HCI: Analysis of Empirical Data

Learning Objective

- In the previous lectures, we discussed the basics of HCI - Empirical Research Methods
  - We discussed about the three themes, namely: Research Question Formulation, Observation and Measurement, and User Study (Experiment Design)
- In this lecture, we shall mainly focus on the Analysis of Empirical Data

Learning Objective

- In particular, we shall learn about the following:
  - The case for Statistical Analysis of Observed Data
  - Introduction to one of the commonly and widely used Statistical Analysis Techniques, namely: one-way ANalysis Of VAriance (ANOVA) test

Answering Empirical Questions

- Suppose, we want to determine if the text entry speed of a proposed text input system is more than an existing system
- We know how to design an experiment and we also know how to observe and measure
- So, we do the following (details are given in the next slide)

Answering Empirical Questions

- We conduct a User Study and measure the performance on each test condition (our proposed system and the existing system) over a group of participants
- For each test condition we compute the mean score (text entry speed) over the group of participants
- We now have the observed (empirical) data. What next?

Answering Empirical Questions

- Now, we are faced with the following three questions:
  - Is there a difference?

    This is true as we are most likely to see some differences. However, can we conclude anything from this difference? This brings us to the second question.

Answering Empirical Questions

- We are faced with the following three questions:
  - Is the difference too large or too small?

    This is more difficult to answer. If we observe a difference of, say, 30%, we can definitely say the difference is too large. However, we can't say anything definite about, say, a 5% difference. Clearly, the difference figure itself can't help us to draw any definite conclusion. This brings us to the third question.

Answering Empirical Questions

- We are faced with the following three questions:
  - Is the difference significant or is it due to chance?

    Even if the observed difference is “small”, it can still lead us to conclude about our design if we can determine the nature of the difference. If the difference is found to be “significant” (not occurred by chance), then we can say something about our design.

Answering Empirical Questions

- It is important to note that the term "significance" is a statistical term
- The test of (statistical) significance is an important aspect of empirical data analysis
- We can use statistical techniques for this purpose
  - The basic technique is ANOVA or ANalysis Of VAriance

Statistical Hypothesis Testing

In statistics, a Hypothesis is a claim or statement about a property of a population.

A Hypothesis Test (or Test of Significance) is a standard procedure for testing a claim about a property of a population.

Basics of Hypothesis Testing

The following components of a statistical hypothesis test are widely used for carrying out the comprehensive procedures.

- **Null and Alternative Hypotheses**
- **Test Statistics**
- **Critical Region and Critical Values**
- **Significance Levels**
- ***P*-values**
- **Decision Criteria**
- **Type I and II Errors**
- **Power of a Hypothesis Test**

Learning Objectives

- **Given a claim, identify the null hypothesis and the alternative hypothesis, and express them both in symbolic form.**
- **Given a claim and sample data, calculate the value of the test statistic.**
- **Given a significance level, identify the critical value(s).**
- **Given a value of the test statistic, identify the *P*-value.**
- **State the conclusion of a hypothesis test in simple, non-technical terms.**

Example: Let's refer to Gender Choice product that was once distributed by ProCare Industries. ProCare claimed that couples using the pink packages of Gender Choice would have girls at a rate that is greater than 50% or 0.5. Let's again consider an experiment whereby 100 couples use Gender Choice in an attempt to have a baby girl; let's assume that the 100 babies include exactly 52 girls, and let's formalize some of the analysis. Under normal circumstances the proportion of girls is 0.5, so a claim that Gender Choice is effective can be expressed as $p > 0.5$. Using a normal distribution as an approximation to the binomial distribution, we find that P(52 or more girls in 100 births) = 0.3821. Figure 1 (next slide) shows that with a probability of 0.5, the outcome of 52 girls in 100 births is not unusual.

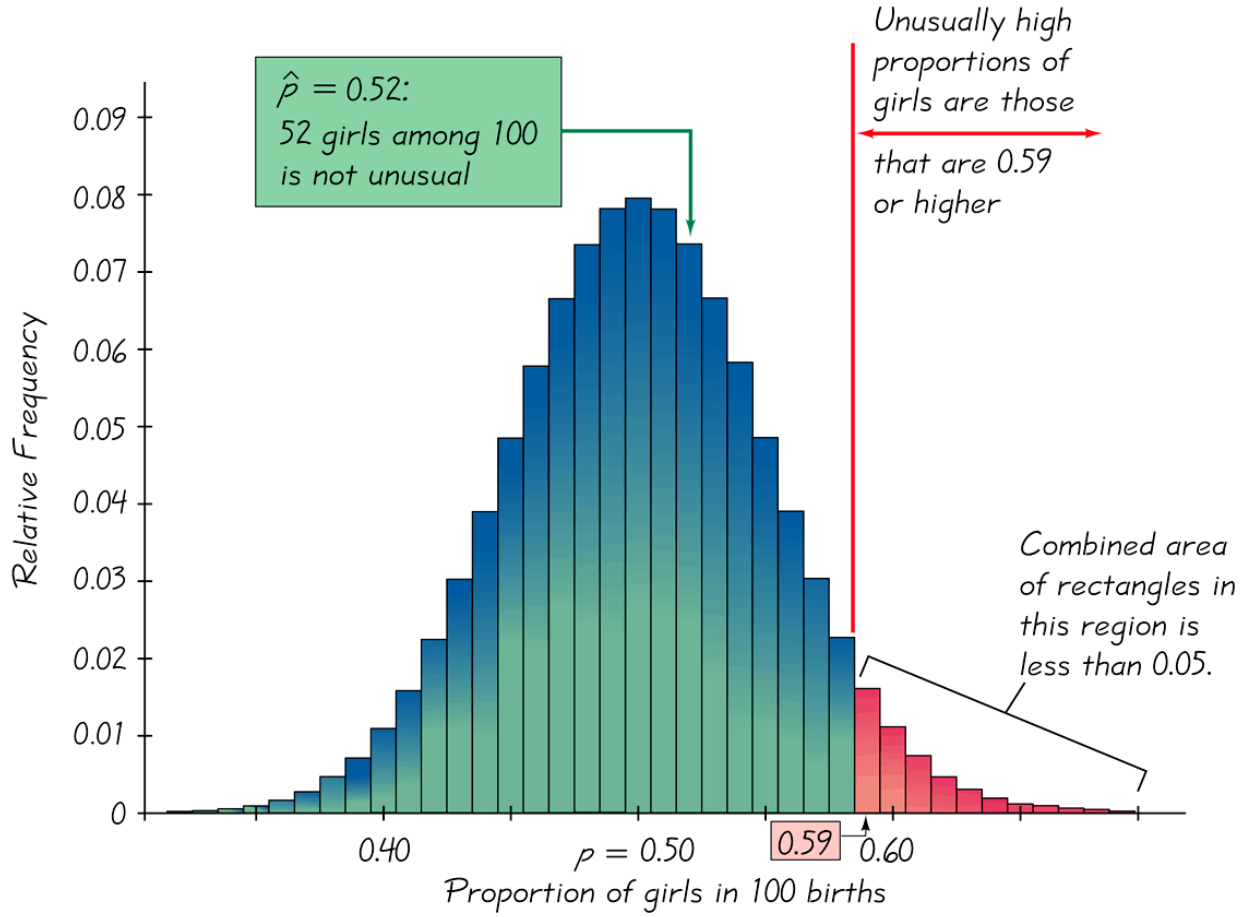


Figure 1

We do not reject random chance as a reasonable explanation. We conclude that the proportion of girls born to couples using Gender Choice is not significantly greater than the number that we would expect by random chance.

Observations

- Claim: For couples using the Gender Choice product, the proportion of girls is $p > 0.5$.
- Working assumption: The proportion of girls is $p = 0.5$ (with no effect from the Gender Choice).
- The sample resulted in 52 girls among 100 births, so the sample proportion is $\hat{p} = 52/100 = 0.52$.
- Assuming that $p = 0.5$, we use a normal distribution as an approximation to the binomial distribution to find that P (at least 52 girls in 100 births) = 0.3821.
- There are two possible explanations for the result of 52 girls in 100 births: Either a random chance event (with probability 0.3821) has occurred, or the proportion of girls born to couples using Gender Choice is greater than 0.5.
- There isn't sufficient evidence to support Gender Choice's claim.

Components of a Formal Statistical Hypothesis Test

Null Hypothesis: H_0

- The **Null Hypothesis** (denoted by H_0) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is **equal to** some claimed value.
- We test the **Null Hypothesis** directly.
- Either reject H_0 or fail to reject (accept) H_0.

Alternative Hypothesis: H_1

- **The Alternative Hypothesis (denoted by H_1 or H_a or H_A) is the statement that the parameter has a value that somehow differs from Null Hypothesis (as defined in the previous slide).**
- **The symbolic form of the alternative hypothesis must use one of these symbols: ≠, <, >.**

Note about Forming Our Own Claims (Hypotheses)

If we are conducting a study and want to use a hypothesis test to support our claim, the claim must be worded so that it becomes the alternative hypothesis.

Note about Identifying H_0 and H_1

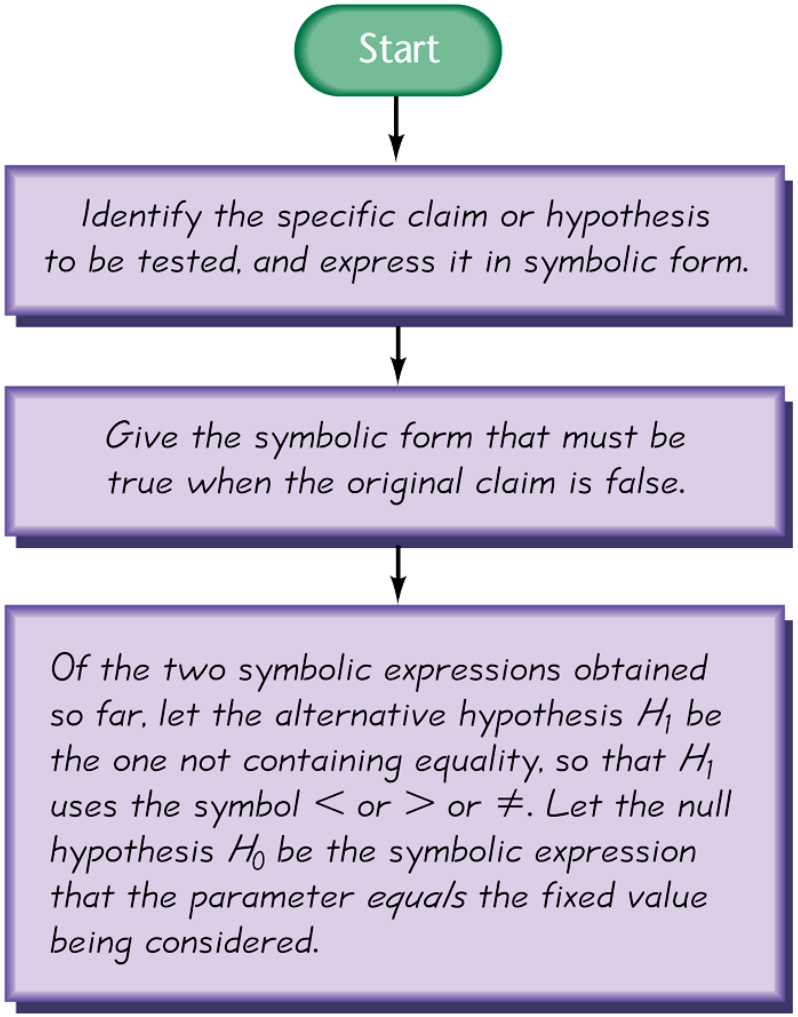


Figure 2

Example: Identify Null and Alternative Hypotheses. Refer to Figure 2 and use the given claims to express the corresponding the null and alternative hypotheses in symbolic form.

a) The proportion of drivers who admit to running red lights is > 0.5.

b) The mean height of professional basketball players is at most 7 ft.

c) The standard deviation of IQ scores of actors is equal to 15.

