

HKBM7000

Approaches to Research and Academic Communications Day 4 (8 August 2021)

Data-Analytic Methods

[1]

Topics

- Descriptive Statistics
 - Frequency Distribution
 - Central Tendency
 - Variability
- Inferential Statistics
 - Checking Assumptions
 - Logic of Inferential Statistics
 - Approaches to Hypothesis Testing
 - General Procedure for Hypothesis Testing
- Statistical Tests
 - Z Tests
 - Comparing Proportions
 - T Tests
 - Analysis of Variance (ANOVA)
 - Correlation
 - Regression Analysis

[2]

DATA ANALYTIC METHODS AND HYPOTHESIS TESTING

What do we need to know?

- Types of data
- Types of statistics
 - Descriptive VS inferential
- Levels of measurement
 - Central tendency
 - Variance
 - P-Value
- Types of statistical tests
 - Difference VS relationship
 - Parametric VS non-parametric



Types of Data

Nominal Numbers Assigned to Runners



Ordinal Rank Order of Winners



Interval Performance Rating on a 0 to 10 Scale

8.2 9.1 9.6

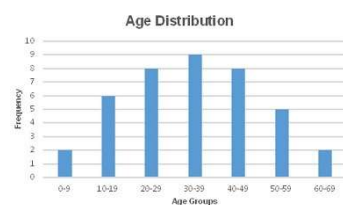
Ratio Time to Finish in Seconds

15.2 14.1 13.4

Descriptive Statistics: Frequency Distribution

What are descriptive statistics?

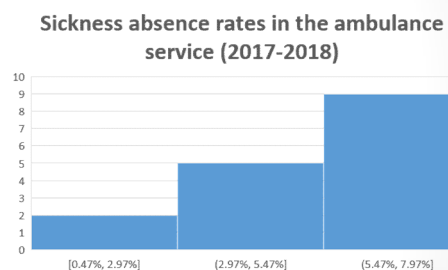
- Procedures that summarise, organise and simplify data to aid with interpretation
- Grouping/ordering the data can tell us useful information



Shoe Size				
	Frequency	Percent	Valid Percent	Cumulative Percent
Valid 6.0	1	6.7	6.7	6.7
6.5	2	13.3	13.3	20.0
7.0	3	20.0	20.0	40.0
7.5	2	13.3	13.3	53.3
8.5	2	13.3	13.3	66.7
9.0	1	6.7	6.7	73.3
9.5	2	13.3	13.3	86.7
10.5	2	13.3	13.3	100.0
Total	15	100.0	100.0	

Frequency Distribution

- Help understand our data
- Build up a picture of what is happening with our data
- Simple, visual method to present and categorise data



Frequency distribution

- How often a score appears in a data set
- We can create a frequency distribution based on counting these occurrences

Grades	Frequency
A	12
B	10
C	6
D	4
Fail	2

Creating a frequency distribution

Summarise data;

8, 7, 5, 6, 4, 5, 7, 10, 7, 6

Rank raw data in order:

4, 5, 5, 6, 6, 7, 7, 7, 8, 10

Participant number	Score
1	8 letters
2	7 letters
3	5 letters
4	6 letters
5	4 letters
6	5 letters
7	7 letters
8	10 letters
9	7 letters
10	6 letters

Creating a frequency distribution

Consider the data and create a frequency distribution table:

4, 5, 5, 6, 6, 7, 7, 7, 8, 10

Score	Frequency (f)
4 letters	1
5 letters	2
6 letters	2
7 letters	3
8 letters	1
9 letters	0
10 letters	1

Grouped frequency distribution

- If range of possible values is large
- Frequency table would consist of empty entries

Length of time in service (X)	f
1 years	15
→ 2 years	0
3 years	0
→ 35 years	15

The principles of grouping data

- Usually limited to 10 intervals (1 – 5, 6 – 10, 11 – 15 etc.)
- Simple numbers are used for intervals: 2, 5, 10, 20, 50, 100

For example:
Years in service

1 – 5
6 – 10
• 11 – 15

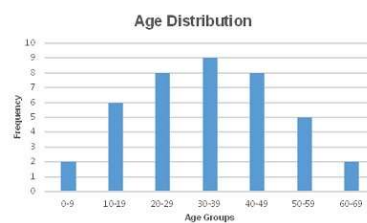
- These must not overlap
- Cover all the data
- Width is the same

$$6 - 1 = 5$$

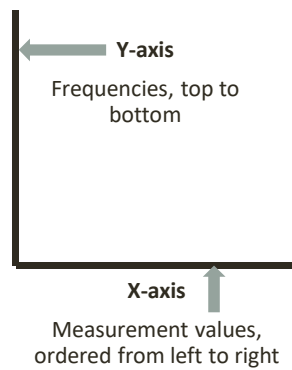
$$11 - 6 = 5$$

Frequency distribution graph

- Picture of the information that's available in a frequency distribution table
- Interval/ratio data – use a **histogram**



Drawing a frequency distribution graph



Using relative frequencies

- When you have a large number of participants
- Can be expressed as a **percentage** or **proportion**

$$\frac{f}{N} \times 100$$

$$\text{percentage} = \frac{\text{frequency}}{\text{sample_size}} \times 100$$

$$\frac{f}{N}$$

$$\text{proportion} = \frac{\text{frequency}}{\text{sample_size}}$$

Using relative frequencies: Example

- Percentage of participants who remembered 4 letters
 - $56 \div 214 \times 100 = 26\%$
- Proportion of participants who remembered 4 letters
 - $56 \div 214 = 0.26$

Score	Frequency (f)
4 letters	56
5 letters	62
6 letters	35
7 letters	23
8 letters	18
9 letters	12
10 letters	8

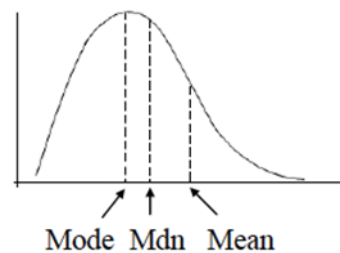
$N = 214$

Descriptive Statistics: Central Tendency

[18]

Central tendency

- Where scores tend to cluster
- 3 main methods of measuring central tendency
 - **Mode**
 - **Median**
 - **Mean**



Mode

Advantages

- Easy to work out
 - Most frequently occurring score

Disadvantages

- You can have more than one
- Not sensitive to all scores in the distribution

1, 1, 1, 2, 2, 2, 2, 2, 4, 4, 4,
5, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7,
7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 10

Mode = 7

Median

In a ranked distribution, the median is the middle score OR the average of the two middle scores in an even numbered distribution

1, 1, 1, 2, 2, 2, 2, 2, 4, 4, 4,
5, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7,
7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 10

Median = 6

Disadvantages

- All other scores in the distribution are irrelevant

Mean

- The average score
- Very common in psychology

Sample mean $\rightarrow \bar{x} = \frac{\sum x}{N}$

Sum total $\rightarrow \sum$

Score $\rightarrow x$

Number of observations $\rightarrow N$

Calculating the mean

Sample mean $\rightarrow \bar{x} = \frac{\sum x}{N}$

Sum total $\rightarrow \sum x$

Score $\rightarrow x$

Number of observations $\rightarrow N$

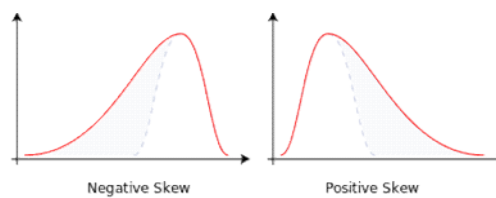
1, 1, 1, 2, 2, 2, 2, 2, 4, 4, 4,
5, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7,
7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 10
= 177

$$177 \div 33 = 5.36$$

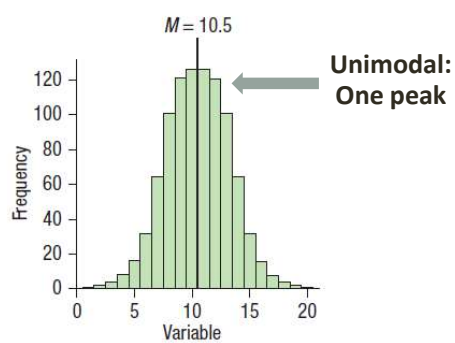
Mean = 5.36

Properties of the mean

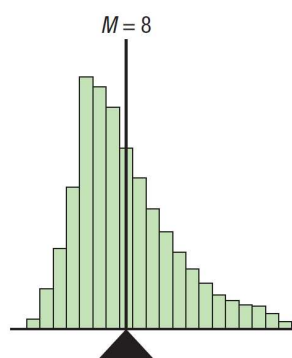
- Sensitive to every score in the distribution
 - Problems when data is **skewed**
 - Extreme scores will provide an unrepresentative mean



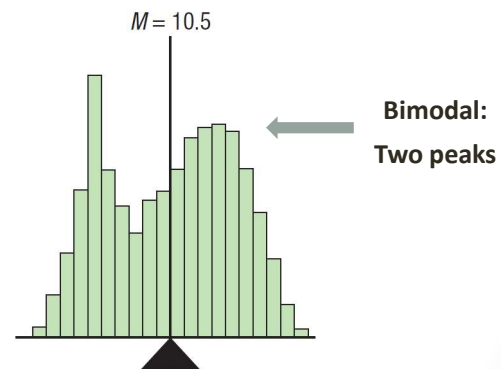
Example: Mean of a unimodal, symmetric distribution



Example: Mean of a unimodal, asymmetric distribution



Example: Mean of a bimodal distribution



Which is preferable?

Mean

- Be careful of outliers, particularly in small samples

Median

- When distribution is skewed
- Ordinal data

Mode

- Nominal data only



What are the problems with only looking at measures of central tendency?

**Descriptive Statistics:
Variability**

[30]

Variability

There are a number of ways in which the variability in a data set can be considered:

- Range
- Interquartile range
- Variance
- Standard deviation
- Standard error



Range

Highest score – lowest score

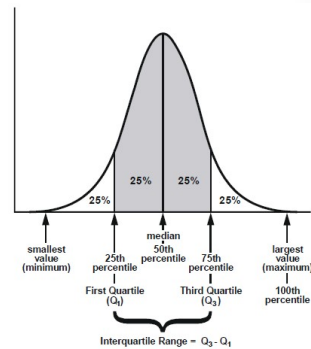
- Useful but simple measure
- Ignores most of the data
- Problems with extreme scores
- Depends on sample size

1, 1, 1, 2, 2, 2, 2, 2, 4, 4, 4,
5, 5, 5, 5, 5, 6, 6, 7, 7, 7, 7,
7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 10

$$\text{Range} = 10 - 1 = 9$$

Interquartile range

- Range of the middle 50% of the scores (75th – 25th percentile)
- No problems with outliers but still ignores 50% of scores



The variance

- A measure of **average squared deviations** between the observations and the sample mean

$$\text{Variance} \rightarrow SD^2 = \frac{\sum (x - M)^2}{n - 1}$$

Diagram illustrating the components of the variance formula:

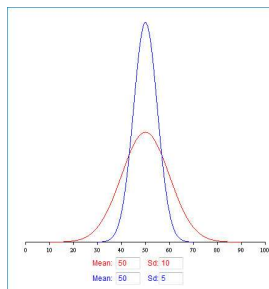
- Sum total** points to the summation symbol $\sum$.
- Score** points to the variable x .
- Mean** points to the variable M .
- Number of cases - 1** points to the denominator $n - 1$.

Standard Deviation (SD)

- The square root of the variance
- Represents the standard amount by which scores vary in a distribution
- Only meaningful for scales with equal intervals

Standard Deviation (SD)

- The larger the standard deviation, the most dispersed the values are



Should the denominator be n or $n-1$?

- Sample mean is an unbiased estimate of the population mean, BUT

$$sD^2 = \frac{\sum (x-M)^2}{n-1}$$

- Sample variance ALWAYS underestimates the population variance.

or

- Using $n-1$ as the denominator when working with sample variance can correct this bias.

$$sD^2 = \frac{\sum (x-M)^2}{n}$$

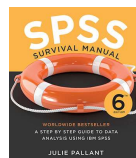
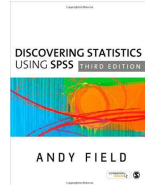
Standard Deviation (SD) Formula

Standard Deviation Formula	
Population	Sample
$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$ <i>X – The Value in the data distribution</i> <i>μ – The population Mean</i> <i>N – Total Number of Observations</i>	$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$ <i>X – The Value in the data distribution</i> <i>̄x – The Sample Mean</i> <i>n – Total Number of Observations</i>

Putting it all together: APA formatting

In text:

Children who viewed the violent cartoon displayed more aggressive responses ($M = 12.4$, $SD = 3.7$) than those who viewed the control cartoon ($M = 4.2$, $SD = 1.0$)



Inferential Statistics: Checking Assumptions

Inferential statistics

- Used to **test hypotheses** and **reach conclusions** that extend beyond the data alone
- Used to **make generalisations** from a sample to a population

Can one factor predict another?

