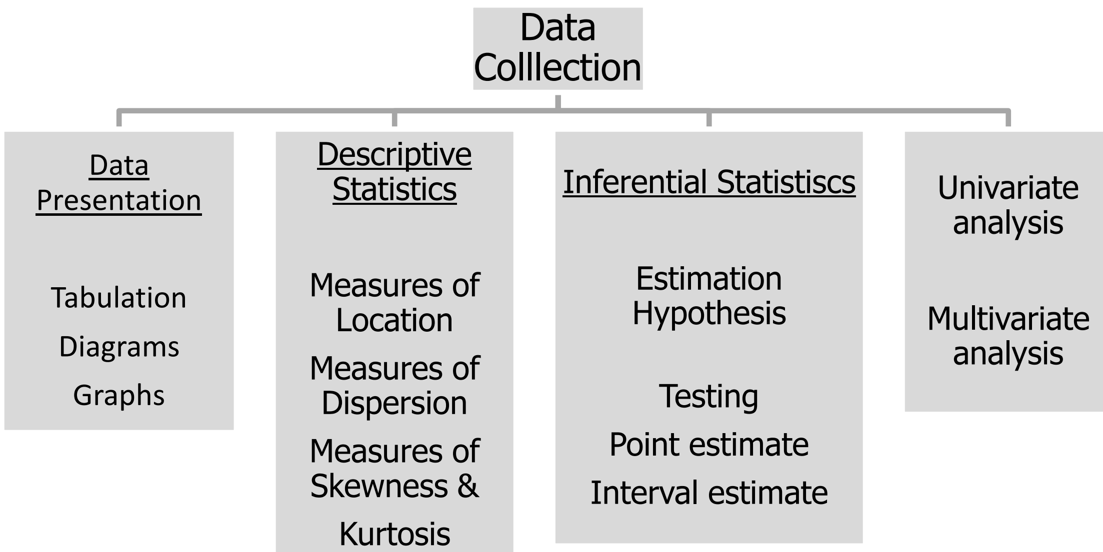
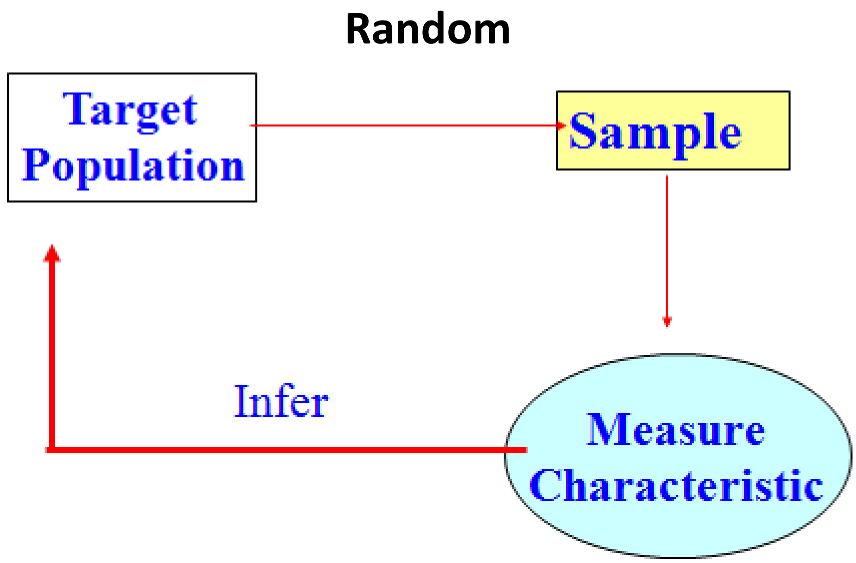
HYPOTHESIS TESTING

INVESTIGATION



- " STATISTICAL INFERENCE IS THE PROCESS OF USING SAMPLES TO MAKE INFERENCES ABOUT THE POPULATION"
    - Parameter Estimation
    - Hypothesis Testing (Test of Significance)

- " An analytic technique for drawing conclusions about the population from an appropriately collected sample "



Concept of test of significance

- A test of significance is a procedure used to obtain answer, on the basis of information from sample observation, to a question of the hypothetical value of the universal parameter
- Hypothesis testing requires the formulation of two opposing hypotheses about the population of interest.
- Data from random samples are used to determine which of the opposing hypothesis is more likely to be correct.