Example: Identify the Null and Alternative Hypotheses. Refer to the Figure 2 and use the given claims to express the corresponding null and alternative hypotheses in symbolic form.

a) The proportion of drivers who admit to running red lights is greater than 0.5. In Step 1 of Figure 2, we express the given claim as $p > 0.5$. In Step 2, we see that if $p > 0.5$ is false, then $p \leq 0.5$ must be true. In Step 3, we see that the expression $p > 0.5$ does not contain equality, so we let the alternative hypothesis H_1 be $p > 0.5$, and we let H_0 be $p = 0.5$.

Example: Identify the Null and Alternative Hypotheses. Refer to the Figure 2 and use the given claims to express the corresponding null and alternative hypotheses in symbolic form.

b) The mean height of professional basketball players is at most 7 ft. In Step 1 of Figure 2, we express "a mean of at most 7 ft" in symbols as $\mu \leq 7$. In Step 2, we see that if $\mu \leq 7$ is false, then $\mu > 7$ must be true. In Step 3, we see that the expression $\mu > 7$ does not contain equality, so we let the alternative hypothesis H_1 be $\mu > 0.5$, and we let H_0 be $\mu = 7$.

Example: Identify the Null and Alternative Hypotheses. Refer to Figure 2 and use the given claims to express the corresponding null and alternative hypotheses in symbolic form.

c) The standard deviation of IQ scores of actors is equal to 15. In Step 1 of Figure 2, we express the given claim as $\sigma = 15$. In Step 2, we see that if $\sigma = 15$ is false, then $\sigma \neq 15$ must be true. In Step 3, we let the alternative hypothesis H_1 be $\sigma \neq 15$, and we let H_0 be $\sigma = 15$.

Test Statistic

The Test Statistic is a value used in making a decision about the null hypothesis, and is found by converting the sample statistic to a score with the assumption that the null hypothesis is true.

Test Statistic - Formulas

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

Test Statistic for Proportions

$$z = \frac{\bar{x} - \mu_{x^-}}{\frac{\sigma}{\sqrt{n}}}$$

Test Statistic for Mean

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Test Statistic for Standard Deviation

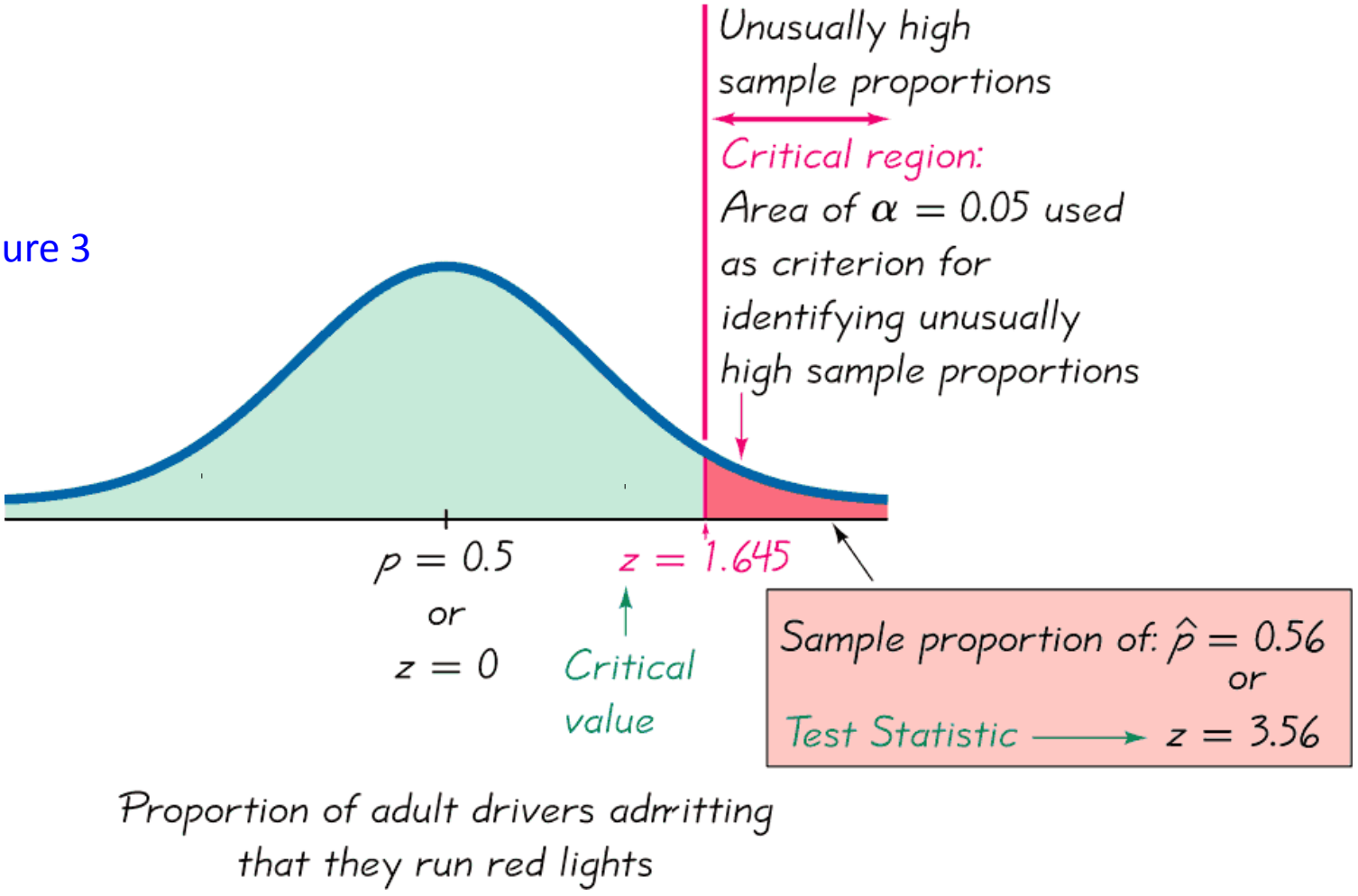
Example: A survey of $n = 880$ randomly selected adult drivers showed that 56% (or $p = 0.56$) of those respondents admitted to running red lights. Find the value of the test statistic for the claim that the majority of all adult drivers admit to running red lights. (Assume that the required assumptions are satisfied and focus on finding the indicated test statistic.)

Solution: The preceding example showed that the given claim results in the following null and alternative hypotheses: H_0: $p = 0.5$ and H_1: $p > 0.5$. Because we work under the assumption that the null hypothesis is true for a value of $p = 0.5$, we get the following test statistic:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} = \frac{0.56 - 0.5}{\sqrt{\frac{(0.5)(0.5)}{880}}} = 3.56$$

Critical Region, Critical Value, Test Statistic

Figure 3



Proportion of adult drivers admitting that they run red lights

Critical Region

The Critical Region (or Rejection Region) is the set of all values of the test statistic that cause us to reject the null hypothesis. For example, see the red-shaded region as shown in Figure 3 (previous slide).

Significance Level

The Significance Level (denoted by α) is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true. Common choices for α are 0.05, 0.01, and 0.10.

Critical Value

A Critical Value is any value that separates the critical region (where we reject the null hypothesis) from the values of the test statistic that do not lead to rejection of the null hypothesis. The critical values depend on the nature of the null hypothesis, the sampling distribution that applies, and the significance level α. See Figure 3 where the critical value of $z = 1.645$ corresponds to a significance level of $\alpha = 0.05$.

Two-tailed, Right-tailed, Left-tailed Tests

The Tails in a distribution are the extreme regions bounded by critical values.