Is there a relationship between one variable and another?

Do groups differ on some outcome variable?

Parametric and non-parametric tests

Parametric tests:

- Parameter: A number that describes how an event happens
- The data have to have certain characteristics (parameters)
- E.g. t-test, ANOVA, Pearson



Nonparametric tests:

- Less fussy about the nature of the data
- E.g. Mann-Whitney, Spearman

Parametric and non-parametric tests

Parametric tests:

- Robust / powerful tests
- E.g. likely to find a difference if there is one



Nonparametric tests:

- Not as strong / powerful
- E.g. Might not find a difference when there is one

Checking assumptions

- Statistical tests have several assumptions
- You must check these assumptions to ensure the data is suitable for the statistical analysis
- Important step in being confident of your interpretation/conclusions

Assumption 1: Data types

What type of data do you have? (e.g. is it continuous, categorical?)



Types of data

Parametric tests:

- Continuous data
- E.g. interval/ratio data



Nonparametric tests:

- Categorical
- Ordinal/nominal data



Assumption 2: Study design

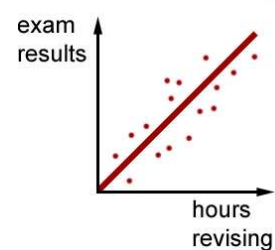
What design is your study?
(i.e. what do you want to
find out?)



Key quantitative study designs

Correlational

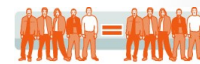
- Is there a relationship between x and y ?
- Pearson's r , Spearman's rank



Key quantitative study designs

Comparing groups

- Does group x differ from group y on an outcome variable
- Looking for differences between groups
 - Independent measures (different people in the groups)
 - 2 groups = IM test, Mann-Whitney test
 - >2 groups = One-way IM ANOVA, Kruskal-Wallis
 - Repeated measures (same people in the groups)
 - 2 groups = RM t-test, Mann-Whitney test
 - >2 groups = One-Way RM ANOVA, Friedman's test



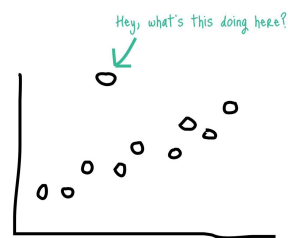
Assumption 3: Outliers

Does your data have any outliers? (e.g. any influential points that may affect the data)

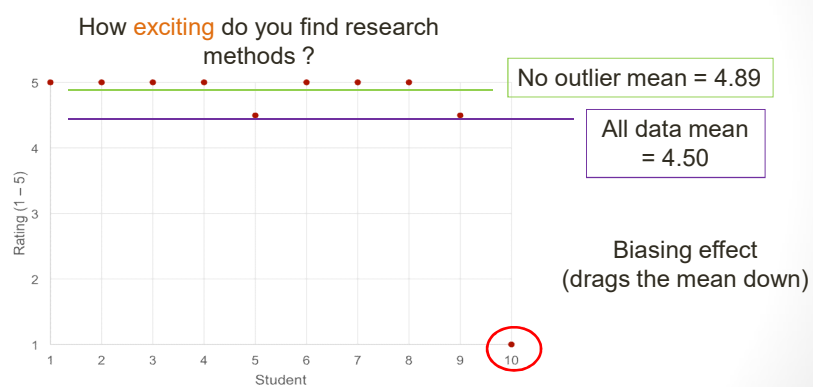


What is an outlier?

- "An observation that lies outside the overall pattern of a distribution of data" (Moore & McCabe, 1999)
- A single observation (or small number of observations) that are different (bigger or smaller) from the rest of the data



What is an outlier?



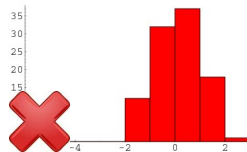
Dealing with outliers

Outliers usually indicate **problems**

What problems might an outlier indicate?

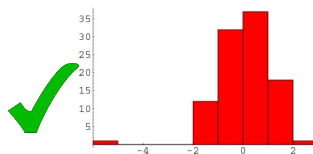


Dealing with outliers



Parametric tests

- No outliers



Nonparametric tests

- Can use with outliers

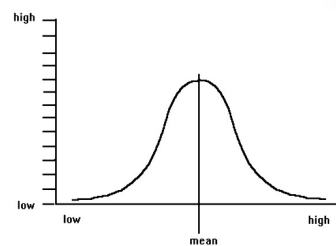
Assumption 4: Normality

Does your data have a normal distribution?



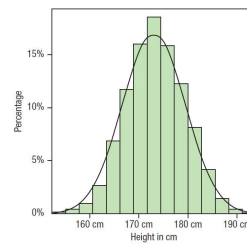
Assessing normality

- Parametric tests assume normal distribution
- It's important to check whether this is true and that your data is normally distributed
- **Normal distribution = 'bell shaped curve'**



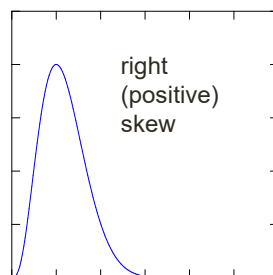
Assessing normality: Skewness

- 0 = perfectly normal
- - 1 to +1 = *may* be normal

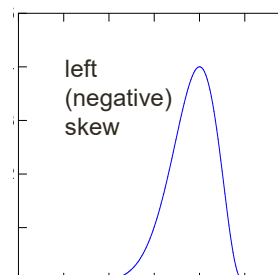


Normal

Assessing normality: Skewness



- $> +1$ = positive skew



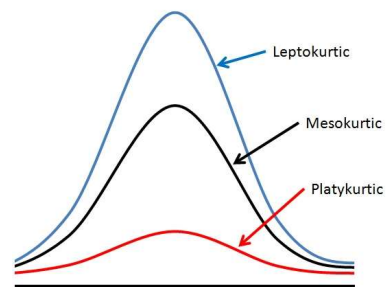
- < -1 = negative skew

Assessing normality: Kurtosis

Kurtosis

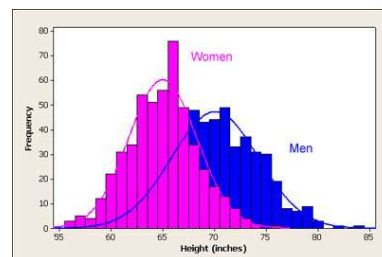
How pointy or flat the distribution is

- More than +1 = pointy
- 0 = perfectly normal
- Less than -1 = flat



Assessing normality: Groups

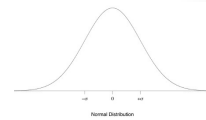
You need to check whether **each group** has normally distributed data



Assessing normality

Parametric tests

- Assume the data is normally distributed (bell-curve)
- Use mean and SD



Nonparametric tests

- No need to assume anything about the shape of the data
- Use ranks of scores *not* mean/SD



Inferential Statistics: Logic of Inferential Statistics

Using Inferential Statistics in Business Research

- Research Question:
 - E.g., Are female managers better or worse than male managers?
- Theory:
 - Theory 0: All people are equal
 - Theory 1: Women are by nature more nurturant than men, making them better managers
- Hypotheses:
 - Ho: All managers are equal, regardless of their gender
 - H1: Female managers are better than male managers

Logic of Inferential Statistics

- The population is too big to measure in its entirety
- Samples (as representative as possible) are drawn from the population and used to infer what is going on in the population
- Null hypothesis – An assumption about the state of the population (when no theory suggests otherwise)
 - E.g., all people are equally good (or bad) managers
- Alternative hypothesis – An expectation (based on theory) about the true state of the population
 - E.g., female managers are better than male managers
- Use the sample statistics to reject the null hypothesis or the alternative hypothesis

Symbols for Population and Sample Variables

Variable	Population	Sample
Mean	μ	$\bar{X}$
Proportion	Π	p
Variance	σ^2	s^2
Standard deviation	σ	s
Size	N	n
Standard error of the mean	σ_x	S_x
Standard error of the proportion	σ_p	S_p
Standardized variate (z)	$(X-\mu)/\sigma$	$(X-\bar{X})/S$
Coefficient of variation (CV)	σ/μ	$S/\bar{X}$

Margin of Error

- Each data point can be represented as the Sample Mean plus a Margin of Error, which is the variability from the point estimate
 - $X = \bar{X} + \text{Error}$
- Sample mean is an unbiased point estimate of the population mean
 - $\bar{X} = \mu + \text{Error}$
- The error can be represented in units of SD from the mean
 - $Z * \text{SD}$, therefore:
 - $X = \bar{X} + Z * \text{SD}$
 - $\bar{X} = \mu + Z * \text{SD}_{\bar{X}}$

Standard Score (z)

- A Z score represents how far a data point is away from the mean in terms of the number of standard deviations (SDs) away from the mean.
- Calculating Z scores is a way of **standardizing** values in a data set.

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

Data point → x μ ← Mean
Standard deviation → σ

This forms the basis of the *Confidence Interval*

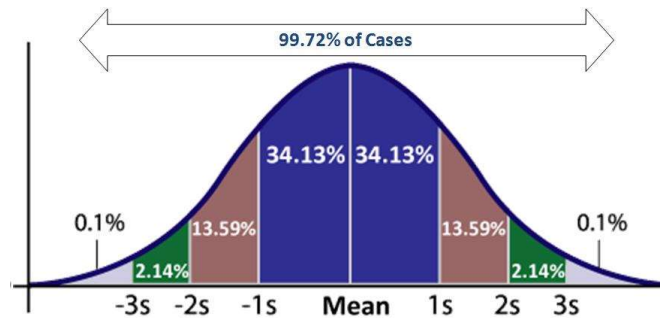
For the population:

$$z = \frac{x - \mu}{\sigma} \quad \Leftrightarrow \quad x = \mu + z * \sigma$$

For the sampling distribution:

$$z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \Leftrightarrow \quad \bar{x} = \mu + z * \sigma_{\bar{x}}$$

Percent of cases within standard deviations

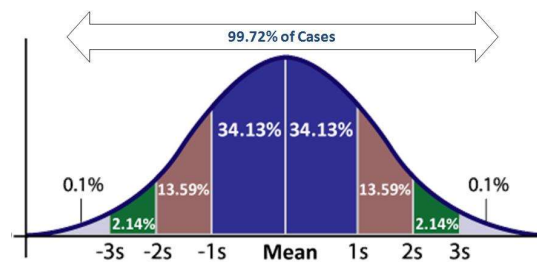


Confidence Level

- Percent of cases within a certain section of a distribution
- E.g., a 95% confidence level refers to a section of the distribution that covers 95% of the cases
- 1 - Significance Level ($1 - \alpha$)

Confidence Interval

- A range of scores within which a certain percent of cases can be found.
- E.g., suppose that the 99.72% confidence level is chosen.
 - The 99.72% confidence interval is marked by 3 SDs below the mean and 3 SDs above the mean.
 - Within this range of scores, we are confident that 99.72% of the cases can be found.