<u>Example:</u>

A current area of research interest is the familial aggregation of cardio vascular risk factors in general and lipid levels in particular.

Suppose we know that the 'average' cholesterol level in children is 175 mg/dl. We identify a group of men who have died from heart disease within the past year and the cholesterol level of their children.

We would like to consider two hypotheses:

1. The average cholesterol level of these children is 175mg/dl. (Null hypothesis 'Ho').
2. The average cholesterol level of these children is greater than 175ml/dl. (Alternative hypothesis 'H_1')

Suppose that the mean cholesterol level of the children in the sample is 180 mg/dl. We need to determine the probability of observing a mean of 180 mg/dl or higher under the assumption that the 'Ho' is true.

If the corresponding probability is considered small enough then we conclude that this is an unlikely finding and therefore that the null hypothesis is unlikely to be true – that is we ‘reject’ the ‘Ho’ in favor of the ‘H1’.

Another way of saying, that the mean is larger than we expect and that this difference has not occurred by chance, so there is evidence that the mean cholesterol is larger than 175 mg/dl.

“ THIS IS THE ESSENCE OF HYPOTHESIS TESTING”

STEPS TAKEN IN HYPOTHESIS TESTING

Testing hypothesis depends on the following logical procedure:

The logic consists of six steps:

1. Generate the Ho and Ha.
2. Generate the sampling distribution of test statistic
3. Check the assumptions of the statistical procedure
4. Set the significance level and formulate the decision rule.
5. Compute the test statistic
6. Apply the decision rule and draw conclusion

Hypothesis

Null hypothesis :

‘The mean sodium concentrations in the two populations are equal.’

Alternative hypothesis:

Logical alternative to the null hypothesis :

‘The mean sodium concentrations in the two populations are different.’

simple, specific, in advance

Null Hypothesis

- "Innocent until proven guilty"
- Null hypothesis (Ho) usually states that no difference between test groups really exists
- Fundamental concept in research is the concept of either "rejecting" or "conceding" the Ho
- State the Ho: An investigator states that a new therapy is similar to the current therapy

Null Hypothesis (Ho): Courtroom Analogy

- The null hypothesis is that the defendant is innocent.
- The alternative is that the defendant is guilty.
- If the jury acquits the defendant, this does not mean that it accepts the defendant's claim of innocence.
- It merely means that innocence is believable because guilt has not been established beyond a reasonable doubt.