Two-tailed Test

H_0: =

H_1: ≠

Means less than or greater than

α is divided equally between the two tails of the critical region

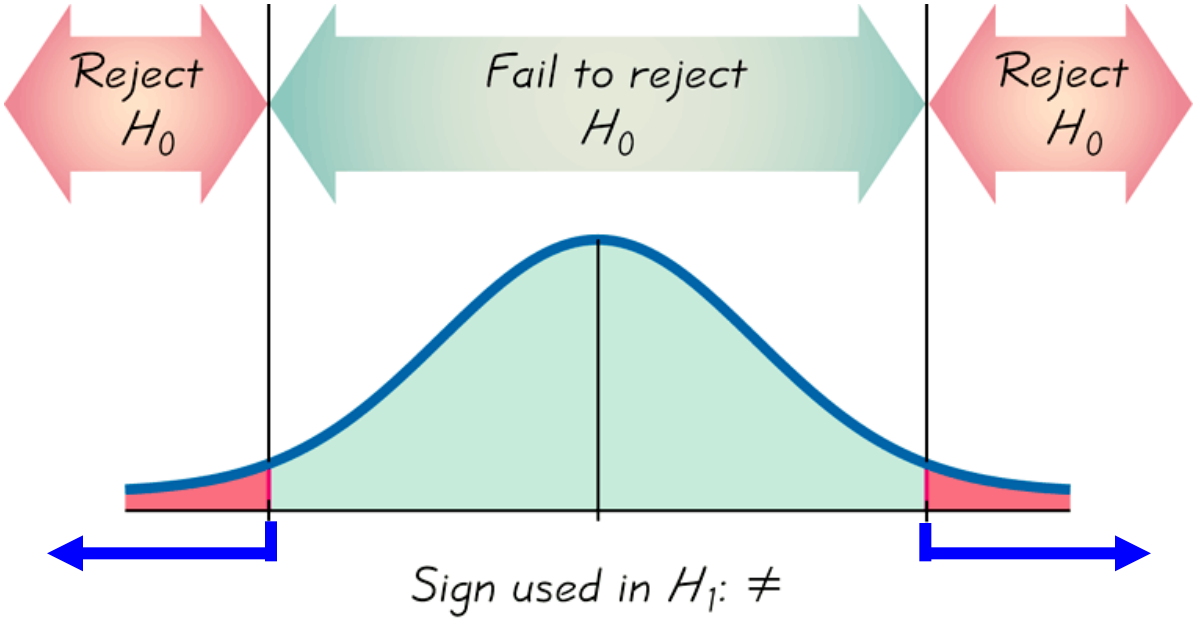


Figure 4

Right-tailed Test

H_0: =

H_1: >

Points Right

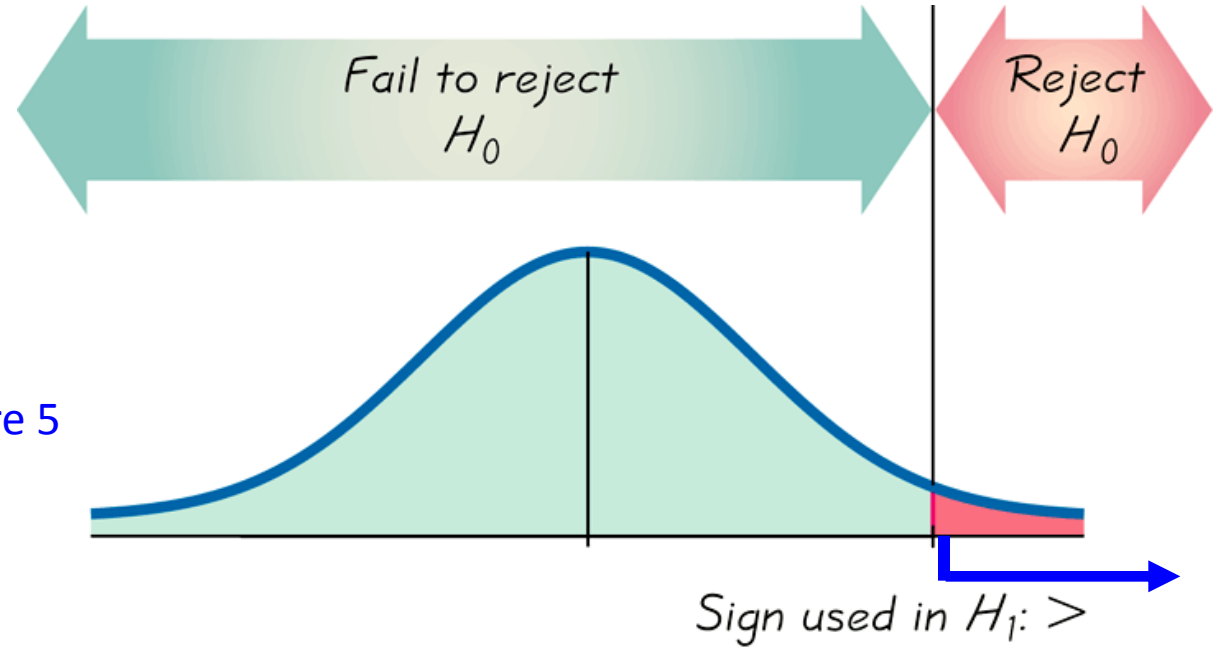


Figure 5

Left-tailed Test

H_0: =

H_1: <

Points Left

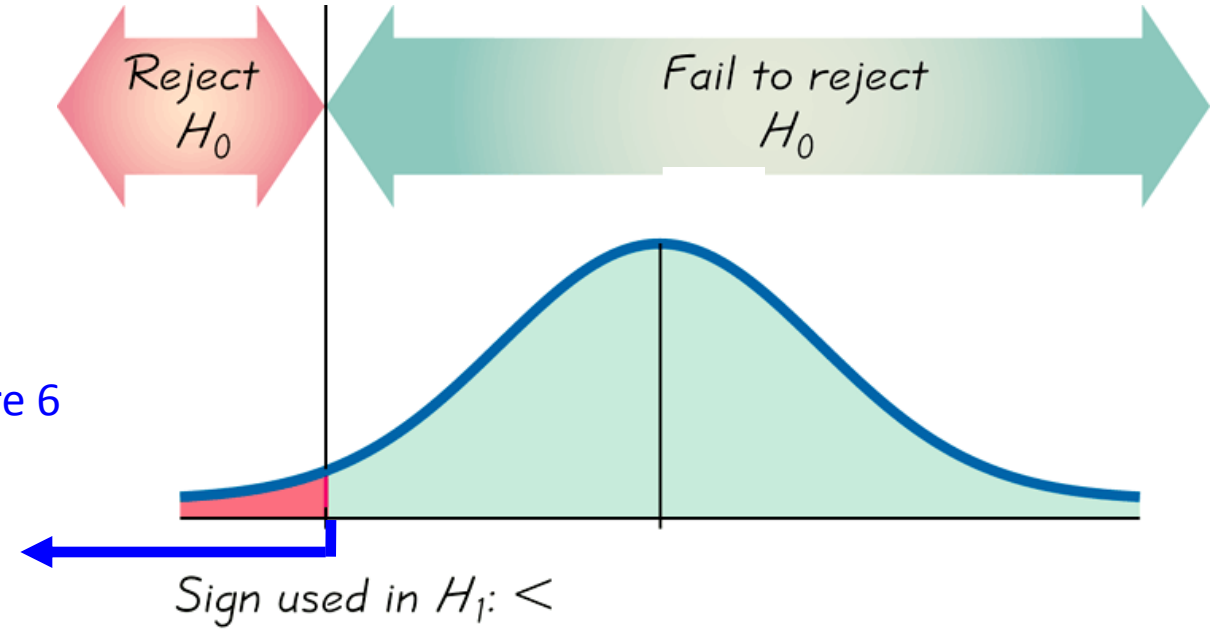


Figure 6

P-Value

The *P*-value (or *p*-value or probability value) is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis is true. The null hypothesis is rejected if the *P*-value is very small, such as 0.05 or less.

Initial Conclusions in Hypothesis Testing

We always test the null hypothesis. The initial conclusion will always be one of the following:

1. Reject the Null Hypothesis.

2. Fail to Reject the Null Hypothesis.

Decision Criterion

Traditional Method:

Reject H_0 if the test statistic falls within the critical region.

Fail to Reject H_0 if the test statistic does not fall within the critical region.

Decision Criterion (Contd…)

P-value Method:

Reject H_0 if the *P*-value $\leq \alpha$ (where α is the significance level, such as 0.05).

Fail to Reject H_0 if the *P*-value > α.

Decision Criterion (Contd…)

Another Option:

Instead of using a significance level such as 0.05, simply identify the *P*-value and leave the decision to the reader.

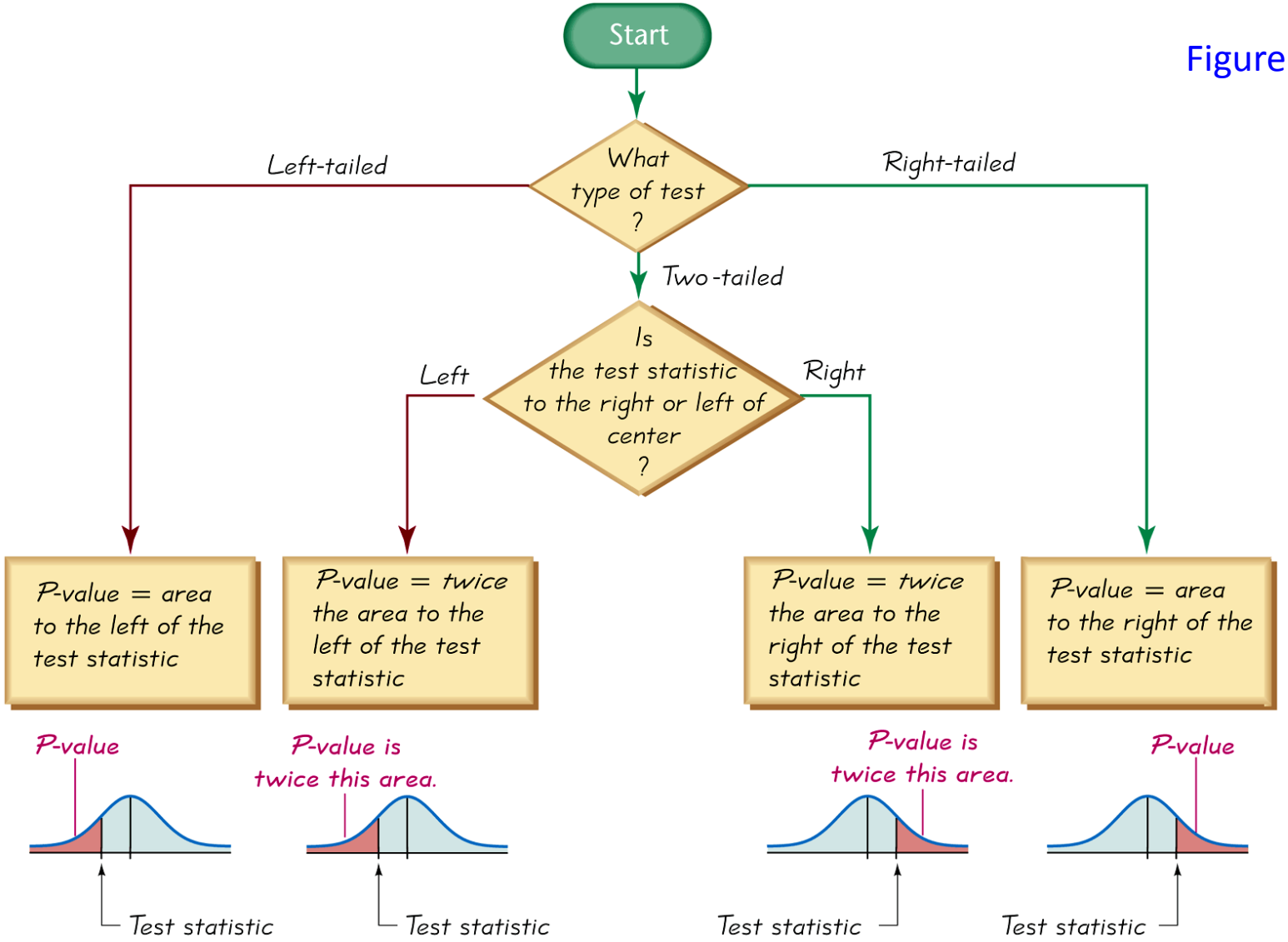
Decision Criterion (Contd…)

Confidence Intervals:

Because a confidence interval estimate of a population parameter contains likely values of that parameter, reject a claim that the population parameter has a value that is not included in the confidence interval.