Sampling Distribution

- Imagine that we draw a sample from a population over and over again
- Each sample has a mean
- The distribution of the means of these samples is the sampling distribution

Central Limit Theorem

- The sampling distribution of the means is a t-distribution and approaches normal distribution (z-distribution) as sample size increases (>30), even if the distribution of the population (from which the many samples were drawn) is not normal.
- The mean of the sampling distribution is an unbiased estimate of the population mean.
- The standard deviation of the sampling distribution is known as **Standard Error**, which is given by:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

This forms the basis of the *Confidence Interval*

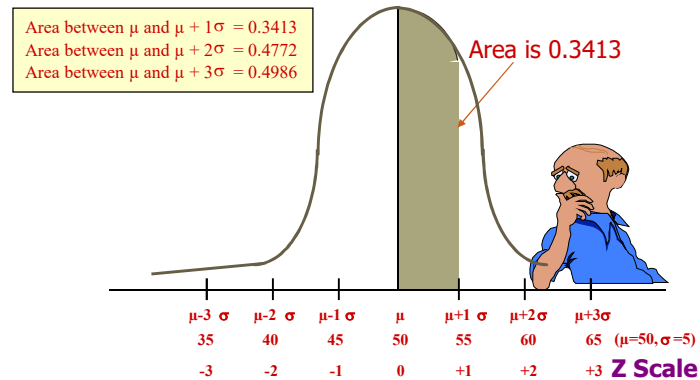
For the population:

$$z = \frac{x - \mu}{\sigma} \quad \Leftrightarrow \quad x = \mu + Z * \sigma$$

For the sampling distribution:

$$Z = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \quad \Leftrightarrow \quad \bar{X} = \mu + Z * \frac{\sigma}{\sqrt{n}}$$

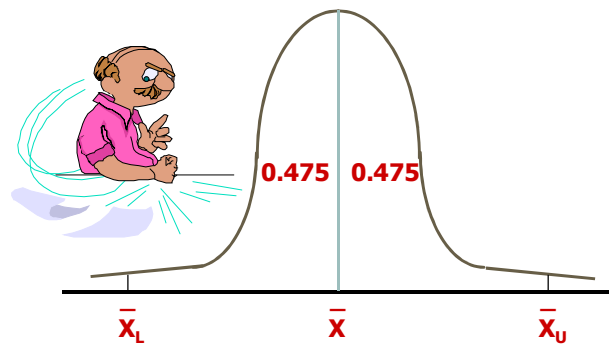
Example: Finding Probabilities Corresponding to Known Values



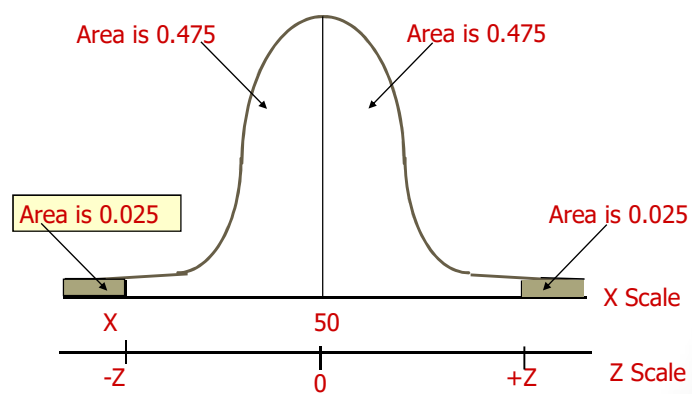
Critical Values

- Values of the test statistic that mark the **confidence interval** and the conventional levels of **statistical significance**
- E.g., 95% confidence interval on a z distribution
 - $Z = \pm 1.96$
- E.g., 99% confidence interval on a z distribution
 - $Z = \pm 2.575$

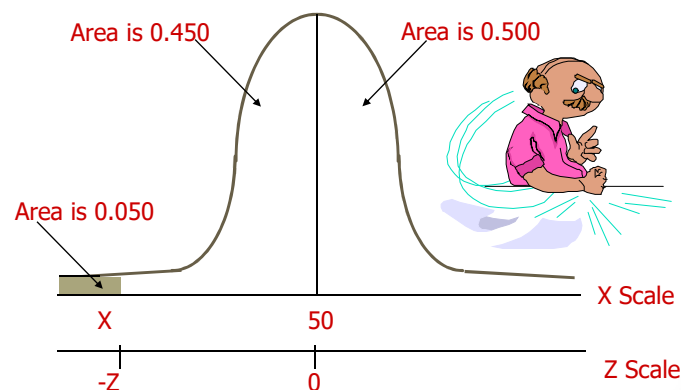
Two-Tailed Test Using the 95% Confidence Interval



95% Confidence Interval, $Z = 1.96$



One-Tailed Test Using the 95% Confidence Interval



Using Confidential Interval for Inferential Statistics

- Using the sample mean to estimate the population mean
 - Assume the sample mean is an unbiased estimate of the population mean
 - Construct a **sampling distribution around the sample mean** using the **standard error** obtained from the sample
 - There is 95% probability that the true population mean is located in the 95% confidence interval (two-tailed)
- Using the sample mean difference to estimate population mean difference
 - Assume the sample mean difference is an unbiased estimate of the population mean difference
 - Construct a **sampling distribution around the sample mean difference** using the **standard error** obtained from the sample
 - There is 95% probability that the true population mean difference is located in the 95% confidence interval (two-tailed)
 - If the 95% confidence interval covers the value of 0, there is 95% probability that the true mean difference in the population is 0
- Using the sample means to test hypotheses about mean difference (Fisher's Approach)
 - Assume the null hypothesis is true
 - Construct a **sampling distribution around 0 (the mean difference specified in the null hypothesis)** and the **standard error** obtained from the sample
 - If the test statistics (i.e., z , or t) of the sample falls in the 95% confidence interval, H_0 is supported
 - If the test statistics (i.e., z , or t) of the sample falls **OUT** of the 95% confidence interval, it is called **statistically significant**. H_0 is rejected.
 - That is, if H_0 is true, probability of finding this test statistics is less than or equal to 5%. Now this test statistics is found, H_0 is unlikely to be true.

Confidence Interval: An Example

Note that μ is estimated by $\bar{X}$. The confidence interval is given by

$$\bar{X} \pm z\sigma_{\bar{X}}$$

We can now set a 95% confidence interval around the sample mean (e.g., \$182). As a first step, we compute the standard error of the mean (e.g., standard deviation = 55):

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = 55/\sqrt{300} = 3.18$$

The central 95% of the normal distribution lies within ± 1.96 z values. The 95% confidence interval is given by

$$\begin{aligned}\bar{X} &\pm 1.96 \sigma_{\bar{X}} \\ &= 182.00 \pm 1.96(3.18) \\ &= 182.00 \pm 6.23\end{aligned}$$

Thus the 95% confidence interval ranges from \$175.77 to \$188.23. The probability of finding the true population mean to be within \$175.77 and \$188.23 is 95%.

*** Bear in Mind ***

- In inferential statistics, we work with the Sampling Distribution (usually of the means) rather than the distribution of the raw scores or the distribution of the population.
- The standard deviation of the sampling distribution of the means is the Standard Error, which is:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Inferential Statistics: Approaches to Hypothesis Testing

[83]

Fisher's Approach to Null Hypothesis Testing

- The null hypothesis represents the original state of nature; that is, nothingness
- If the null hypothesis is true, the probability of finding a test statistic out of the confidence interval is α
- α is small enough to be considered unlikely.
- If the test statistic that is found is actually out of the confidence interval, this test statistics is considered surprising.
 - If the null hypothesis is true, this test statistic is unlikely to be found.
 - If this test statistic is found, the null hypothesis is unlikely to be true.
- There is nothing about the alternative hypothesis.

Problems of Fisher's Approach

- Reversibility of inference is dubious (Cohen, 1994). That is, the syllogism "If A is true, then B is not true. If B is true, then A is not true" is correct. But the syllogism "if A is true, then B is unlikely to be true. If B is true, then A is unlikely to be true" is incorrect.
- Emphasis on H_0 is inadequate because in the first place H_0 is theoretically unimportant, and in the second place H_0 as a hypothesis involving a single real number on a random variable is doomed to be false (Cohen, 1994; Hunter, 1997).

Neyman & Pearson's Approach to Statistical Hypothesis Testing

- Alternative Hypothesis (H_1)
 - A hypothesis that states the relationship anticipated by a theory
- Procedure of statistical hypothesis testing:
 - Draw two *sampling* distributions, one for the condition in which H_0 is true, the other for the condition in which H_1 is true.
 - Considering the statistical power you want to achieve and the budget you are willing to pay, determine the α (type I error), β (type II error), and the sample size.
 - Mark a cut-off point on the dependent measure for statistical decision making. This cut-off point is determined by the α (type I error) and β (type II error). This cut-off point essentially marks the confidence interval.
 - Collect sample statistics to see whether the sample's standing is below (within the confidence interval) or above the cut-off point. If below, conclude that H_0 is true. If above, conclude that H_1 is true.
 - α is taken as part of the test, not part of the data.

Level of Significance

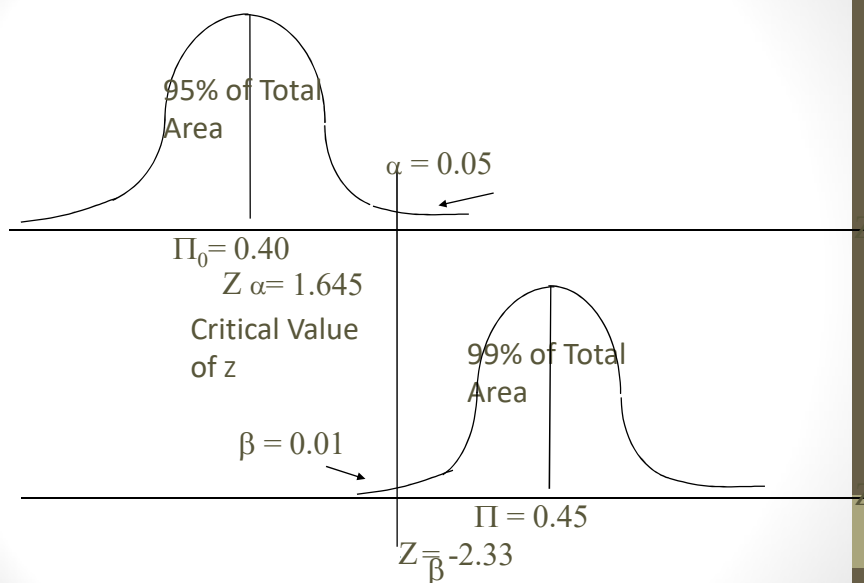
Type I Error

- **Type I error** occurs when the sample results lead to the rejection of the null hypothesis when it is in fact true.
- The probability of type I error (α) is also called the **level of significance**.
- Significance Level = 1 - Confidence Level

Type II Error

- **Type II error** occurs when, based on the sample results, the null hypothesis is not rejected when it is in fact false.
- The probability of type II error is denoted by β .
- Unlike α , which is specified by the researcher, the magnitude of β depends on the actual value of the population parameter.

Probabilities of Type I & Type II Error



Logic of Testing Alternative Hypothesis

- To determine β , one has to know the population parameter.
- Unfortunately, the population parameter is unknown.
- Even if one sets an alternative hypothesis that specifies the value of the population parameter, this hypothesis will be rejected for sure, given a sample large enough.

[89]

Power of a Test

- The **power of a test** is the probability ($1 - \beta$) of rejecting the null hypothesis when it is false and should be rejected.
- Although β is unknown, it is related to α . An extremely low value of α (e.g., $\alpha = 0.001$) will result in intolerably high β errors.
- So it is necessary to balance the two types of errors.

The Hybrid Approach (Currently in Use)

- Alternative Hypothesis (H1)
 - A hypothesis that is mutually exclusive with the null hypothesis. If one of them is rejected, the other is automatically supported.
- Problems
 - The same problems of Fisher's approach.
 - $(1 - \alpha)$ is often taken as the confidence you can have on the data, or the probability that the same sample statistics is replicable. Both interpretations are wrong.
 - The approach is mechanical and presumes that there is one single answer to the substantive. In fact, there may be many more.

P-Value

- If H_0 is true, p-value is the probability of finding a test statistic equal to or larger than the one found in your sample.
- Also referred to as the **Significance Level** of the finding

Paired Samples Test								
	Paired Differences			95% Confidence Interval of the Difference		t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error	Lower	Upper			
	Mean							
Pair 1 Neuroticism - Extroversion	-3.16667	9.74990	2.81455	-9.36146	3.02813	-1.125	11	.284

Correlations			
		Neuroticism	Extroversion
Neuroticism	Pearson Correlation	1	-.267
	Sig. (2-tailed)		.402
	N	12	12
Extroversion	Pearson Correlation	-.267	1
	Sig. (2-tailed)	.402	
	N	12	12

These statements about p-value are inaccurate

- Probability of your result being due to chance
- Probability statement telling us how likely it is that any **observed difference was due to chance alone**



Degree of Freedom (df)

- SD and variance should be interpreted as the deviation from the mean **per degree of freedom**
 - N represents the degree of freedom in the population
 - (n - 1) represents the degree of freedom in the sample
- This is because the sample variance calculation uses the sample mean to estimate the deviations, but the calculation of the sample mean has used up one degree of freedom. The denominator has to be adjusted by subtracting that degree of freedom.
- That means, the average is obtained by dividing by the degree of freedom, or the degree of randomness.
- If the population mean is known, N correctly represents the degree of freedom

Standard Deviation Formula	
Population	Sample
$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$	$s = \sqrt{\frac{\sum(X - \bar{x})^2}{n - 1}}$
X – The Value in the data distribution μ – The population Mean N – Total Number of Observations	X – The Value in the data distribution $\bar{x}$ – The Sample Mean n – Total Number of Observations