Null hypothesis

H_0: The two treatments (or 2 groups) **are not different .**

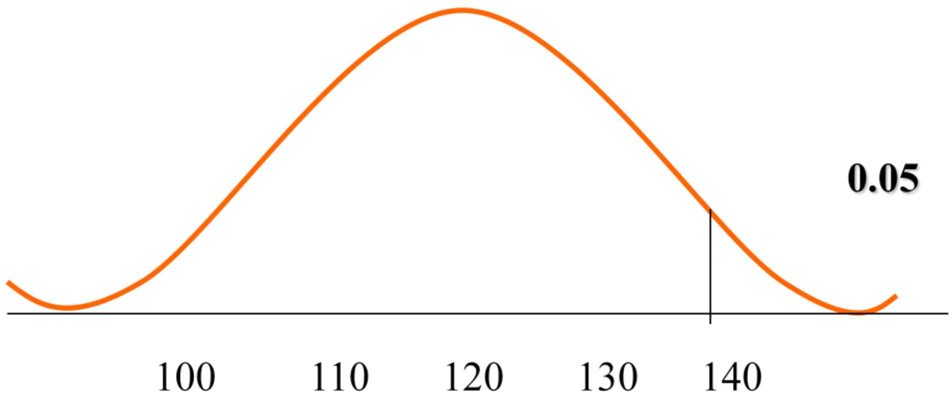
Alternative hypothesis

H_A or H_1: The two treatments (or 2 groups) **are different .**

One-tail test

Ho:μ= μo
Ha: μ> μo or μ< μo

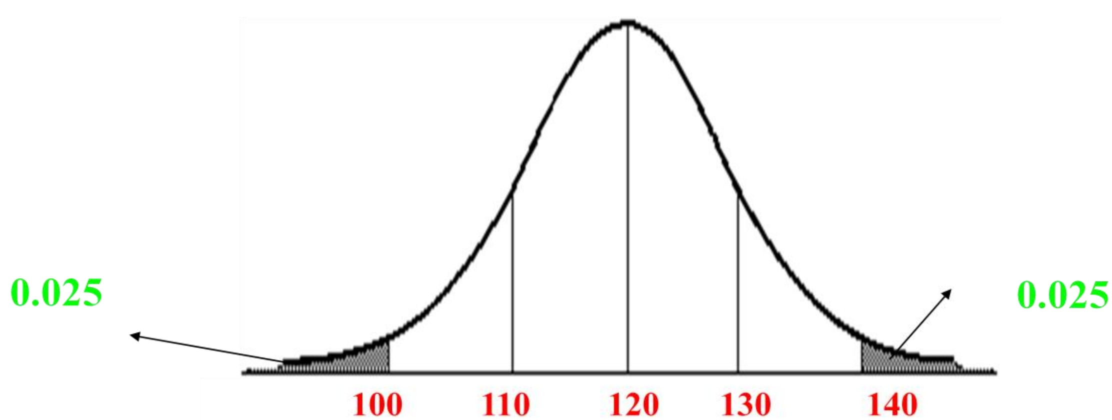
<u>Alternative Hypothesis:</u> Mean systolic BP of Nephrology patients is significantly **higher (or lower)** than the mean systolic BP of normal patients.



Two-tail test

Ho:μ= μo
Ha:μ# μo

<u>Alternative Hypothesis :</u> Mean systolic BP of Nephrology patients are significantly **different** from mean systolic BP of normal patients.



ONE vs TWO SIDED HYPOTHESIS

- If you don't know which therapy or test will yield lower values you have a two sided hypothesis.
- If you know that one must by biologic principles be lower then it will be a one sided hypothesis.
- A trail of a cholesterol lower drug versus placebo. (It won't raise cholesterol).

TYPE I & TYPE II ERRORS

- Every decisions making process will commit two types of errors.
- We may conclude that the difference is significant when in fact there is not real difference in the population, and so reject the null hypothesis when it is true. This is error is known as type-I error, whose magnitude is denoted by the Greek letter 'α'.
- On the other hand, we may conclude that the difference is not significant, when in fact there is real difference between the populations, that is the null hypothesis is not rejected when actually it is false. This error is called type-II error, whose magnitude is denoted by 'β'.

Type I and Type II Errors

	Actual Situation	
Decision of Court	***Guilty***	**Innocent**
Guilty	**Correct decision**	**Type I error**
Innocent	**Type II error**	**Correct decision**

Ho: Defendant is innocent
Ha: Defendant is guilty

Here Type-I is more important is more serious than Type-II error

Diagnostic Test situation

Disease (Gold Standard)

Test Result		Present	Absent	Total
	Positive	Correct A	False Positive B	a + b
	Negative	False Negative C	Correct D	c + d
	Total	a + c	b + d	a + b + c + d

Type I and Type II Errors

Conclusion	Actual Situation H_o False	Actual Situation H_o True
Reject H_o	Correct decision	Type I error
Accept H_o	Type II error	Correct decision

α = probability of rejecting the H_o when H_o is true (Type I error)

β = probability of failing to rejecting the H_o when H_o is false (Type II error)

- Alpha = probability of Type I error = α
  - ✓ Significance level; 1 - α is the confidence level
  - ✓ Probability of rejecting a true null hypothesis

- Beta = probability of Type II error = β
  - ✓ Probability of not rejecting a false null hypothesis
  - ✓ Probability of not detecting a real difference
  - ✓ 1 - β is the power of the test

- p-value = posterior significance level

Examples:

1. In treating 'TB' there are lot of drugs available with not much of difference (ie., similar efficacy) hence Type-I error is important.