Procedure for Finding *P*-Values

Figure 7



Wording of Final Conclusion

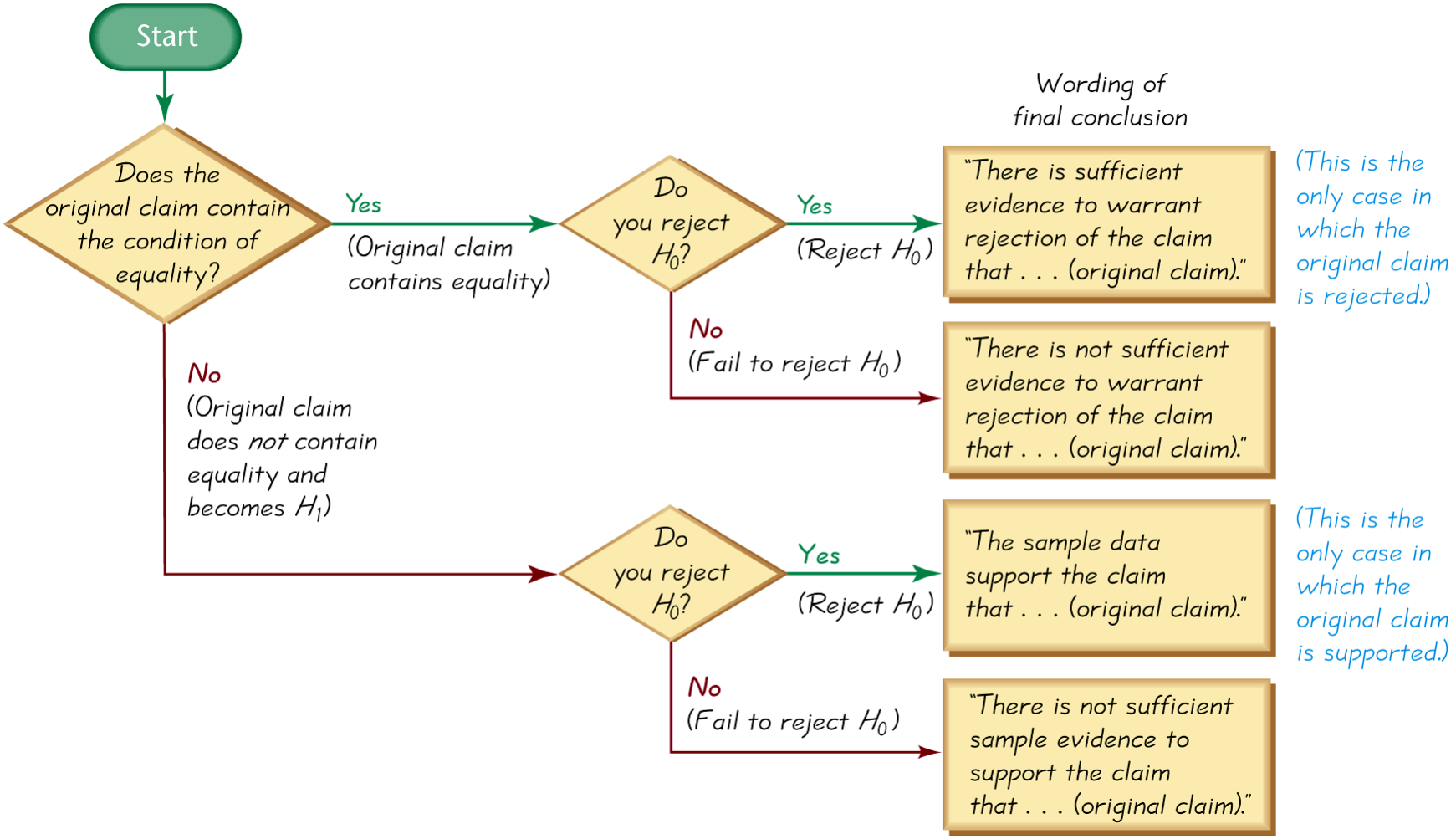


Figure 8

Accept Versus Fail to Reject

- **Some texts use "accept the null hypothesis."**
- **We are not proving the null hypothesis.**
- **The sample evidence is not strong enough to warrant rejection (such as not enough evidence to convict a suspect).**

Type I Error

- **A Type I Error is the mistake of rejecting the null hypothesis when it is true.**

- **The symbol α (alpha) is used to represent the probability of a type I error.**

Type II Error

- **A Type II Error is the mistake of failing to reject the null hypothesis when it is false.**
- **The symbol β (beta) is used to represent the probability of a type II error.**

Type I and Type II Errors

Table: Type I and Type II Errors

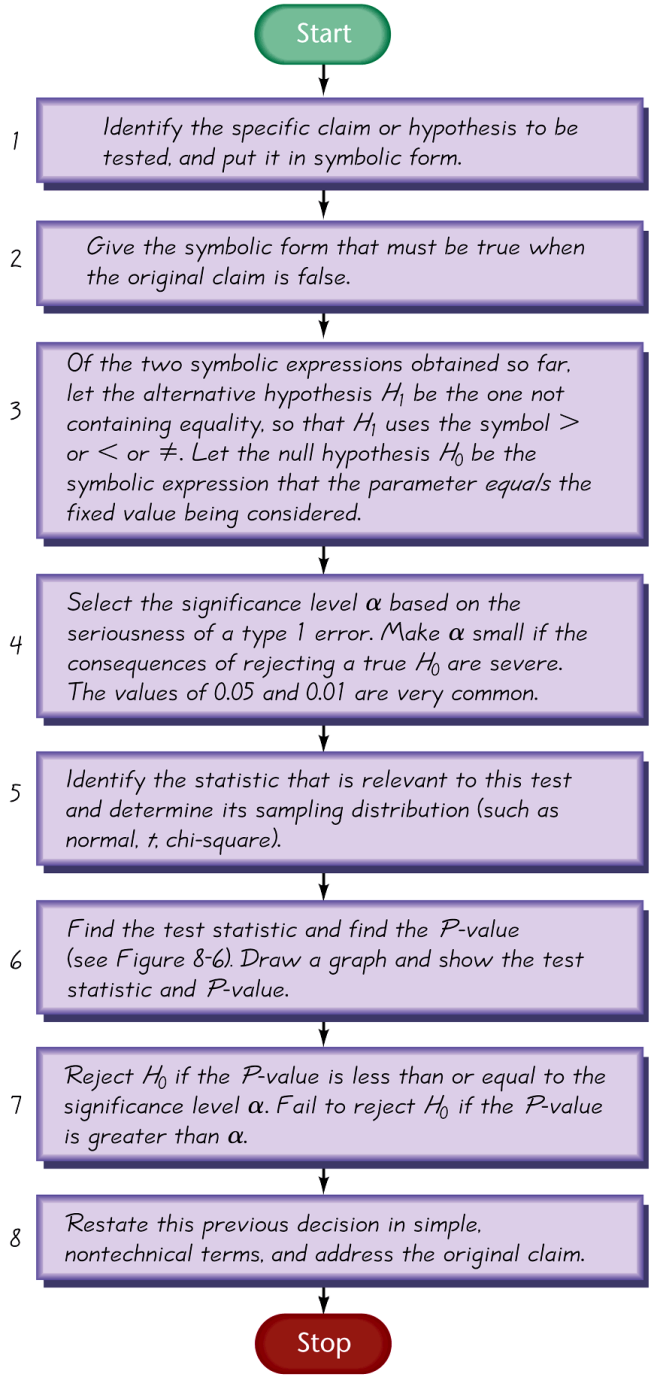
		True State of Nature	
		The null hypothesis is true	The null hypothesis is false
Decision	We decide to reject the null hypothesis	**Type I error** (rejecting a true null hypothesis) α	Correct decision
	We fall to reject the null hypothesis	Correct decision	**Type II error** (failing to reject a false null hypothesis) β

Controlling Type I and Type II Errors

- **For any fixed α, an increase in the sample size *n* will cause a decrease in β.**
- **For any fixed sample size *n*, a decrease in α will cause an increase in β. Conversely, an increase in α will cause a decrease in β.**
- **To decrease both α and β, increase the sample size.**

Power of a Hypothesis Test

The Power of a Hypothesis Test is the probability (1 - β) of rejecting a false null hypothesis, which is computed by using a particular significance level α and a particular value of the population parameter that is an alternative to the value assumed true in the null hypothesis. That is, the power of the hypothesis test is the probability of supporting an alternative hypothesis that is true.



Comprehensive Hypothesis Test: *P*-Value Method

Figure 9

Comprehensive Hypothesis Test: Traditional Method

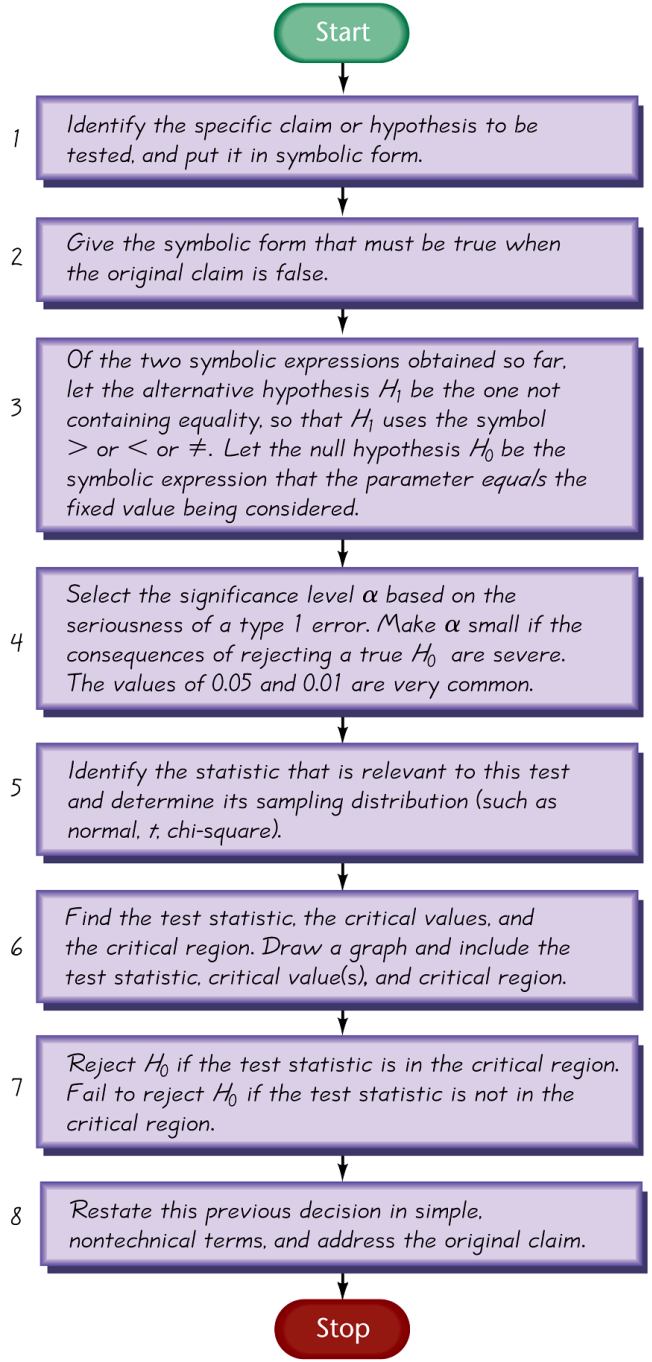


Figure 10

Comprehensive Hypothesis Test

A confidence interval estimate of a population parameter contains the likely values of that parameter. We should therefore reject a claim that the population parameter has a value that is not included in the confidence interval.