Inferential Statistics: General Procedure for Hypothesis Testing

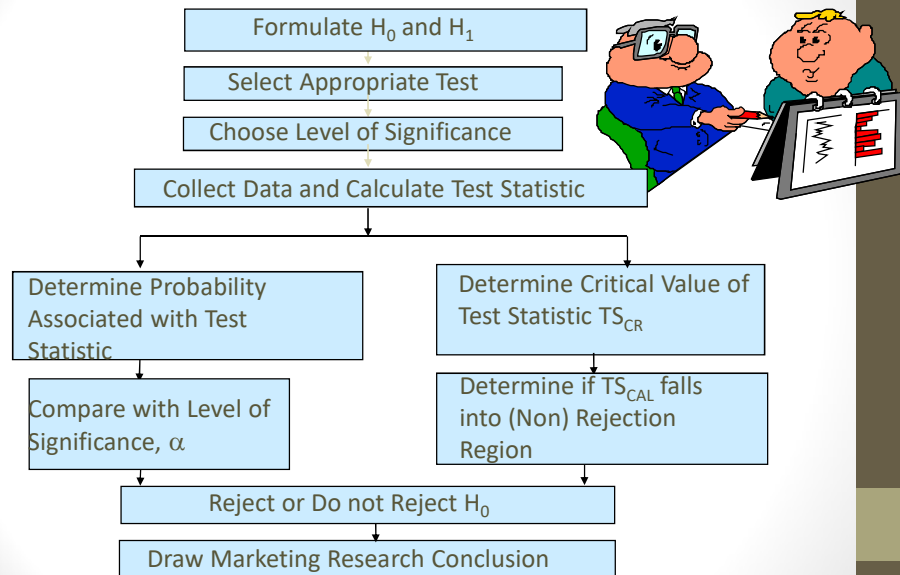
[95]

General Procedure

- Step 1: Formulate the hypothesis
- Step 2: Select an appropriate test
- Step 3: Choose a level of significance
- Step 4: Collect data and calculate test statistic
- Step 5: Determine the probability (critical value)
- Step 6: Compare the probability (critical value)
- Step 7: Make the decision (accept or reject the hypothesis)
- Step 8: Draw research conclusion

[96]

Steps Involved in Hypothesis Testing



Example: Internet Usage Data

Respondent Number	Sex	Familiarity	Internet Usage	Attitude Toward Internet	Technology	Usage of Internet Shopping	Banking
1	1.00	7.00	14.00	7.00	6.00	1.00	1.00
2	2.00	2.00	2.00	3.00	3.00	2.00	2.00
3	2.00	3.00	3.00	4.00	3.00	1.00	2.00
4	2.00	3.00	3.00	7.00	5.00	1.00	2.00
5	1.00	7.00	13.00	7.00	7.00	1.00	1.00
6	2.00	4.00	6.00	5.00	4.00	1.00	2.00
7	2.00	2.00	2.00	4.00	5.00	2.00	2.00
8	2.00	3.00	6.00	5.00	4.00	2.00	2.00
9	2.00	3.00	6.00	6.00	4.00	1.00	2.00
10	1.00	9.00	15.00	7.00	6.00	1.00	2.00
11	2.00	4.00	3.00	4.00	3.00	2.00	2.00
12	2.00	5.00	4.00	6.00	4.00	2.00	2.00
13	1.00	6.00	9.00	6.00	5.00	2.00	1.00
14	1.00	6.00	8.00	3.00	2.00	2.00	2.00
15	1.00	6.00	5.00	5.00	4.00	1.00	2.00
16	2.00	4.00	3.00	4.00	3.00	2.00	2.00
17	1.00	6.00	9.00	5.00	3.00	1.00	1.00
18	1.00	4.00	4.00	5.00	4.00	1.00	2.00
19	1.00	7.00	14.00	6.00	6.00	1.00	1.00
20	2.00	6.00	6.00	6.00	4.00	2.00	2.00
21	1.00	6.00	9.00	4.00	2.00	2.00	2.00
22	1.00	5.00	5.00	5.00	4.00	2.00	1.00
23	2.00	3.00	2.00	4.00	2.00	2.00	2.00
24	1.00	7.00	15.00	6.00	6.00	1.00	1.00
25	2.00	6.00	6.00	5.00	3.00	1.00	2.00
26	1.00	6.00	13.00	6.00	6.00	1.00	1.00
27	2.00	5.00	4.00	5.00	5.00	1.00	1.00
28	2.00	4.00	2.00	3.00	2.00	2.00	2.00
29	1.00	4.00	4.00	5.00	3.00	1.00	2.00
30	1.00	3.00	3.00	7.00	5.00	1.00	2.00

A General Procedure for Hypothesis Testing

Step 1: Formulate the Hypothesis

- A **null hypothesis** is a statement of the status quo, one of no difference or no effect. If the null hypothesis is not rejected, no changes will be made.
- An **alternative hypothesis** is one in which some difference or effect is expected. Accepting the alternative hypothesis will lead to changes in opinions or actions.
- The null hypothesis refers to a specified value of the population parameter (e.g., μ , σ , π), not a sample statistic (e.g., $\bar{X}$).

A General Procedure for Hypothesis Testing

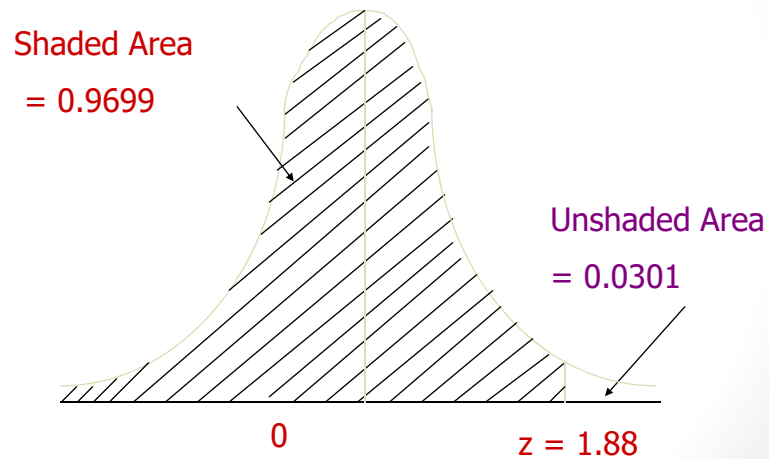
Step 1: Formulate the Hypothesis

- One-tailed test vs. two-tailed test
- The test of the null hypothesis below is a **one-tailed test**, because the alternative hypothesis is expressed directionally. If that is not the case, then a **two-tailed test** would be required, and the hypotheses would be expressed as:

$$H_0: \pi = 0.40$$

$$H_1: \pi \neq 0.40$$

Probability of z with a One-Tailed Test



A General Procedure for Hypothesis Testing Step 2: Select an Appropriate Test

- The **test statistic** measures how close the sample has come to the null hypothesis.
- The test statistic often follows a well-known distribution, such as the normal, t , or chi-square distribution.
- In our example, the **z statistic**, which follows the standard normal distribution, would be appropriate.

$$z = \frac{p - \pi}{\sigma_p}$$

where

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

Some commonly used test statistics

- z-test
 - Based on the z distribution (i.e., normal distribution)
- t-test
 - Based on the t distribution
- F-test
 - Based on the F distribution
- Chi-square test
 - Based on the chi-square distribution

A General Procedure for Hypothesis Testing Step 3: Choose a Level of Significance

P-value Interpretation	
>0.05 ($\geq$ More Than)	Insufficient evidence against null hypothesis
<0.05 ($\leq$ Less Than)	Some evidence against null hypothesis, statistically significant
<0.01	Strong evidence against null hypothesis, very statistically significant

A General Procedure for Hypothesis Testing Step 4: Collect Data and Calculate Test Statistic

- The required data are collected and the value of the test statistic computed.
- In our example, the value of the sample proportion is $\hat{p} = 17/30 = 0.567$.
- The value of $\sigma_{\hat{p}}$ can be determined as follows:

$$\begin{aligned}\sigma_{\hat{p}} &= \sqrt{\frac{\pi(1 - \pi)}{n}} \\ &= \sqrt{\frac{(0.40)(0.6)}{30}} \\ &= 0.089\end{aligned}$$

A General Procedure for Hypothesis Testing Step 4: Collect Data and Calculate Test Statistic

The test statistic z can be calculated as follows:

$$\begin{aligned}z &= \frac{\hat{p} - \pi}{\sigma_{\hat{p}}} \\ &= \frac{0.567 - 0.40}{0.089} \\ &= 1.88\end{aligned}$$

A General Procedure for Hypothesis Testing

Step 5: Determine the Probability (Critical Value)

- Using standard normal tables, the probability of obtaining a z value of 1.88 can be calculated (see Figure 15.5).
- The shaded area between $-\infty$ and 1.88 is 0.9699. Therefore, the area to the right of $z = 1.88$ is $1.0000 - 0.9699 = 0.0301$.
- Alternatively, the critical value of z , which will give an area to the right side of the critical value of 0.05, is between 1.64 and 1.65 and equals 1.645.
- Note, in determining the critical value of the test statistic, the area to the right of the critical value is either α or $\alpha/2$. It is α for a one-tail test and $\alpha/2$ for a two-tail test.

A General Procedure for Hypothesis Testing Steps 6 & 7:

Compare the Probability (Critical Value) & Making the Decision

- If the probability associated with the calculated or observed value of the test statistic (TSCAL) is less than the level of significance (α), the null hypothesis is rejected.
- The probability associated with the calculated or observed value of the test statistic is 0.0301. This is the probability of getting a p value of 0.567 when $\pi = 0.40$. This is less than the level of significance of 0.05. Hence, the null hypothesis is rejected.
- Alternatively, if the absolute calculated value of the test statistic ($|TS_{CAL}|$) is greater than the absolute critical value of the test statistic ($|TS_{CR}|$), the null hypothesis is rejected.

A General Procedure for Hypothesis Testing Steps 6 & 7:
Compare the Probability (Critical Value) & Making the Decision

- The calculated value of the test statistic $z = 1.88$ lies in the rejection region, beyond the value of 1.645. Again, the same conclusion to reject the null hypothesis is reached.
- Note that the two ways of testing the null hypothesis are equivalent but mathematically opposite in the direction of comparison.
- If the probability of $TS_{CAL} < \text{significance level } (\alpha)$ then reject H_0 but if $|TS_{CAL}| > |TS_{CR}|$ then reject H_0 .

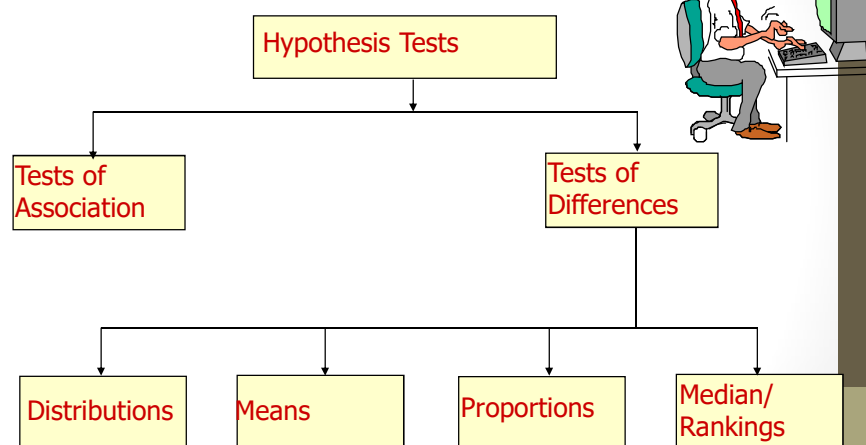
A General Procedure for Hypothesis Testing
Step 8: Draw Research Conclusion

- The conclusion reached by hypothesis testing must be expressed in terms of the research problem.
- In our example, we conclude that there is evidence that the proportion of Internet users who shop via the Internet is significantly greater than 0.40. Hence, the recommendation to the department store would be to introduce the new Internet shopping service.