2. In treating 'CANCER' very few drugs are available with different efficacy rates, hence Type-II error is important.

Extrapolation of Research Findings

- Sample vs. Population
- If your study shows that treatment A is better than treatment B
  - ✓ You cannot conclude that treatment A is ALWAYS better than treatment B
  - ✓ You only sampled a small portion of the entire population, so there is always a chance that your observation was a chance event
- With any research study, there is a possibility that the observed differences were a chance event
- The only way to know that a difference is really present with certainty, the entire population would need to be studied
- The research community and statisticians had to pick a level of uncertainty at which they could live

- At what point are we comfortable concluding that there is a difference between the groups in our sample
- In other words, what is the false-positive rate that we are willing to accept
- What is this called in statistical terms?

Definition of p-value

- This level of uncertainty is called type 1 error or a false-positive rate (α)
- More commonly called a p-value
- In general, $p \leq 0.05$ is the agreed upon level
- In other words, the probability that the difference that we observed in our sample occurred by chance is less than 5%
    - Therefore we can reject the Ho

- **Stating the Conclusions of our Results**
    - When the p-value is small, we reject the null hypothesis or, equivalently, we accept the alternative hypothesis.
        - "Small" is defined as a p-value $\leq$ a, where a = acceptable false (+) rate (usually 0.05).
    - When the p-value is not small, we conclude that we cannot reject the null hypothesis or, equivalently, there is not enough evidence to reject the null hypothesis.
        - "Not small" is defined as a p-value > a, where a = acceptable false (+) rate (usually 0.05).

P-value

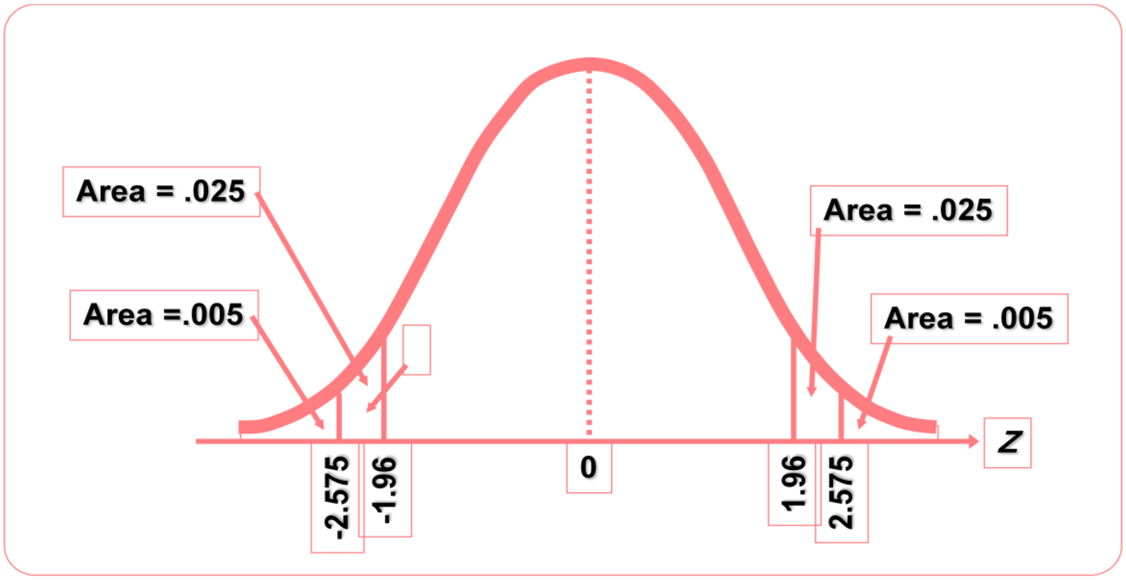
- A standard device for reporting quantitative results in research where variability plays a large role.
- Measures the dissimilarity between two or more sets of measures or between one set of measurements and a standard.
- " the probability of obtaining the study results by chance if the null hypothesis is true"
- "The probability of obtaining the observed value (study results) as extreme as possible"
- " The p-value is actually a probability, normally the probability of getting a result as extreme as or more extreme than the one observed if the dissimilarity is entirely due to variability of measurements or patients response, or to sum up, due to chance alone".
  - ✓ **Small p value → the rare event has occurred**
  - ✓ **Large p value → likely event**