Table Confidence Level for Confidence Interval

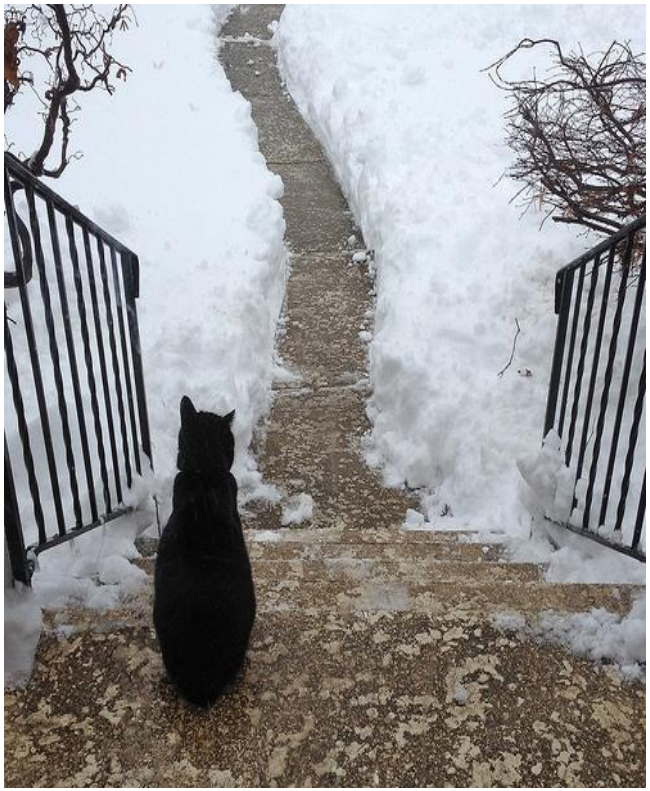
		Two-Tailed Test	One-Tailed Test
Significance Level for Hypothesis Test	0.01	99%	98%
	0.05	95%	90%
	0.10	90%	80%

Summary of Hypothesis Testing

An **objective** method of making decisions or **inferences** from sample data (evidence)

Sample data used to choose between two choices i.e. **Hypotheses** or Statements about a population

We do this by comparing what we have observed to what we expected if one of the statements (**Null Hypothesis**) was true



Summary of Hypothesis Testing

Always we have Two Hypotheses:

H_1: Research (Alternative) Hypothesis

What we aim to gather evidence of

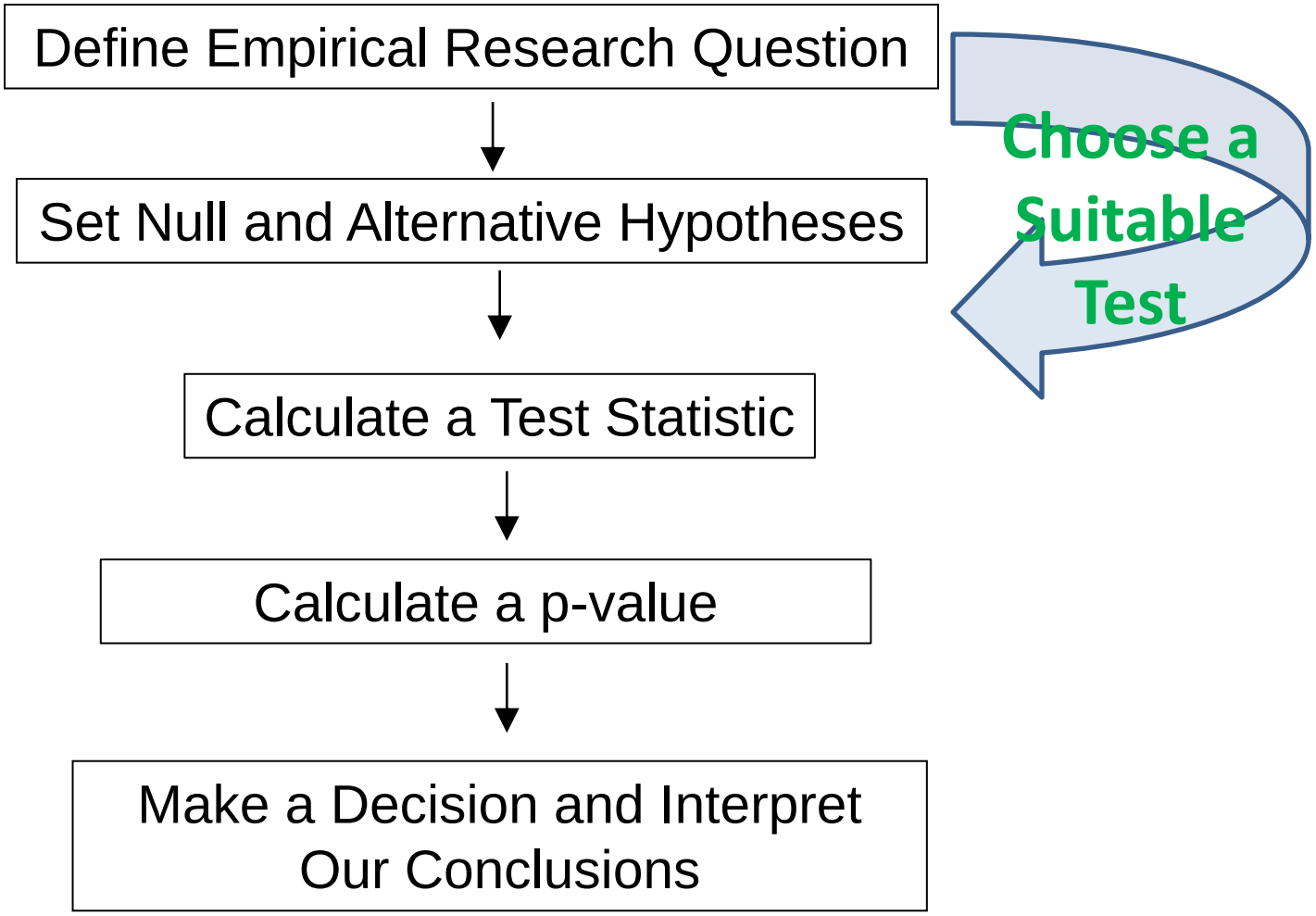
Typically that there **is** a difference/effect/relationship etc.

H_0: Null Hypothesis

What we assume is true to begin with

Typically that there is **no** difference/effect/relationship etc.

Summary of Hypothesis Testing



Summary of Hypothesis Testing

We can use statistical software to undertake a hypothesis test
e.g. SPSS (Statistical Package for Social Sciences)
One part of the output is the p-value (P)

If P < **0.05 reject** H_0 (Reject Null Hypothesis) => **Evidence** of H_1 being true (i.e. **IS** association (Alternative Hypothesis))

If P > **0.05 do not** reject H_0 (Accept the Null Hypothesis)
(i.e. **NO** association (Null Hypothesis))

Summary of Hypothesis Testing (Choosing the Right Test)

1) A clearly defined research question

2) What is the dependent variable and what type of variable is it?

3) How many independent variables are there and what data types are they?

4) Are you interested in comparing means or investigating relationships?

5) Do you have repeated measurements of the same variable for each subject?

Summary of Hypothesis Testing (Choosing the Right Test)

- Clarity of research questions with measurable quantities
- Which variables will help answer these research questions
- Think about what test is needed before carrying out a study so that the right type of variables are collected

Summary of Hypothesis Testing (Choosing the Right Test)

- How many variable are involved?
- Two – interested in the relationship
- One dependent and one independent
- One dependent and several independent variables: some may be controls
- Relationships between more than two: multivariate techniques (not covered here)

Summary of Hypothesis Testing (Choosing the Right Test)

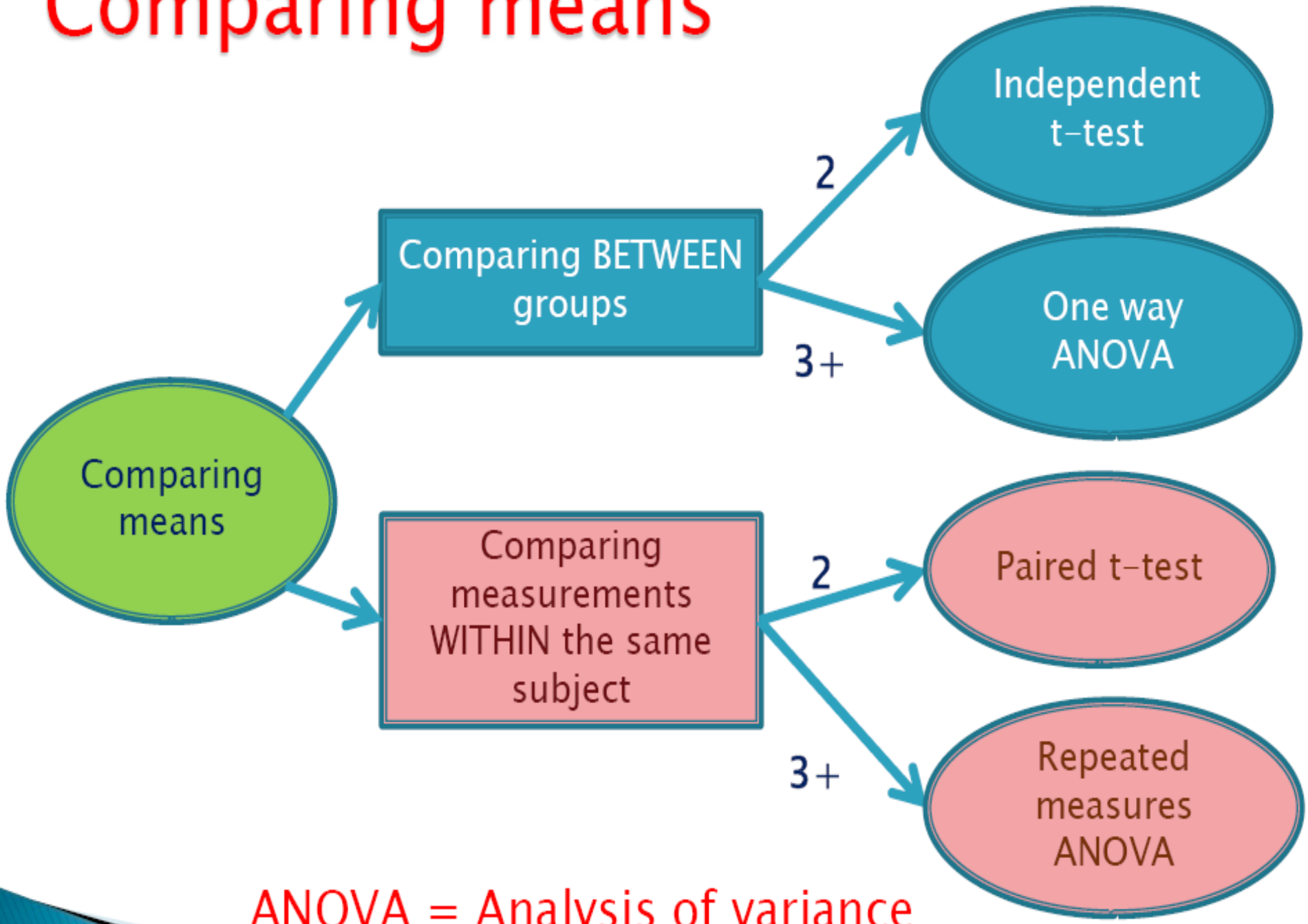
Comparing the Means

- Dependent = Scale
- Independent = Categorical

- How many means are we comparing?

- Do we have independent groups or repeated measurements on each person?

Comparing means



ANOVA = Analysis of variance

Exercise – Comparing the Means

Research question	Dependent variable	Independent variable	Test
Do women do more housework than men?	Housework (hrs per week) (Scale)	Gender (Nominal)	Independent t-test
Does Margarine X reduce cholesterol? Everyone has cholesterol measured on 3 occasions	Cholesterol (Scale)	Occasion (Nominal)	Repeated measures ANOVA
Which of 3 diets is best for losing weight?	Weight lost on diet (Scale)	Diet (Nominal)	One-way ANOVA

Parametric or non-parametric?