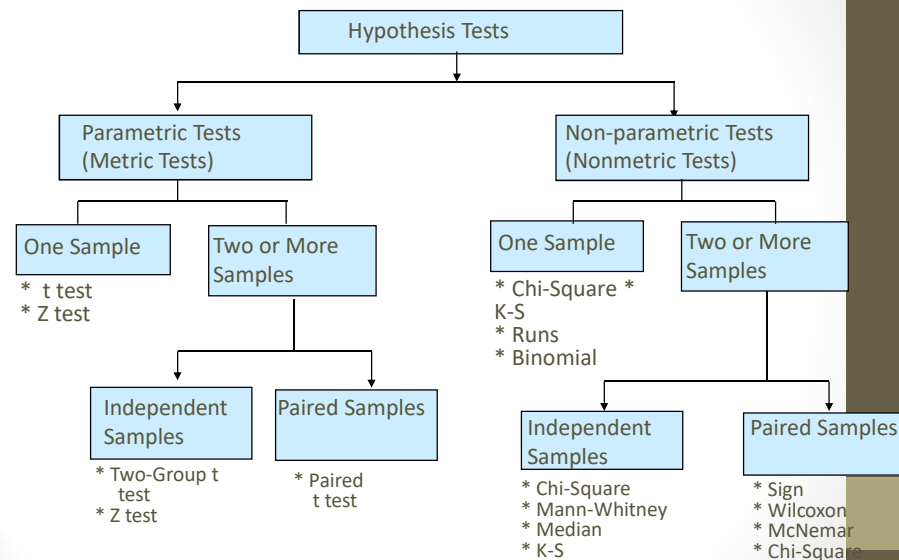
Statistical Tests

(111)

A Broad Classification of Hypothesis Tests



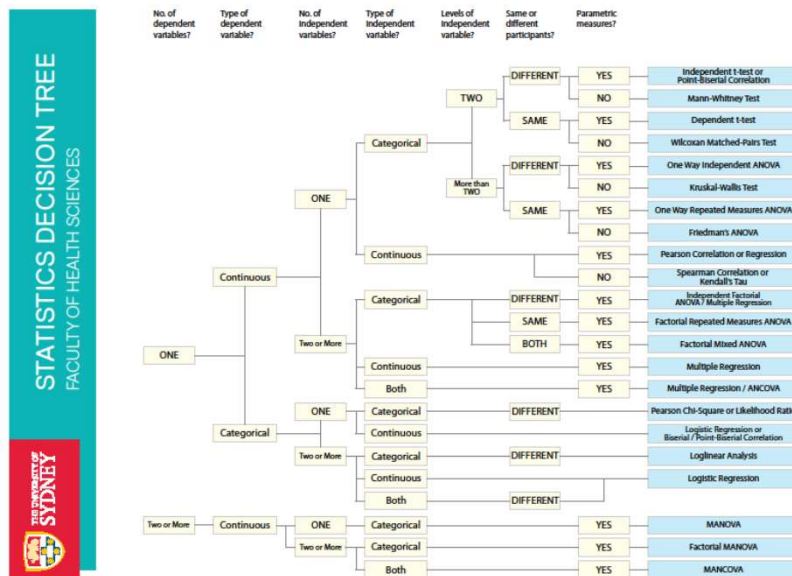
A Classification of Hypothesis Testing Procedures for Examining Differences



Hypothesis Testing Related to Differences

- **Parametric tests** assume that the variables of interest are measured on at least an interval scale.
- **Nonparametric tests** assume that the variables are measured on a nominal or ordinal scale.
- These tests can be further classified based on whether one or two or more samples are involved.
- The samples are **independent** if they are drawn randomly from different populations. For the purpose of analysis, data pertaining to different groups of respondents, e.g., males and females, are generally treated as independent samples.
- The samples are **paired** when the data for the two samples relate to the same group of respondents.

Which statistical test to use?



Statistical Tests: Z Test

z tests

- Using the z distribution (i.e., normal distribution) to construct the confidence interval
- This particular z distribution is the sampling distribution of the means
- Confidence interval of 95% = Mean $\pm$ 1.96 of the standard deviation
- The standard deviation is represented by $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$$\text{Z Test} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

- Appropriate
 - When the population SD is known
 - When sample size is large ($n \geq 30$)
 - For testing whether a raw score comes from a certain population
 - For testing proportions

Statistical Tests: Comparing Proportions

Two Independent Samples Proportions

The case involving proportions for two independent samples is also illustrated using the data of Table 15.1, which gives the number of males and females who use the Internet for shopping. Is the proportion of respondents using the Internet for shopping the same for males and females?

The null and alternative hypotheses are:

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 \neq \pi_2$$

A Z test is used as in testing the proportion for one sample. However, in this case the test statistic is given by:

$$Z = \frac{P_1 - P_2}{S_{P_1 - P_2}}$$

Two Independent Samples Proportions

In the test statistic, the numerator is the difference between the proportions in the two samples, P_1 and P_2 . The denominator is the standard error of the difference in the two proportions and is given by

$$S_{P_1 - P_2} = \sqrt{P(1-P) \left[\frac{1}{n_1} + \frac{1}{n_2} \right]}$$

where

$$P = \frac{n_1 P_1 + n_2 P_2}{n_1 + n_2}$$

Two Independent Samples Proportions

A significance level of $\alpha = 0.05$ is selected. Given the data of Table 15.1, the test statistic can be calculated as:

$$\begin{aligned} P_1 - P_2 &= (11/15) - (6/15) \\ &= 0.733 - 0.400 = 0.333 \end{aligned}$$

$$P = (15 \times 0.733 + 15 \times 0.4) / (15 + 15) = 0.567$$

$$S_{p1-p2} = \sqrt{0.567 \times 0.433 \left[\frac{1}{15} + \frac{1}{15} \right]} = 0.181$$

$$Z = 0.333 / 0.181 = 1.84$$

Two Independent Samples Proportions

Given a two-tail test, the area to the right of the critical value is 0.025. Hence, the critical value of the test statistic is 1.96. Since the calculated value is less than the critical value, the null hypothesis can not be rejected. Thus, the proportion of users (0.733 for males and 0.400 for females) is not significantly different for the two samples. Note that while the difference is substantial, it is not statistically significant due to the small sample sizes (15 in each group).

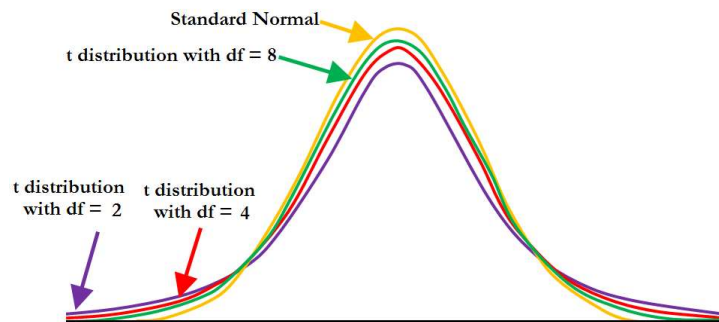
Statistical Tests: t Tests

[123]

t tests

- Using the t distribution construct the confidence interval
- The shape of the t distribution, and thus the confidence interval, depends on the degree of freedom, which is a function of the sample size ($n - 1$).
- When sample size is larger than 30, t distribution and z distribution are very similar. Z test can be used instead.
- t tests are used when
 - Sample size < 30
 - Population variance is not known

t Distribution



t test vs z test

- t test and z test basically are the same
- The **t distribution** is used to obtain the percent of cases within the confidence interval marked by the t score, whereas the **z distribution** is used to obtain the percent of cases within the confidence interval marked by the z score.

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} \qquad \mathbf{Z\ Test} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

What are t-tests?

- Used for comparing two groups/conditions
- Does group x differ from group y on an outcome variable

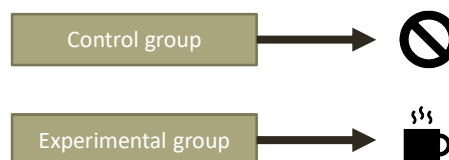
Example

Does caffeine impact IQ scores?



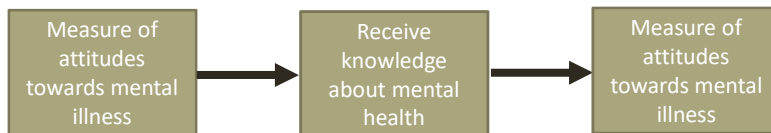
Independent vs. related samples

- Two types of t-test
 - Independent samples (different people in the groups)
 - Also known as unrelated OR between-subjects



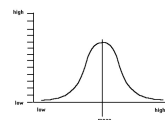
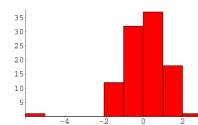
Independent vs. related samples

- Two types of t-test
 - Related samples (same people in the groups)
 - Also known as dependent or within-subjects



Assumptions for independent t-tests

- No significant outliers
- Dependent variable is “approximately normally distributed” for each category of your independent variable
- There are homogeneity (equality) of variances



Two Independent Samples Means

- In the case of means for two independent samples, the hypotheses take the following form.

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

- The two populations are sampled and the means and variances computed based on samples of sizes n_1 and n_2 . If both populations are found to have the same variance, a pooled variance estimate is computed from the two sample variances as follows:

$$s^2 = \frac{\sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2 + \sum_{i=1}^{n_2} (X_{i2} - \bar{X}_2)^2}{n_1 + n_2 - 2} \quad \text{or} \quad s^2 = \frac{(n_1 - 1) s_1^2 + (n_2 - 1) s_2^2}{n_1 + n_2 - 2}$$

Two Independent Samples Means

The standard deviation of the test statistic can be estimated as:

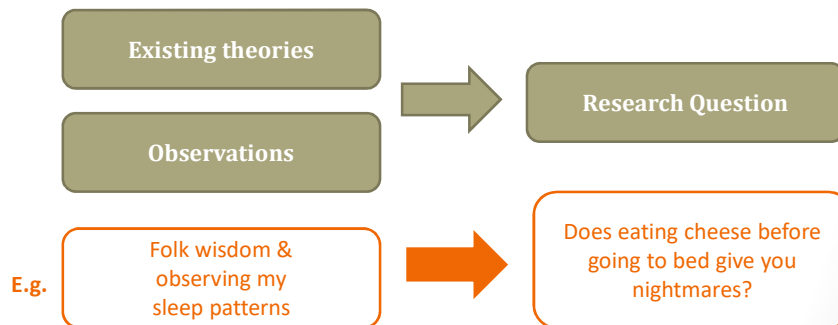
$$s_{\bar{X}_1 - \bar{X}_2} = \sqrt{s^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}$$

The appropriate value of t can be calculated as:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{\bar{X}_1 - \bar{X}_2}}$$

The degrees of freedom in this case are $(n_1 + n_2 - 2)$.

Independent-Samples T-test: Example



How can we find the answer?

Experiment

A powerful technique for investigating links between two variables

Cheese
condition

vs.

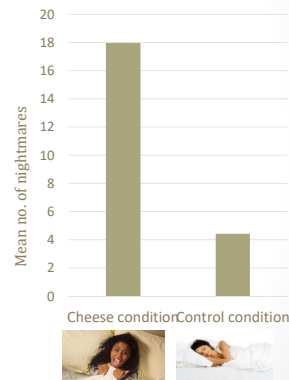
Control
condition



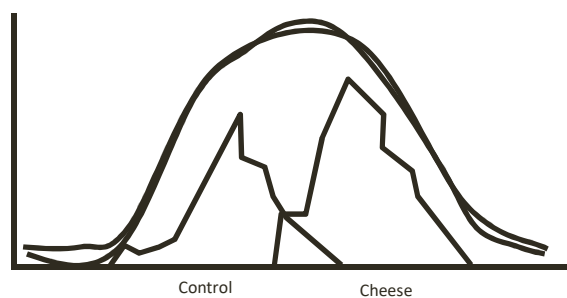
Record the number of
nightmares

Results of an experiment

- We have a difference in means
 - How could you interpret this?
 - Is there a 'real' difference between the means?
 - OR, is it a sampling error or due to chance?

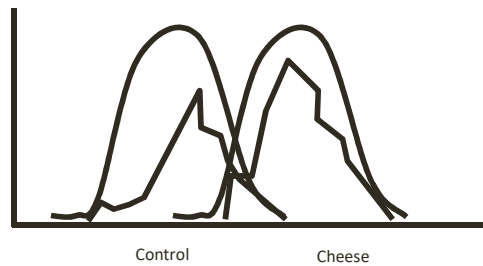


Results of the experiment



One possibility: Difference is due to sampling variability (i.e. null hypothesis is true and two sets of scores come from the same population)

Results of the experiment

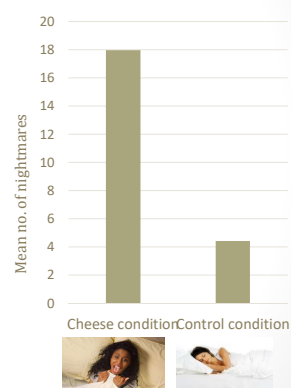


OR: It may be that our difference reflects a **real difference** between the two populations

Results of the experiment

T-test

- How likely is it that the difference we found between the two groups could be the result of chance or error?
- If that likelihood is low then we say we have a **real difference between the groups**



Group statistics table

Number of participants

Mean (group average)

Standard error (how representative the sample is likely to be of the population)

SD (dispersion from average)

DV

IV (2 conditions/groups)

Group	N	Mean	Std. Deviation	Std. Error Mean
Nightmares cheese	50	17.9800	1.60979	.22766
Nightmares no cheese	50	4.4400	.95105	.13450

Independent samples table

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
Nightmares	Equal variances assumed	6.610	.012	51.206	98	.000	13.54000	.26442	13.01527	14.06473
	Equal variances not assumed			51.206	79.491	.000	13.54000	.26442	13.01373	14.06627

F	Sig.
6.610	.012

- One of assumptions for t-tests is quality (homogeneity) of variances
- Test using Levene's test
- Produced automatically in SPSS

Levene's test

- Levene's test asks: **are the variances in the different groups equal?**
- Significant result ($p \leq .05$) = variances are *significantly different* from equal (e.g. NOT equal variances)

Levene's Test for Equality of Variances	
F	Sig.
6.610	.012

Q - Does this mean I have to conduct a non-parametric test?