p value   - 0.05

- **Advantages:**
  - ✓ It gives specific level to keep in mind, objectively chosen
  - ✓ It may be easier to say whether a p-value is smaller or larger than 0.05 than to compute the exact probability

- **Disadvantages**
  - ✓ It suggest a rather mindless cutoff point having nothing to do with the importance of the decision or the costs and losses associated with the outcomes.
  - ✓ Reporting of 'greater than' or 'less than' 0.05 is not as informative as reporting the actual level.

Determining the p-Value



Application of Test of Significance

1. To test sample proportion is equal to population proportion
   H0: p=P HIV; Diabetes; Hypertension; Anemia

2. To test whether proportion of sample I is equal to proportion of sample II
   H0: p1=p2 Sex, State, Disease wise

3. Test sample mean is equal to predefined (population) mean
   H0: x = µ Hb, Creatine, Cholesterol.,

4. Test whether a mean of a sample I is equal to the mean of the sample II
   H0:x1=x2

5. Test whether post treatment observation is significantly higher than pre treatment observation :
   H0: No change ↑ Hb; ↓ ESR

6. To find association between two categorical variables
   Smoking → Lung cancer
   Alcohol → Liver disease
   Genital Ulcer → HIV

Test for single prop. with population prop.

- **<u>Problem</u>**

In an otological examination of school children, out of 146 children examined 21 were found to have some type of otological abnormalities. Does it confirm with the statement that 20% of the school children have otological abnormalities?

a . Question to be answered:

Is the sample taken from a population of children with 20% otological abnormality

b. Null hypothesis :

The sample has come from a population with 20% otological abnormal children

c. Test statistics

$$z = \frac{p - P}{\sqrt{\frac{pq}{n}}} = \frac{14.4 - 20.0}{\sqrt{\frac{14.4 * 85.6}{146}}} = 1.69$$

P – Population. Prop.
p- sample prop.
n- number of samples

d.Comparison with theoritical value :

$Z \sim N(0,1)$; $Z_{0.05} = 1.96$

The prob. of observing a value equal to or greater than 1.69 by chance is more than 5%. We therefore do not reject the Null Hypothesis.

e. Inference :

There is no evidence to show that the sample is not taken from a population of children with 20% abnormalities.

Comparison of two sample proportions

- **Problem :**

In a hearing survey among 246 town school children, 36 were found with conductive hearing loss and among 349 village school children 61 were found with conductive hearing loss. Does this present any evidence that conductive hearing loss is as common among town children as among village children?

a. Question to be answered:

Is there any difference in the proportion of hearing loss between children living in town and village?

Given Data	Sample 1	Sample 2
Size	246	342
Hearing loss	36	61
% Hearing Loss	14.6%	17.5%

b. Null Hypothesis :

There is no difference between the proportions of conductive hearing loss cases among the town children and among the village children

c. Test statistics **q= 1- p**

$$z = \frac{p1 - p2}{\sqrt{\frac{p1q1}{n1} + \frac{p2q2}{n2}}} = \frac{14.6 - 17.5}{\sqrt{\frac{14.4 * 85.6}{246} + \frac{17.5 * 82.5}{342}}} = 1.81$$

p1, p2 are sample proportions,

n1,n2 are subjects in sample 1 & 2

d. Comparison with theoretical value

Z ~ N (0,1); Z 0.05 = 1.96

The prob. of observing a value equal to or greater than 1.81 by chance is more than 5%. We therefore do not reject the Null Hypothesis

e. Inference

There is no evidence to show that the two sample proportions are statistically significantly different.