Statistical tests fall into two types:

Parametric tests

Assume data follows a particular distribution e.g. normal

Non-parametric

Nonparametric techniques are usually based on ranks/ signs rather than actual data

Non-parametric Tests

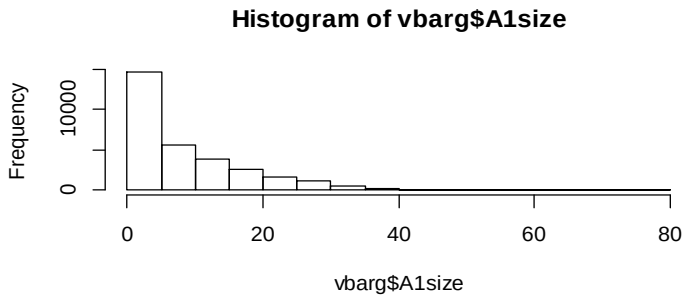
- Non-parametric methods are used when:
  - Data is ordinal
  - Data does not seem to follow any particular shape or distribution (e.g. Normal)
  - Assumptions underlying parametric test not met
  - A plot of the data appears to be very skewed
  - There are potential influential outliers in the dataset
  - Sample size is small

Note: Parametric tests are fairly robust to non-normality.
Data has to be very skewed to be a problem

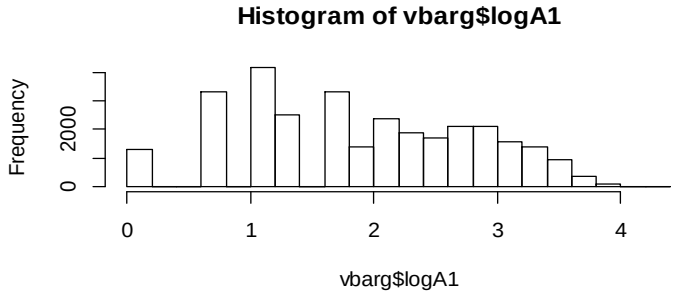
What can be done about non-normality?

If the data are not normally distributed, there are two options:

1. **Use a non-parametric test**
2. **Transform the dependent variable**



For positively skewed data, taking the log of the dependent variable often produces normally distributed values

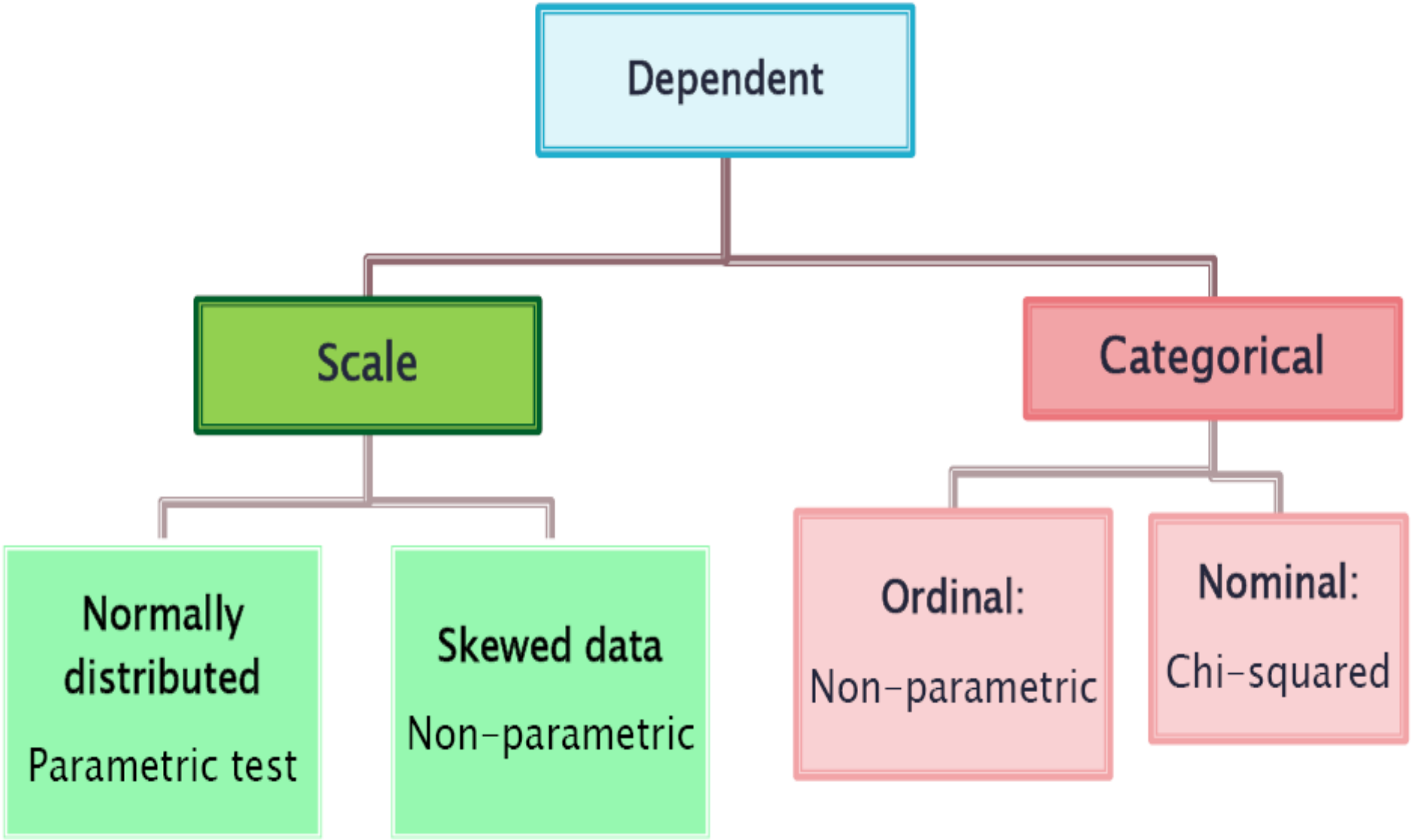


Non-parametric tests

Parametric test	What to check for normality	Non-parametric test
Independent t-test	Dependent variable by group	Mann-Whitney test
Paired t-test	Paired differences	Wilcoxon signed rank test
One-way ANOVA	Residuals/Dependent	Kruskal-Wallis test
Repeated measures ANOVA	Residuals	Friedman test
Pearson's Correlation Co-efficient	At least one of the variables should be normal	Spearman's Correlation Co-efficient
Linear Regression	Residuals	None - transform the data

Notes: The residuals are the differences between the observed and expected values.

Summary


Dependent
Scale
Categorical
Normally distributed
Parametric test
Skewed data
Non-parametric
Ordinal:
Non-parametric
Nominal:
Chi-squared

Statistical Hypothesis Testing: T-tests
(Paired or Independent (Unpaired) Data?)

T-tests are used to compare two population means

- **Paired Data:** same individuals studied at two different times or under two conditions PAIRED T-TEST
- **Independent (Unpaired) Data:** data collected from two separate groups INDEPENDENT SAMPLES T-TEST (UNPAIRED T-TEST)

What is the t-distribution?

- The t-distribution is similar to the standard normal distribution but has an additional parameter called degrees of freedom (df or v)

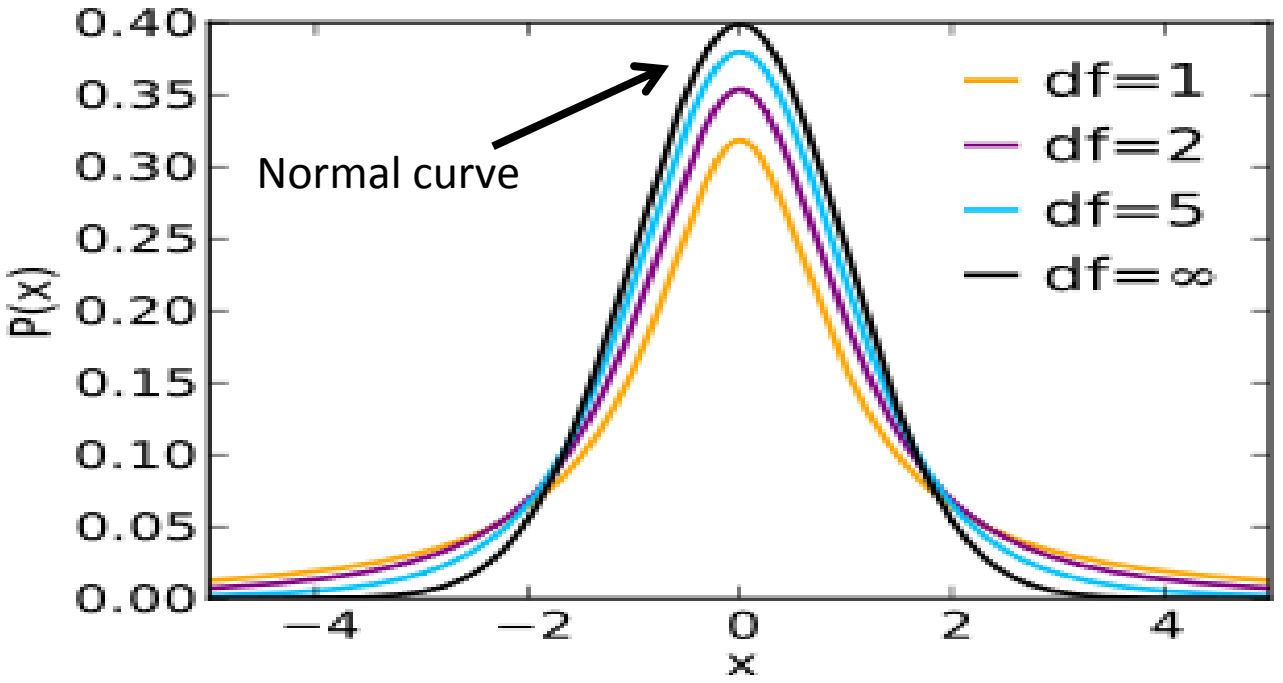
For a paired t-test, v = number of pairs – 1

For an independent t-test, $v = n_{group1} + n_{group2} - 2$

- Used for small samples and when the population standard deviation is not known
- Small sample sizes have heavier tails

Relationship to Normal Distrbution

As the sample size gets big, the t-distribution matches the Normal Distribution



Assumptions in t-Tests

Normality: Plot histograms

- One plot of the paired differences for any paired data
- Two (One for each group) for independent samples
- Don't have to be perfect, just roughly symmetric

Equal Population variances: Compare sample standard deviations

- As a rough estimate, one should be no more than twice the other
- Do an F-test (Levene's in SPSS) to formally test for differences

However the *t*-test is very robust to violations of the assumptions of Normality and equal variances, particularly for moderate (i.e. >30) and larger sample sizes

What if the assumptions are not met?

There are alternative tests which do not have these assumptions

Test	Check	Equivalent non-parametric test
Independent t-test	Histograms of data by group	Mann-Whitney
Paired t-test	Histogram of paired differences	Wilcoxon signed rank

ANOVA Test

- Let us go through the procedure for one-way ANOVA
  - That means, one independent variable
- Multi-way ANOVA computations are very cumbersome to do manually
  - So it is better to do computations using statistical packages

ANOVA Test

Compares the means of several groups.