A - No . . . ! You simply need to interpret the results of your **SPSS output** in a different way

Levene's test

Levene's test = non-significant result = assumption *not* violated = read top line

Independent Samples Test										
		Levene's Test for Equality of Variances			t-test for Equality of Means					
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error	95% Confidence Interval of the Difference	
									Lower	Upper
Numbers	Equal variances assumed	6.610	.012	24.000	98	.000	13.54000	.26442	13.01527	14.06473
	Equal variances not assumed			31.200	79.491	.000	13.54000	.26442	13.01373	14.06627

Levene's test = Significant result = assumption violated = read bottom line

Independent samples table

P-value – what is the probability that the difference was obtained due to chance? *Is the difference statistically significant?*

t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
					Lower	Upper
51.206	38	.000	13.34000	.26442	13.01327	14.00473
51.206	79.491	.000	13.54000	.26442	13.01373	14.06627

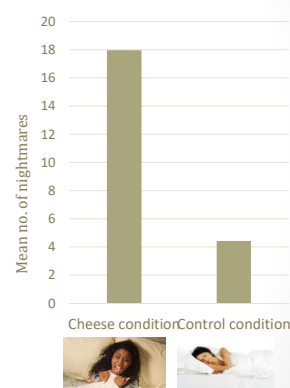
t-statistic = is there a difference between the groups?

df = number of individual scores than can vary without changing the sample mean

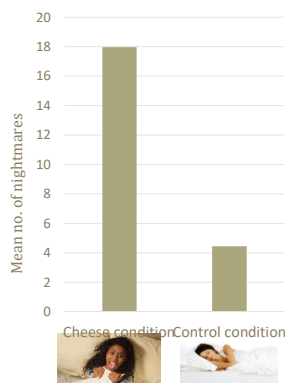
- Value between -1 and +1 = no difference
- Value between less than -2 and more than +2 = a reliable difference between the groups

Statistical significance and our hypothesis

$P < .05$ = the likelihood of the obtained difference in number of nightmares arising by chance is lower than 5 times out of 100 (or 1 in 20), if there really were no difference in population



Interpreting your results

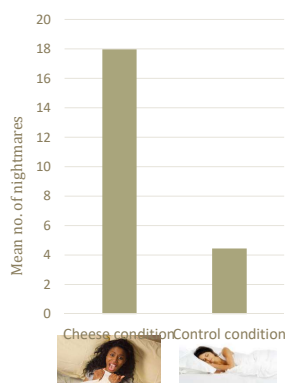


Null hypothesis: Individuals in the cheese condition will have **about the same number of nightmares** as those in the control condition

Accept or reject ?

$t(79.49) = 51.21, p < .001$

Interpreting your results



Null hypothesis: Individuals in the cheese condition will have **about the same number of nightmares** as those in the control condition

Experimental: Individuals in the cheese condition **will experience more nightmares** than those in the control condition

Reporting results: APA style

The study found that participants who ate cheese just before bed had significantly more nightmares ($M = 17.98, SD = 1.61$) than control participants who did not eat any cheese before bed ($M = 4.44, SD = .95$), $t(79.49) = 51.21, p < .001$

Report **descriptive statistics** for both groups (Mean and SD)

Report **inferential statistics** (i.e. results of t-test)

Interpretation

- What do these results mean?
- What could you interpret from these findings?

The study found that participants who ate cheese just before bed had significantly more nightmares ($M = 17.98, SD = 1.61$) than control participants who did not eat any cheese before bed ($M = 4.44, SD = .95$), $t(79.49) = 51.21, p < .001$

Paired Samples

The difference in these cases is examined by a **paired samples t test**. To compute t for paired samples, the paired difference variable, denoted by D , is formed and its mean and variance calculated. Then the t statistic is computed. The degrees of freedom are $n - 1$, where n is the number of pairs. The relevant formulas are:

$$H_0: \mu_D = 0$$

$$H_1: \mu_D \neq 0$$

$$t_{n-1} = \frac{\bar{D} - \mu_D}{\frac{s_D}{\sqrt{n}}}$$

continued...

Paired Samples

Where:

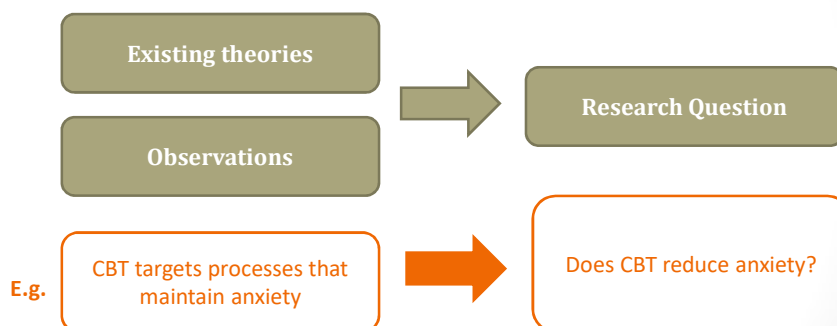
$$\bar{D} = \frac{\sum_{i=1}^n D_i}{n}$$

$$s_D = \sqrt{\frac{\sum_{i=1}^n (D_i - \bar{D})^2}{n - 1}}$$

$$S_{\bar{D}} = \frac{s_D}{\sqrt{n}}$$

In the Internet usage example (Table 15.1), a paired t test could be used to determine if the respondents differed in their attitude toward the Internet and attitude toward technology. The resulting output is shown in Table 15.15.

Related-Samples T-test: Example



How can we find the answer?



Our experimental hypothesis

- What would we expect to happen in our experiment if our idea that CBT reduces anxiety **is correct**?

Pre CBT

Vs.

Post CBT



- Individuals **will be significantly less anxious** after taking part in a CBT intervention

Our null hypothesis

- What would we expect to happen in our experiment if our idea that CBT reduces anxiety **is wrong** and CBT is not an effective treatment?

Pre CBT

Vs.

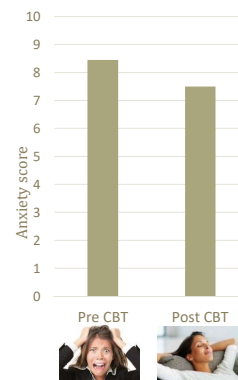
Post CBT



- Individuals **anxiety levels will be about the same** before and after taking part in a CBT intervention

Results of the experiment

- We have some difference in means
 - How could you interpret this?
 - Is there a 'real' difference between the means?
 - OR, is it a sampling error or due to chance?



Descriptive statistics

Mean (group average)

Number of participants

Standard error (how representative the sample is likely to be of the population)

SD (dispersion from average)

2 groups

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Pre_CBT	8.4500	20	1.09904	.24575
	Post_CBT	7.5000	20	1.10024	.24602

Independent samples table

P-value – what is the probability that the difference was obtained due to chance? *Is the difference statistically significant?*

Paired Samples Test									
		Paired Differences				t	df	Sig. (2-tailed)	
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower				Upper
Pair 1	Pre_CBT - Post_CBT	95000	1.66938	.37329	.16870	1.73130	2.545	.020	

t-statistic = is there a difference between the groups?

- How big is this difference?

df = number of individual scores than can vary without changing the sample mean

Reporting results: APA style

Participants levels of anxiety fell significantly after taking part in a CBT intervention (Pre-CBT: $M = 8.45$, $SD = 1.10$ vs. Post-CBT: $M = 7.50$, $SD = 1.10$), $t(19) = 2.55$, $p < .05$.

Report **descriptive statistics** for both groups (Mean and SD)

Report **inferential statistics** (i.e. results of t-test)

Reporting the t-test results

Confirming you've
used a t-test

t-value

Notation used with p values:

$<$ = less than
 $>$ = greater than
 $\leq$ = less than or equal to
 $\geq$ = greater than or equal to

$t(19) = 2.55, p < .05$

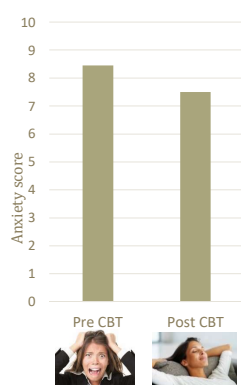
Degrees of
freedom (df)

p-value

If non-sig results:

p = exact p value
 e.g. " $p = .07$ "

Interpreting your results

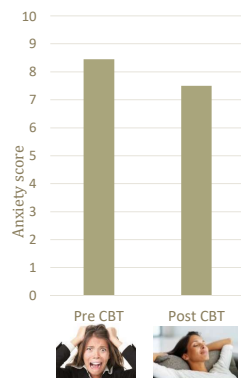


Null hypothesis: Individuals
**anxiety levels will be about the
 same** before and after taking
 part in a CBT intervention

Accept or reject ?

$t(19) = 2.55, p < .05$

Interpreting your results



Null hypothesis: Individuals **anxiety level will be about the same** before and after taking part in a CBT intervention

Experimental: Individuals will be **significantly less anxious** after taking part in a CBT intervention

Interpreting results

- The results of the t-test only indicate whether the differences were likely to be due to chance or not
- NOT whether this change is meaningful or important
- We still need to think about what these results mean



Interpretation

- What do these results mean?
- What could you interpret from these findings?

Participants levels of anxiety fell significantly after taking part in a CBT intervention (Pre-CBT: $M = 8.45$, $SD = 1.10$ vs. Post-CBT: $M = 7.50$, $SD = 1.10$), $t(19) = 2.55$, $p < .05$.

Statistical Tests: Analysis of Variance (ANOVA)

[164]

Two Independent Samples F Test

An **F test** of sample variance may be performed if it is not known whether the two populations have equal variance. In this case, the hypotheses are:

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

Two Independent Samples F Statistic

The **F statistic** is computed from the sample variances as follows:

$$\text{where } F_{(n_1-1),(n_2-1)} = \frac{s_1^2}{s_2^2}$$

n_1 = size of sample 1

n_2 = size of sample 2

n_1-1 = degrees of freedom for sample 1

n_2-1 = degrees of freedom for sample 2

s_1^2 = sample variance for sample 1

s_2^2 = sample variance for sample 2

Using the data of Table 15.1, suppose we wanted to determine whether Internet usage was different for males as compared to females. A two-independent-samples *t* test was conducted. The results are presented in Table 15.14.

One-Way Analysis of Variance

Researchers are often interested in examining the differences in the mean values of the dependent variable for several categories of a single independent variable or factor. For example:

- Do the various segments differ in terms of their volume of product consumption?
- Do the brand evaluations of groups exposed to different commercials vary?
- What is the effect of consumers' familiarity with the store (measured as high, medium, and low) on preference for the store?

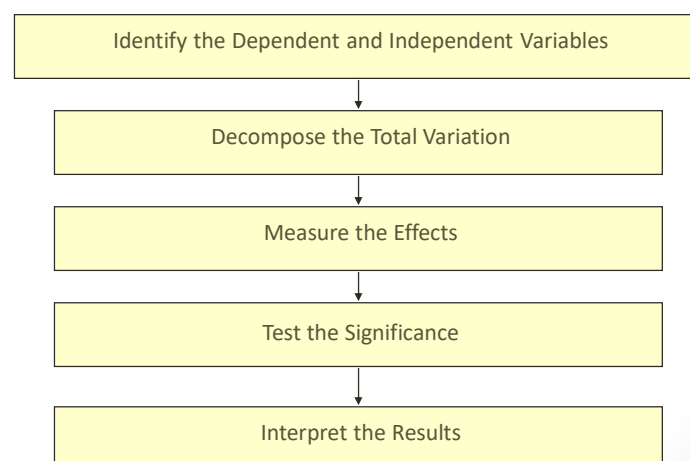
Statistics Associated with One-Way Analysis of Variance

- **η^2 (eta squared)**. The strength of the effects of X (independent variable or factor) on Y (dependent variable) is measured by **η^2 (eta squared)**. The value of η^2 varies between 0 and 1.
- **F statistic**. The null hypothesis that the category means are equal in the population is tested by an **F statistic** based on the ratio of mean square related to X and mean square related to error.
- **Mean square**. This is the sum of squares divided by the appropriate degrees of freedom.