Example : Weight Loss for Diet vs Exercise

Did dieters lose more fat than the exercisers?

Diet Only:

sample mean = 5.9 kg
sample standard deviation = 4.1 kg
sample size = *n* = 42
standard error = SEM_1 = 4.1/ Ö42 = 0.633

Exercise Only:

sample mean = 4.1 kg
sample standard deviation = 3.7 kg
sample size = *n* = 47
standard error = SEM_2 = 3.7/ Ö47 = 0.540
measure of variability = $[(0.633)^2 + (0.540)^2] = 0.83$

Step 1. Determine the null and alternative hypotheses.

Null hypothesis: No difference in average fat lost in population for two methods. Population mean difference is *zero*.

Alternative hypothesis: There is a difference in average fat lost in population for two methods. Population mean difference is not *zero*.

Step 2. Sampling distribution: Normal distribution (z-test)

Step 3. Assumptions of test statistic (sample size > 30 in each group)

Step 4. Collect and summarize data into a test statistic.

The sample mean difference = 5.9 – 4.1 = 1.8 kg and the standard error of the difference is 0.83.

So the *test statistic:* $z = 1.8 - 0 / 0.83 = 2.17$

Step 5. Determine the p-value.

Recall the alternative hypothesis was two-sided.

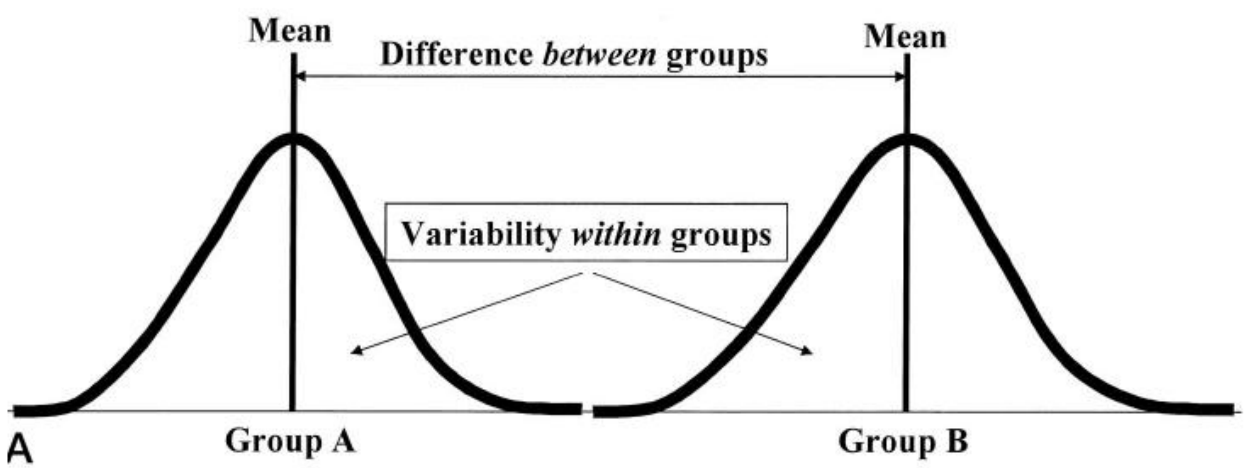
- ✓ *p*-value = 2 × [proportion of bell-shaped curve above 2.17]
- ✓ Z-test table => proportion is about 2 × 0.015 = 0.03.

Step 6. Make a decision.

The *p*-value of 0.03 is less than or equal to 0.05, so …

- ✓ If really no difference between dieting and exercise as fat loss methods, would see such an extreme result only 3% of the time, or 3 times out of 100.
- ✓ Prefer to believe truth does not lie with null hypothesis. We conclude that there is a *statistically significant difference between average fat loss for the two methods*.