- Which diet is the best?
  Dependent: Weight lost (Scale)
  Independent: Diet 1, 2 or 3 (Nominal)
- Null Hypothesis: The mean weight lost on diets 1, 2 and 3 is the same
- Alternative Hypothesis: The mean weight lost on diets 1, 2 and 3 are not all the same

Summary Statistics

	Overall	Diet 1	Diet 2	Diet 3
Mean	3.85	3.3	3.03	5.15
Standard deviation	2.55	2.24	2.52	2.4
Number in group	78	24	27	27

- Which diet was the best?
- Are the standard deviations similar?

ANOVA Test

ANOVA = ANalysis Of VAriance

We compare variation **between** groups relative to variation **within** groups

Population variance estimated in two ways:

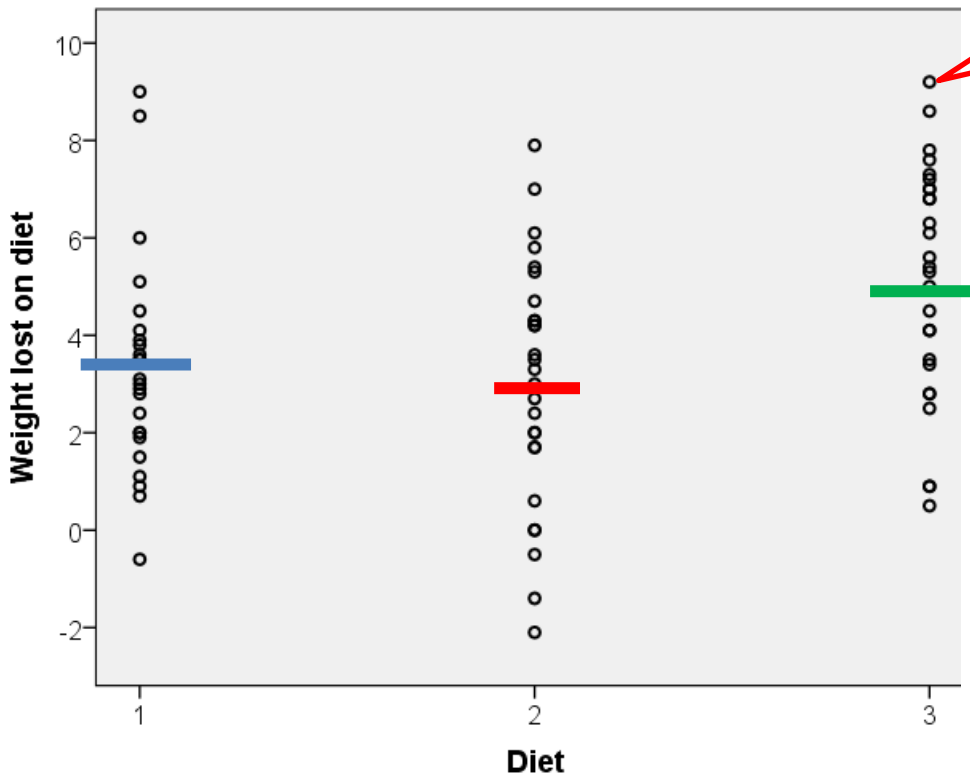
One based on variation **between** groups we call the **M**ean **S**quare due to **T**reatments/ **MST/ $MS_{between}$**

Other based on variation **within** groups we call the **M**ean **S**quare due to **E**rror/ **MSE/ MS_{within}**

Within the Group Variation

Residual =difference between an individual and the group mean

SS_{within}=sum of squared residuals

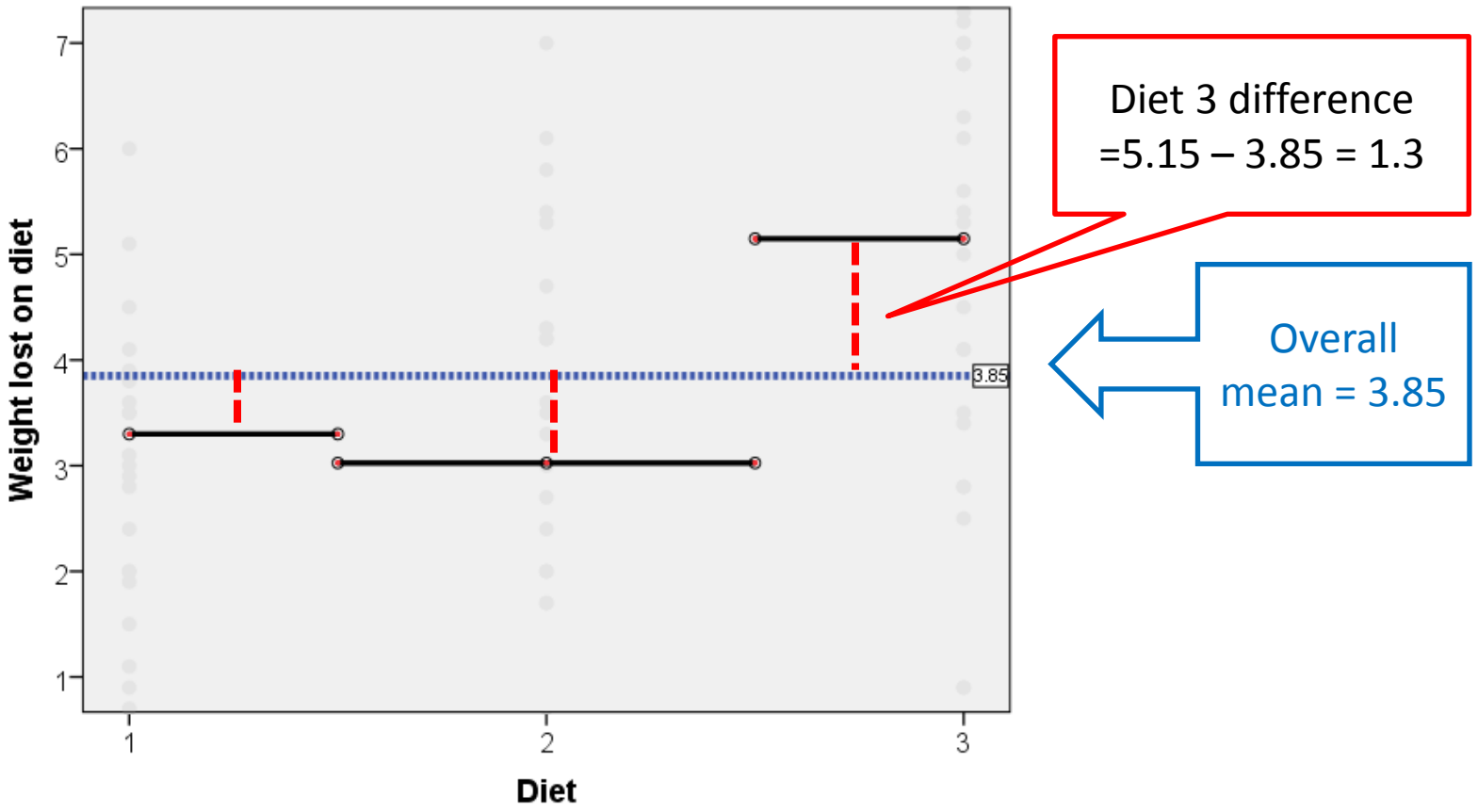


Person lost 9.2kg kg, hence residual =9.2 – 5.15 = 4.05

Mean weight lost on diet 3 = 5.15kg

Between the Group Variation

Differences between each group mean and the overall mean



Sum of Squares Calculations

K = Number of Groups

$$SS_{within} = \sum_{j=1}^{k}\sum_{i=1}^{n}\left(x_{ij} - \overline{x}_j\right)^2$$

$$= \sum_{i=1}^{24}\left(x_i - 3.3\right)^2 + \sum_{i=1}^{27}\left(x_i - 3.03\right)^2 + \sum_{i=1}^{27}\left(x_i - 5.15\right)^2 = 430.179$$

$$SS_{Between} = \sum_{j=1}^{k} n_j\left(\overline{x}_j - \overline{x}_T\right)^2$$

$$= 24\left(3.3 - 3.85\right)^2 + 27\left(3.03 - 3.85\right)^2 + 27\left(5.15 - 3.85\right)^2 = 71.094$$

ANOVA Test Statistics

Summary ANOVA

Source	Sum of Squares	Degrees of Freedom	Variance Estimate (Mean Square)	*F* Ratio
Between	SS_B	$K-1$	$MS_B = \frac{SS_B}{K-1}$	$\frac{MS_B}{MS_W}$
Within	SS_W	$N-K$	$MS_W = \frac{SS_W}{N-K}$	
Total	$SS_T = SS_B + SS_W$	$N-1$		

Test Statistic (usually reported)

N = total observations in all groups,

K = number of groups

Test Statistic (by hand)

Filling in the boxes

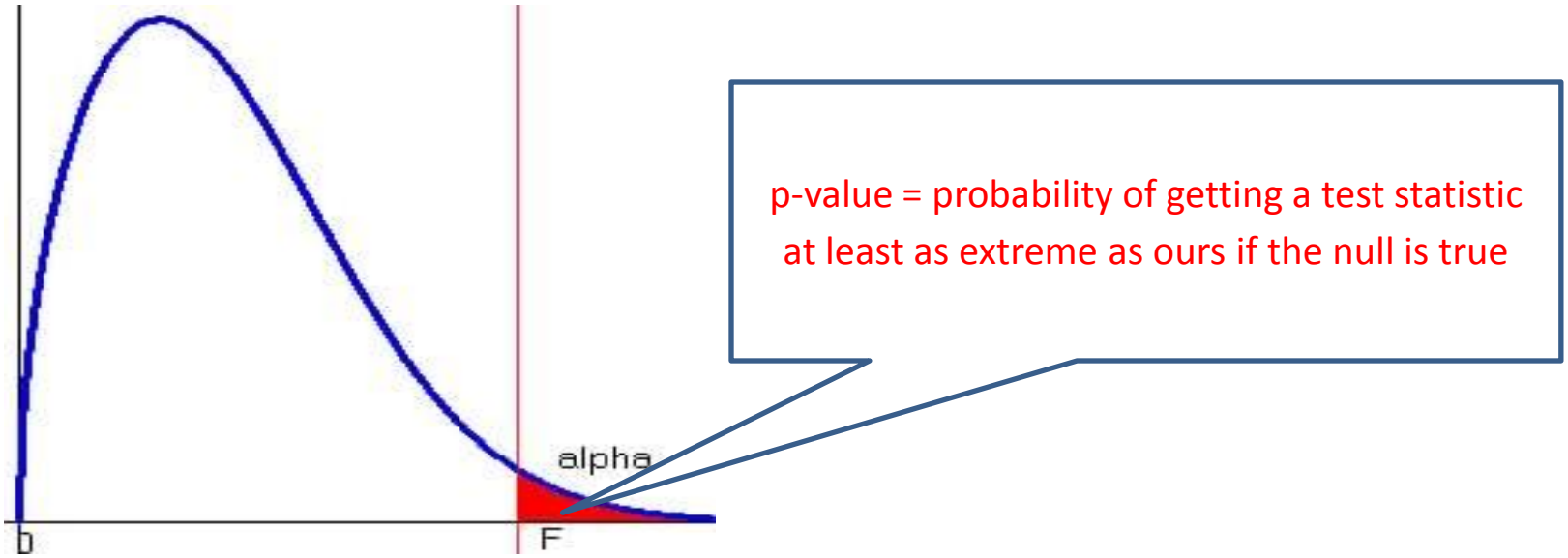
	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio (Test Statistic)
$SS_{between}$	71.045	2	35.522	**6.193**
SS_{within}	430.180	75	5.736	
SS_{total}	501.275	77		

$$\text{F-ratio} = \frac{\text{Mean between group sum of squared differences}}{\text{Mean within group sum of squared differences}}$$

If F-ratio > 1, then there is a bigger difference between the groups than within the groups

P-value

- The p-value for ANOVA is calculated using the F-distribution
- If we repeated the experiment several times, then we would get a variety of test statistics



Test Statistic

One Way ANOVA

$$\text{Test Statistic} = \frac{\text{between group variation}}{\text{within group variation}} = \frac{MS_{Diet}}{MS_{Error}} = 6.197$$

Tests of Between-Subjects Effects

Dependent Variable: Weight lost on diet (kg)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	71.094[a]	2	35.547	6.197	.003
Intercept	1137.494	1	1137.494	198.317	.000
Diet	71.094	2	35.547	6.197	.003
Error	430.179	75	5.736		
Total	1654.350	78			
Corrected Total	501.273	77			

a. R Squared = .142 (Adjusted R Squared = .119)

$MS_{between}$

MS_{within}

There was a significant difference in weight lost between the diets (p=0.003)

Post-hoc Tests

If there is a significant ANOVA result, pairwise comparisons are made

They are t-tests with adjustments to keep the type 1 error to a minimum

- Tukey's and Scheffe's tests are the most commonly used post-hoc tests.
- Hochberg's GT2 is better where the sample sizes for the groups are very different.