Statistics Associated with One-Way Analysis of Variance

- SS_{between} . Also denoted as SS_X , this is the variation in Y related to the variation in the means of the categories of X . This represents variation between the categories of X , or the portion of the sum of squares in Y related to X .
- SS_{within} . Also referred to as SS_{error} , this is the variation in Y due to the variation within each of the categories of X . This variation is not accounted for by X .
- SS_Y . This is the total variation in Y .

Conducting One-Way ANOVA



Conducting One-Way Analysis of Variance Decompose the Total Variation

The total variation in Y , denoted by SS_y , can be decomposed into two components:

$$SS_y = SS_{between} + SS_{within}$$

where the subscripts *between* and *within* refer to the categories of X . $SS_{between}$ is the variation in Y related to the variation in the means of the categories of X . For this reason, $SS_{between}$ is also denoted as SS_x . SS_{within} is the variation in Y related to the variation within each category of X . SS_{within} is not accounted for by X . Therefore it is referred to as SS_{error} .

Conducting One-Way Analysis of Variance Decompose the Total Variation

The total variation in Y may be decomposed as:

$$SS_y = SS_x + SS_{error}$$

Where $SS_y = \sum_{i=1}^N (Y_i - \bar{Y})^2$

$$SS_x = \sum_{j=1}^c n (\bar{Y}_j - \bar{Y})^2$$

$$SS_{error} = \sum_{j=1}^c \sum_{i=1}^n (Y_{ij} - \bar{Y}_j)^2$$

Y_i = individual observation

$\bar{Y}_j$ = mean for category j

$\bar{Y}$ = mean over the whole sample, or grand mean

Y_{ij} = i th observation in the j th category

Decomposition of the Total Variation: One-Way ANOVA

Independent Variable					X
Categories					Total Sample
X_1	X_2	X_3	...	X_c	
Y_1	Y_1	Y_1		Y_1	Y_1
Y_2	Y_2	Y_2		Y_2	Y_2
$\vdots$					$\vdots$
Y_n	Y_n	Y_n		Y_n	Y_N
$\bar{Y}_1$	$\bar{Y}_2$	$\bar{Y}_3$		$\bar{Y}_c$	$\bar{Y}$

Within Category Variation = SS_{within}
 Category Mean
 Between Category Variation = $SS_{between}$
 Total Variation = SS_y

Conducting One-Way Analysis of Variance

In analysis of variance, we estimate two measures of variation: within groups (SS_{within}) and between groups ($SS_{between}$). Thus, by comparing the Y variance estimates based on between-group and within-group variation, we can test the null hypothesis.

Measure the Effects

The strength of the effects of X on Y are measured as follows:

$$\eta^2 = SS_x / SS_y = (SS_y - SS_{error}) / SS_y$$

The value of η^2 varies between 0 and 1.

Conducting One-Way Analysis of Variance Test of Significance

In one-way analysis of variance, the interest lies in testing the null hypothesis that the category means are equal in the population.

$$H_0: \mu_1 = \mu_2 = \mu_3 = \dots = \mu_c$$

Under the null hypothesis, SS_x and SS_{error} come from the same source of variation. In other words, the estimate of the population variance of Y ,

$$\begin{aligned} &= SS_x / (c - 1) \\ &= \text{Mean square due to } X \\ S_y^2 &= MS_x \end{aligned}$$

or

$$\begin{aligned} &= SS_{error} / (N - c) \\ S_y^2 &= \text{Mean square due to error} \\ &= MS_{error} \end{aligned}$$

Conducting One-Way Analysis of Variance Test of Significance

The null hypothesis may be tested by the **F statistic** based on the ratio between these two estimates:

$$F = \frac{SS_x / (c - 1)}{SS_{error} / (N - c)} = \frac{MS_x}{MS_{error}}$$

This statistic follows the **F distribution**, with $(c - 1)$ and $(N - c)$ degrees of freedom (df).

Conducting One-Way Analysis of Variance Interpret the Results

- If the null hypothesis of equal category means is not rejected, then the independent variable does not have a significant effect on the dependent variable.
- On the other hand, if the null hypothesis is rejected, then the effect of the independent variable is significant.
- A comparison of the category mean values will indicate the nature of the effect of the independent variable.

Illustrative Applications of One-Way Analysis of Variance

We illustrate the concepts discussed in this chapter using the data presented in Table 16.2.

The department store is attempting to determine the effect of in-store promotion (X) on sales (Y). For the purpose of illustrating hand calculations, the data of Table 16.2 are transformed in Table 16.3 to show the store sales (Y_{ij}) for each level of promotion.

The null hypothesis is that the category means are equal:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

Effect of Promotion and Clientele on Sales

Table 16.2

Store Number	Coupon Level	In-Store Promotion	Sales	Clientele Rating
1	1.00	1.00	10.00	9.00
2	1.00	1.00	9.00	10.00
3	1.00	1.00	10.00	8.00
4	1.00	1.00	8.00	4.00
5	1.00	1.00	9.00	6.00
6	1.00	2.00	8.00	8.00
7	1.00	2.00	8.00	4.00
8	1.00	2.00	7.00	10.00
9	1.00	2.00	9.00	6.00
10	1.00	2.00	6.00	9.00
11	1.00	3.00	5.00	8.00
12	1.00	3.00	7.00	9.00
13	1.00	3.00	6.00	6.00
14	1.00	3.00	4.00	10.00
15	1.00	3.00	5.00	4.00
16	2.00	1.00	8.00	10.00
17	2.00	1.00	9.00	6.00
18	2.00	1.00	7.00	8.00
19	2.00	1.00	7.00	4.00
20	2.00	1.00	6.00	9.00
21	2.00	2.00	4.00	6.00
22	2.00	2.00	5.00	8.00
23	2.00	2.00	5.00	10.00
24	2.00	2.00	6.00	4.00
25	2.00	2.00	4.00	9.00
26	2.00	3.00	2.00	4.00
27	2.00	3.00	3.00	6.00
28	2.00	3.00	2.00	10.00
29	2.00	3.00	1.00	9.00
30	2.00	3.00	2.00	8.00



Illustrative Applications of One-Way Analysis of Variance

Table 16.3

EFFECT OF IN-STORE PROMOTION ON SALES

Store No.	Level of In-store Promotion		
	High	Medium	Low
	Normalized Sales		
1	10	8	5
2	9	8	7
3	10	7	6
4	8	9	4
5	9	6	5
6	8	4	2
7	9	5	3
8	7	5	2
9	7	6	1
10	6	4	2
Column Totals	83	62	37
Category means: $\bar{y}_j$	83/10 = 8.3	62/10 = 6.2	37/10 = 3.7
Grand mean, $\bar{Y}$	= (83 + 62 + 37)/30 = 6.067		

Illustrative Applications of One-Way Analysis of Variance

To test the null hypothesis, the various sums of squares are computed as follows:

$$\begin{aligned}
 SS_y &= (10-6.067)^2 + (9-6.067)^2 + (10-6.067)^2 + (8-6.067)^2 + (9-6.067)^2 \\
 &+ (8-6.067)^2 + (9-6.067)^2 + (7-6.067)^2 + (7-6.067)^2 + (6-6.067)^2 \\
 &+ (8-6.067)^2 + (8-6.067)^2 + (7-6.067)^2 + (9-6.067)^2 + (6-6.067)^2 \\
 &+ (4-6.067)^2 + (5-6.067)^2 + (5-6.067)^2 + (6-6.067)^2 + (4-6.067)^2 \\
 &+ (5-6.067)^2 + (7-6.067)^2 + (6-6.067)^2 + (4-6.067)^2 + (5-6.067)^2 \\
 &+ (2-6.067)^2 + (3-6.067)^2 + (2-6.067)^2 + (1-6.067)^2 + (2-6.067)^2 \\
 &= (3.933)^2 + (2.933)^2 + (3.933)^2 + (1.933)^2 + (2.933)^2 \\
 &+ (1.933)^2 + (2.933)^2 + (0.933)^2 + (0.933)^2 + (-0.067)^2 \\
 &+ (1.933)^2 + (1.933)^2 + (0.933)^2 + (2.933)^2 + (-0.067)^2 \\
 &+ (-2.067)^2 + (-1.067)^2 + (-1.067)^2 + (-0.067)^2 + (-2.067)^2 \\
 &+ (-1.067)^2 + (0.933)^2 + (-0.067)^2 + (-2.067)^2 + (-1.067)^2 \\
 &+ (-4.067)^2 + (-3.067)^2 + (-4.067)^2 + (-5.067)^2 + (-4.067)^2 \\
 &= 185.867
 \end{aligned}$$

Illustrative Applications of One-Way Analysis of Variance

$$\begin{aligned}
 SS_x &= 10(8.3-6.067)^2 + 10(6.2-6.067)^2 + 10(3.7-6.067)^2 \\
 &= 10(2.233)^2 + 10(0.133)^2 + 10(-2.367)^2 \\
 &= 106.067
 \end{aligned}$$

$$\begin{aligned}
 SS_{error} &= (10-8.3)^2 + (9-8.3)^2 + (10-8.3)^2 + (8-8.3)^2 + (9-8.3)^2 \\
 &+ (8-8.3)^2 + (9-8.3)^2 + (7-8.3)^2 + (7-8.3)^2 + (6-8.3)^2 \\
 &+ (8-6.2)^2 + (8-6.2)^2 + (7-6.2)^2 + (9-6.2)^2 + (6-6.2)^2 \\
 &+ (4-6.2)^2 + (5-6.2)^2 + (5-6.2)^2 + (6-6.2)^2 + (4-6.2)^2 \\
 &+ (5-3.7)^2 + (7-3.7)^2 + (6-3.7)^2 + (4-3.7)^2 + (5-3.7)^2 \\
 &+ (2-3.7)^2 + (3-3.7)^2 + (2-3.7)^2 + (1-3.7)^2 + (2-3.7)^2 \\
 &= (1.7)^2 + (0.7)^2 + (1.7)^2 + (-0.3)^2 + (0.7)^2 \\
 &+ (-0.3)^2 + (0.7)^2 + (-1.3)^2 + (-1.3)^2 + (-2.3)^2 \\
 &+ (1.8)^2 + (1.8)^2 + (0.8)^2 + (2.8)^2 + (-0.2)^2 \\
 &+ (-2.2)^2 + (-1.2)^2 + (-1.2)^2 + (-0.2)^2 + (-2.2)^2 \\
 &+ (1.3)^2 + (3.3)^2 + (2.3)^2 + (0.3)^2 + (1.3)^2 \\
 &+ (-1.7)^2 + (-0.7)^2 + (-1.7)^2 + (-2.7)^2 + (-1.7)^2 \\
 &= 79.80
 \end{aligned}$$

Illustrative Applications of One-Way Analysis of Variance

It can be verified that

$$SS_y = SS_x + SS_{error}$$

as follows:

$$185.867 = 106.067 + 79.80$$

The strength of the effects of X on Y are measured as follows:

$$\begin{aligned}\eta^2 &= SS_x / SS_y \\ &= 106.067 / 185.867 \\ &= 0.571\end{aligned}$$

In other words, 57.1% of the variation in sales (Y) is accounted for by in-store promotion (X), indicating a modest effect. The null hypothesis may now be tested.

$$F = \frac{SS_x / (c - 1)}{SS_{error} / (N - c)} = \frac{MS_x}{MS_{error}}$$

$$\begin{aligned}F &= \frac{106.067 / (3 - 1)}{79.800 / (30 - 3)} \\ &= 17.944\end{aligned}$$

Illustrative Applications of One-Way Analysis of Variance

- From Table 5 in the Statistical Appendix we see that for 2 and 27 degrees of freedom, the critical value of F is 3.35 for $\alpha = 0.05$. Because the calculated value of F is greater than the critical value, we reject the null hypothesis.
- We now illustrate the analysis of variance procedure using a computer program. The results of conducting the same analysis by computer are presented in Table 16.4.

Assumptions in Analysis of Variance

The salient assumptions in analysis of variance can be summarized as follows:

1. Ordinarily, the categories of the independent variable are assumed to be fixed. Inferences are made only to the specific categories considered. This is referred to as the **fixed-effects model**.
2. The error term is **normally** distributed, with a **zero mean** and a **constant variance**. The error is not related to any of the categories of X .
3. The error terms are **uncorrelated**. If the error terms are correlated (i.e., the observations are not independent), the F ratio can be seriously distorted.