Two Sample Tests: Continuous Variable



Inference based on Hypothesis

- If the null hypothesis is rejected
  - ✓ conclude that there is a statistically significant difference between the treatments
  - ✓ the difference is not due to chance
- If the null hypothesis is not rejected
  - ✓ conclude that there is not a statistically significant difference between the treatments
  - ✓ any observed difference may be due to chance
  - ✓ the difference is not necessarily negligible
  - ✓ the groups are not necessarily the same

STATISTICALLY SIGNIFICANT AND NOT STATISTICALLY SINGIFICANT

Statistically significant	**Not statistically significant**
Reject Ho	Do not reject Ho
Sample value not compatible with Ho	Sample value compatible with Ho
Sampling variation is an unlikely explanation of discrepancy between Ho and sample value	Sampling variation is an likely explanation of discrepancy between Ho and sample value

Points of clarification

- There are two approaches to reporting the decision made on the basis of test of significance.
  1. One approach is to fix a level of significance which we denote by α (0.05) and define the rejection regions (or tails of the distribution) which include an area of size α according to whether the test is one-sided or two-sided . If the test statistics falls in these regions the null hypothesis is rejected otherwise we fail to reject the Ho.

2. The other approach is to calculate the p-value or probability corresponding to the observed value of the test statistic and use this p-value as a measure of evidence in favor of Ho. The p-value is defined as the probability of getting a value of extreme or more extreme than that observed in the same direction (for a one-sided) or in either direction (for a two-sided test).

STATISTICAL SIGNIFICANCE

Vs

MEDICAL/CLINICAL/BIOLOGICAL SIGNIFICANCE

- In Hypothesis testing we concerned about minimizing the probability of making type-I error (rejecting Ho when in fact it is true), since we concerned the size of 'α' and formulate the decision rule for rejecting Ho.
- Sample size 'n' (or n1 and n2) occurs in the denominator of the standard error of the sample statistic of interest that the larger the sample size, the smaller the S.Error and so the larger the test statistic regardless of the size of the numerator (the difference between the sample estimate and hypothesized value).
- Thus it follows that one could reject Ho for even very small differences if the sample size is large.
- This leads to the consideration of a difference between treatment effects or differences between the observed and hypothesized values that are clinically important as well as statistically significant.

Example:

- ✓ If a new antihypertensive therapy reduced the SBP by 1mmHg as compared to standard therapy we are not interested in swapping to the new therapy.
- ✓ However, if the decrease was as large as 10 mmHg, then you would be interested in the new therapy.
- ✓ Thus, it is important to not only consider whether the difference is statistically significant by the possible magnitude of the difference should also be considered.

Statistical Significance Versus Clinical Importance

- **Statistical significance**
  - The difference is real (not due to chance)
- **Clinical (practical) importance**
  - The difference is important or large

Statistically Significant	Clinically Important	Result
Yes	Yes	Good Study
No	No	-------
Yes	No	Too Many Subjects
No	Yes	Not Enough Subjects

Home MESSAGE

- VARIABILITY EXISTS IN EVERY ASPECT OF DATA. AND IN ANY COMPARISON MADE IN A CLINICAL CONTEXT, DIFFERENCES ARE ALMOST BOUND TO OCCUR. THESE DIFFERENCES MAY BE DUE TO REAL EFFECTS, RANDOM VARIATION OR BOTH. IT IS THE JOB OF ANALYST TO DECIDE HOW MUCH VARIATION SHOULD BE ASCRIBED TO CHANCE, SO THAT ANY REMAINING VARIATION CAN BE ASSUMED TO BE DUE TO A REAL EFFECT.
- **"THIS IS THE ART OF INFERENTIAL STATISTICS"**

Drawbacks to Hypothesis Tests

- Statistical significance is **NOT** the same as practical importance!
- Hypothesis test
  - ✓ only addresses Type I Error
  - ✓ answers the question: "Is the observed difference between treatments due only to chance?"
- Confidence interval
  - ✓ contains all treatment differences that would not be rejected by the corresponding hypothesis test
  - ✓ addresses both Type I and II Error
  - ✓ answers the question: "Is the size of that difference of any clinical (practical) importance?