Post-hoc Tests

Which diets are significantly different?

Multiple Comparisons

Dependent Variable: Weight lost on diet (kg)

	(I) Diet	(J) Diet	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	1	2	.2741	.67188	.912	-1.3325	1.8806
		3	-1.8481*	.67188	.020	-3.4547	-.2416
	2	1	-.2741	.67188	.912	-1.8806	1.3325
		3	-2.1222*	.65182	.005	-3.6808	-.5636
	3	1	1.8481*	.67188	.020	.2416	3.4547
		2	2.1222*	.65182	.005	.5636	3.6808

Write up the results and conclude with which diet is the best.

Pairwise Comparisons

Test	p-value
Diet 1 vs Diet 2	P = 0.912
Diet 1 vs Diet 3	P = 0.02
Diet 2 vs Diet 3	P = 0.005

There is no significant difference between Diets 1 and 2 but there is between diet 3 and diet 1 ($p = 0.02$) and diet 2 and diet 3 ($p = 0.005$).

The mean weight lost on Diets 1 (3.3kg) and 2 (3kg) are less than the mean weight lost on diet 3 (5.15kg).

Assumptions for ANOVA Test

Assumption	How to check	What to do if assumption not met
Normality: The residuals (difference between observed and expected values) should be normally distributed	Histograms/ QQ plots/ normality tests of residuals	Do a Kruskall-Wallis test which is non-parametric (does not assume the normality)
Homogeneity of variance (each group should have a similar standard deviation)	Levene's test	Welch test instead of ANOVA and Games-Howell for post-hoc or Kruskall-Wallis

ANOVA Illustrated

- Let's illustrate the idea with the following example:

  Suppose we have designed a new text entry technique for mobile phones. We think the design is good. In fact, we feel that our method is *better* than the most widely used state-of-the-art techniques, namely: multi-tap and T9. We decide to undertake some empirical research to evaluate our design invention and to compare it with these current techniques?

  Suppose "better" is defined in terms of error rate

Empirical Data

- In order to ascertain the validity of our claim, we conducted the experiments and thus collected the following empirical data (error rate of participants under different test conditions)

Participants	Our Method	Multi-tap	T9
1	3	5	7
2	2	2	4
3	1	4	5
4	1	2	3
5	4	3	6

ANOVA Steps - 1

- Now Let's compute means, standard deviations (SD) and variances for each test condition (over all participants)

	Our Method	Multi-tap	T9
Mean	**2.20**	**3.20**	**5.00**
SD	**1.30**	**1.30**	**1.58**
Variance	**1.70**	**1.70**	**2.50**

ANOVA Steps - 1

- Also calculate “grands” – values involving all irrespective of groups
  - Grand Mean (mean of means) = 3.467
  - Grand SD (w.r.t. grand mean) = 1.767
  - Grand Variance (w.r.t. grand mean) = 3.124

ANOVA Steps - 2

- Calculate “total sum of squares (SS_T)”

$$SS\_T = \sum (x\_i - mean\_grand)^2$$
$$= 43.74$$

Where, *x_i* is the error rate value of the i-th participant (among all)

ANOVA Steps - 2

- An associated concept is the degrees of freedom (DoF (df)), which is the number of observations that are free to vary
- DoF (df) can be calculated simply as the (number of things used to calculate – 1)
  - For SS_T calculation, DoF (df) = N-1

ANOVA Steps - 3

- Next calculate the “model sum of square (SS_M)”
  - Calculate (mean_group_*i*-mean_grand) for the i-th group
  - Square the above
  - Multiply by *n_i*, the number of participants in the *i*-th group
  - Sum for all groups

ANOVA Steps - 3

- In the example,

$$SS\_M = 5(2.200 - 3.467)^2 + 5(3.200 - 3.467)^2 + 5(5.000 - 3.467)^2$$

$$= 20.135$$

- DoF (df) = number of group means – 1

= 3 - 1 = 2 (in our example)

ANOVA Steps - 4

- Calculate the “residual sum of square (SS_R)” and the corresponding DoF

$$SS\_R = SS\_T - SS\_M$$

$$DoF\ (SS\_R) = DoF\ (SS\_T) - DoF\ (SS\_M)$$

- Thus, in this example,

$$SS\_R = 43.74 - 20.14 = 23.60$$

$$DoF\ (SS\_R) = 14 - 2 = 12$$

ANOVA Steps - 5

- Calculate two “average sum of squares” or “mean squares (MS)”
- Model MS (MS_M) = SS_M/DoF(SS_M)

  = 20.135/2 = 10.067 (for our example)
- Residue MS (MS_R) = SS_R/DOF(SS_R)

  = 23.60/12 = 1.967 (for our example)

ANOVA Steps - 6

- Calculate the “F-ratio” (simply divide MS_M by MS_R)
  - F-ratio (F) = 10.067/1.967 = 5.12 (for our example)
- DoFs associated with the F-ratio are the DoFs used to calculate the two mean squares [that is DoF(SS_M) and DoF(SS_R)]
  - In our case, these are 2, 12 respectively
- Hence, in our case, the F-ratio would be written as F(2, 12) i.e. F-ratio (F) = 5.12

ANOVA Steps - 6

- Look up the critical value of F-ratio (F)
  - The critical values for different “significance levels” / thresholds (α) are available in a tabular form
  - The critical values signifies the value of F that we would expect to get by chance for α% of tests

ANOVA Steps - 6

- Example (for more details, please see the next slide)
  - To find the critical value of F(2, 12) from the Table for α=.05, look at 2nd column, 12th row for .05
  - Which is 3.89
  - That means, 3.89 is the F-value we would expect to get by chance for 5% of the tests.

df (Denominator)	p	df (Numerator)																
		1	2	3	4	5	6	7	8	9	10	15	20	25	30	40	50	1000
1	.05	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95	248.01	249.26	250.10	251.14	251.77	254.19
	.01	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6157.31	6208.74	6239.83	6260.65	6286.79	6302.52	6362.70
2	.05	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.43	19.45	19.46	19.46	19.47	19.48	19.49
	.01	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40	99.43	99.45	99.46	99.47	99.47	99.48	99.50
3	.05	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.70	8.66	8.63	8.62	8.59	8.58	8.53
	.01	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.69	26.58	26.50	26.41	26.35	26.14
4	.05	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86	5.80	5.77	5.75	5.72	5.70	5.63
	.01	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.91	13.84	13.75	13.69	13.47
5	.05	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62	4.56	4.52	4.50	4.46	4.44	4.37
	.01	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.45	9.38	9.29	9.24	9.03
6	.05	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94	3.87	3.83	3.81	3.77	3.75	3.67
	.01	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.30	7.23	7.14	7.09	6.89
7	.05	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51	3.44	3.40	3.38	3.34	3.32	3.23
	.01	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	6.06	5.99	5.91	5.86	5.66
8	.05	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22	3.15	3.11	3.08	3.04	3.02	2.93
	.01	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.26	5.20	5.12	5.07	4.87
9	.05	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01	2.94	2.89	2.86	2.83	2.80	2.71
	.01	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.71	4.65	4.57	4.52	4.32
10	.05	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85	2.77	2.73	2.70	2.66	2.64	2.54
	.01	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.31	4.25	4.17	4.12	3.92
11	.05	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72	2.65	2.60	2.57	2.53	2.51	2.41
	.01	9.65	[illegible]	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	4.01	3.94	3.86	3.81	3.61
12	.05	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62	2.54	2.50	2.47	2.43	2.40	2.30
	.01	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.76	3.70	3.62	3.57	3.37
13	.05	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53	2.46	2.41	2.38	2.34	2.31	2.21
	.01	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.57	3.51	3.43	3.38	3.18
14	.05	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46	2.39	2.34	2.31	2.27	2.24	2.14
	.01	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.41	3.35	3.27	3.22	3.02
15	.05	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40	2.33	2.28	2.25	2.20	2.18	2.07
	.01	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.28	3.21	3.13	3.08	2.88
16	.05	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35	2.28	2.23	2.19	2.15	2.12	2.02
	.01	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.16	3.10	3.02	2.97	2.76
17	.05	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31	2.23	2.18	2.15	2.10	2.08	1.97
	.01	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.07	3.00	2.92	2.87	2.66
18	.05	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27	2.19	2.14	2.11	2.06	2.04	1.92
	.01	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.98	2.92	2.84	2.78	2.58
19	.05	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23	2.16	2.11	2.07	2.03	2.00	1.88
	.01	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.91	2.84	2.76	2.71	2.50
20	.05	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20	2.12	2.07	2.04	1.99	1.97	1.85
	.01	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.84	2.78	2.69	2.64	2.43
22	.05	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15	2.07	2.02	1.98	1.94	1.91	1.79
	.01	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.73	2.67	2.58	2.53	2.32

Implication

- Thus, we get the critical value = 3.89 for F(2,12), α=.05
- Note that F(2, 12)=5.12 > the critical value
  - Implies that the effect of test conditions has a significant effect on the outcome w.r.t. α=.05

Reporting F-Statistic

- We can report the result as “our proposed method has a significant effect on reducing user errors [$F(2,12)=5.12$, $p<.05$] as compared to the other methods.”
- If it is found that the effect is not significant, it is reported as “our method has no significant effect on reducing user errors [$F(1,9)=0.634$, ns] as compared to the other methods.”

A Note of Caution

- ANOVA requires that
  - Empirical Data should have normally distributed sampling distribution and from a normally distributed population
  - Variances in each experimental condition are fairly similar
  - Observations should be independent
  - Dependent variables should be measured on at least an interval scale
- The first two may be ignored if group sizes are equal
  - Otherwise, ALL conditions MUST have to be met