N-Way Analysis of Variance

In marketing research, one is often concerned with the effect of more than one factor simultaneously. For example:

- How do advertising levels (high, medium, and low) interact with price levels (high, medium, and low) to influence a brand's sale?
- Do educational levels (less than high school, high school graduate, some college, and college graduate) and age (less than 35, 35-55, more than 55) affect consumption of a brand?
- What is the effect of consumers' familiarity with a department store (high, medium, and low) and store image (positive, neutral, and negative) on preference for the store?

N-Way Analysis of Variance

Consider the simple case of two factors X_1 and X_2 having categories c_1 and c_2 . The total variation in this case is partitioned as follows:

$$SS_{total} = SS \text{ due to } X_1 + SS \text{ due to } X_2 + SS \text{ due to interaction of } X_1 \text{ and } X_2 + SS_{within}$$

or

$$SS_y = SS_{x_1} + SS_{x_2} + SS_{x_1x_2} + SS_{error}$$

The strength of the joint effect of two factors, called the overall effect, or **multiple η^2** , is measured as follows:

$$\text{multiple } \eta^2 = (SS_{x_1} + SS_{x_2} + SS_{x_1x_2}) / SS_y$$

N-Way Analysis of Variance

The **significance of the overall effect** may be tested by an F test, as follows:

$$\begin{aligned} F &= \frac{(SS_{x_1} + SS_{x_2} + SS_{x_1x_2}) / df_n}{SS_{error} / df_d} \\ &= \frac{SS_{x_1, x_2, x_1x_2} / df_n}{SS_{error} / df_d} \\ &= \frac{MS_{x_1, x_2, x_1x_2}}{MS_{error}} \end{aligned}$$

where

df_n	=	degrees of freedom for the numerator
	=	$(c_1 - 1) + (c_2 - 1) + (c_1 - 1)(c_2 - 1)$
	=	$c_1c_2 - 1$
df_d	=	degrees of freedom for the denominator
	=	$N - c_1c_2$
MS	=	mean square

N-Way Analysis of Variance

If the overall effect is significant, the next step is to examine the **significance of the interaction effect**. Under the null hypothesis of no interaction, the appropriate F test is:

$$F = \frac{SS_{x_1x_2}/df_n}{SS_{error}/df_d}$$

$$= \frac{MS_{x_1x_2}}{MS_{error}}$$

Where

$$df_n = (c_1 - 1)(c_2 - 1)$$

$$df_d = N - c_1c_2$$

N-Way Analysis of Variance

The **significance of the main effect of each factor** may be tested as follows for X_1 :

$$F = \frac{SS_{x_1}/df_n}{SS_{error}/df_d}$$

$$= \frac{MS_{x_1}}{MS_{error}}$$

where

$$df_n = c_1 - 1$$

$$df_d = N - c_1c_2$$

Two-Way Analysis of Variance

Table 16.5

Source of Variation	Sum of squares	df	Mean square	F	Sig. of F	ω^2
Main Effects						
Promotion	106.067	2	53.033	54.862	0.000	0.557
Coupon	53.333	1	53.333	55.172	0.000	0.280
Combined	159.400	3	53.133	54.966	0.000	
Two-way interaction	3.267	2	1.633	1.690	0.226	
Model	162.667	5	32.533	33.655	0.000	
Residual (error)	23.200	24	0.967			
TOTAL	185.867	29	6.409			

Two-Way Analysis of Variance

Table 16.5, cont.

Cell Means

Promotion	Coupon	Count	Mean
High	Yes	5	9.200
High	No	5	7.400
Medium	Yes	5	7.600
Medium	No	5	4.800
Low	Yes	5	5.400
Low	No	5	2.000

TOTAL

30

Factor Level Means

Promotion	Coupon	Count	Mean
High		10	8.300
Medium		10	6.200
Low		10	3.700
	Yes	15	7.400
	No	15	4.733
Grand Mean		30	6.067



Analysis of Covariance

When examining the differences in the mean values of the dependent variable related to the effect of the controlled independent variables, it is often necessary to take into account the influence of uncontrolled independent variables. For example:

- In determining how different groups exposed to different commercials evaluate a brand, it may be necessary to control for prior knowledge.
- In determining how different price levels will affect a household's cereal consumption, it may be essential to take household size into account. We again use the data of Table 16.2 to illustrate analysis of covariance.
- Suppose that we wanted to determine the effect of in-store promotion and couponing on sales while controlling for the effect of clientele. The results are shown in Table 16.6.

Analysis of Covariance

Table 16.6

Source of Variation	Sum of Squares	df	Mean Square	F	Sig. of F
Covariance					
Clientele	0.838	1	0.838	0.862	0.363
Main effects					
Promotion	106.067	2	53.033	54.546	0.000
Coupon	53.333	1	53.333	54.855	0.000
Combined	159.400	3	53.133	54.649	0.000
2-Way Interaction					
Promotion* Coupon	3.267	2	1.633	1.680	0.208
Model	163.505	6	27.251	28.028	0.000
Residual (Error)	22.362	23	0.972		
TOTAL	185.867	29	6.409		
Covariate	Raw Coefficient				
Clientele	-0.078				

Issues in Interpretation

Important issues involved in the interpretation of ANOVA results include interactions, relative importance of factors, and multiple comparisons.

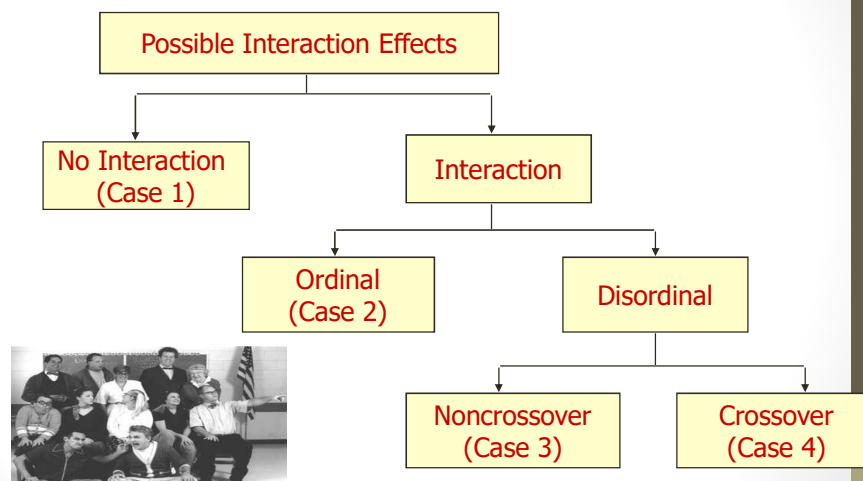
Interactions

- The different interactions that can arise when conducting ANOVA on two or more factors are shown in Figure 16.3.

Relative Importance of Factors

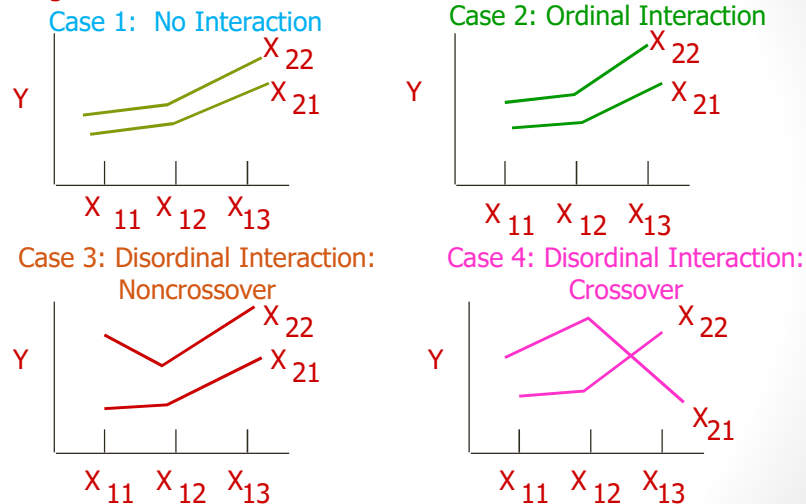
- Experimental designs are usually balanced, in that each cell contains the same number of respondents. This results in an orthogonal design in which the factors are uncorrelated. Hence, it is possible to determine unambiguously the relative importance of each factor in explaining the variation in the dependent variable.

A Classification of Interaction Effects



Patterns of Interaction

Fig. 16.4



Issues in Interpretation

- The most commonly used measure in ANOVA is **omega squared, ω^2** . This measure indicates what proportion of the variation in the dependent variable is related to a particular independent variable or factor. The relative contribution of a factor X is calculated as follows:

$$\omega_x^2 = \frac{SS_x - (df_x \times MS_{error})}{SS_{total} + MS_{error}}$$

- Normally, ω^2 is interpreted only for statistically significant effects. In Table 16.5, ω^2 associated with the level of in-store promotion is calculated as follows:

$$\begin{aligned} \omega_p^2 &= \frac{106.067 - (2 \times 0.967)}{185.867 + 0.967} \\ &= \frac{104.133}{186.834} \\ &= 0.557 \end{aligned}$$

Issues in Interpretation

Note, in Table 16.5, that

$$SS_{total} = 106.067 + 53.333 + 3.267 + 23.2 \\ = 185.867$$

Likewise, the ω^2 associated with couponing is:

$$\omega_c^2 = \frac{53.333 - (1 \times 0.967)}{185.867 + 0.967}$$

$$= \frac{52.366}{186.834}$$

$$= 0.280$$

As a guide to interpreting, a large experimental effect produces an index of 0.15 or greater, a medium effect produces an index of around 0.06, and a small effect produces an index of 0.01. In Table 16.5, while the effect of promotion and couponing are both large, the effect of promotion is much larger.

Issues in Interpretation - Multiple comparisons

- If the null hypothesis of equal means is rejected, we can only conclude that not all of the group means are equal. We may wish to examine differences among specific means. This can be done by specifying appropriate **contrasts**, or comparisons used to determine which of the means are statistically different.
- **A priori contrasts** are determined before conducting the analysis, based on the researcher's theoretical framework. Generally, a priori contrasts are used in lieu of the ANOVA *F* test. The contrasts selected are orthogonal (they are independent in a statistical sense).

Issues in Interpretation - Multiple Comparisons

- **A posteriori contrasts** are made after the analysis. These are generally **multiple comparison tests**. They enable the researcher to construct generalized confidence intervals that can be used to make pairwise comparisons of all treatment means. These tests, listed in order of decreasing power, include least significant difference, Duncan's multiple range test, Student-Newman-Keuls, Tukey's alternate procedure, honestly significant difference, modified least significant difference, and Scheffe's test. Of these tests, least significant difference is the most powerful, Scheffe's the most conservative.

Repeated Measures ANOVA

One way of controlling the differences between subjects is by observing each subject under each experimental condition (see Table 16.7). Since repeated measurements are obtained from each respondent, this design is referred to as within-subjects design or **repeated measures analysis of variance**. Repeated measures analysis of variance may be thought of as an extension of the paired-samples t test to the case of more than two related samples.

Decomposition of the Total Variation: Repeated Measures ANOVA

Table 16.7

		Independent Variable					X	
		Subject	Categories					Total
		No.						Sample
Between People Variation =SS _{between people}			X ₁	X ₂	X ₃	...	X _c	
	1	Y ₁₁	Y ₁₂	Y ₁₃			Y _{1c}	Y ₁
	2	Y ₂₁	Y ₂₂	Y ₂₃			Y _{2c}	Y ₂
		⋮						⋮
	n	Y _{n1}	Y _{n2}	Y _{n3}			Y _{nc}	Y _N
Category Mean			Y ₁	Y ₂	Y ₃		Y _c	Y
		Within People Category Variation = SS_{within people}						
		Total Variation = SS_y						

Repeated Measures ANOVA

In the case of a single factor with repeated measures, the total variation, with $nc - 1$ degrees of freedom, may be split into between-people variation and within-people variation.

$$SS_{\text{total}} = SS_{\text{between people}} + SS_{\text{within people}}$$

The between-people variation, which is related to the differences between the means of people, has $n - 1$ degrees of freedom. The within-people variation has $n(c - 1)$ degrees of freedom. The within-people variation may, in turn, be divided into two different sources of variation. One source is related to the differences between treatment means, and the second consists of residual or error variation. The degrees of freedom corresponding to the treatment variation are $c - 1$, and those corresponding to residual variation are $(c - 1)(n - 1)$.

Repeated Measures ANOVA

Thus,

$$SS_{\text{within people}} = SS_x + SS_{\text{error}}$$

A test of the null hypothesis of equal means may now be constructed in the usual way:

$$F = \frac{SS_x / (c - 1)}{SS_{\text{error}} / ((n - 1)(c - 1))} = \frac{MS_x}{MS_{\text{error}}}$$

So far we have assumed that the dependent variable is measured on an interval or ratio scale. If the dependent variable is nonmetric, however, a different procedure should be used.

Nonmetric Analysis of Variance

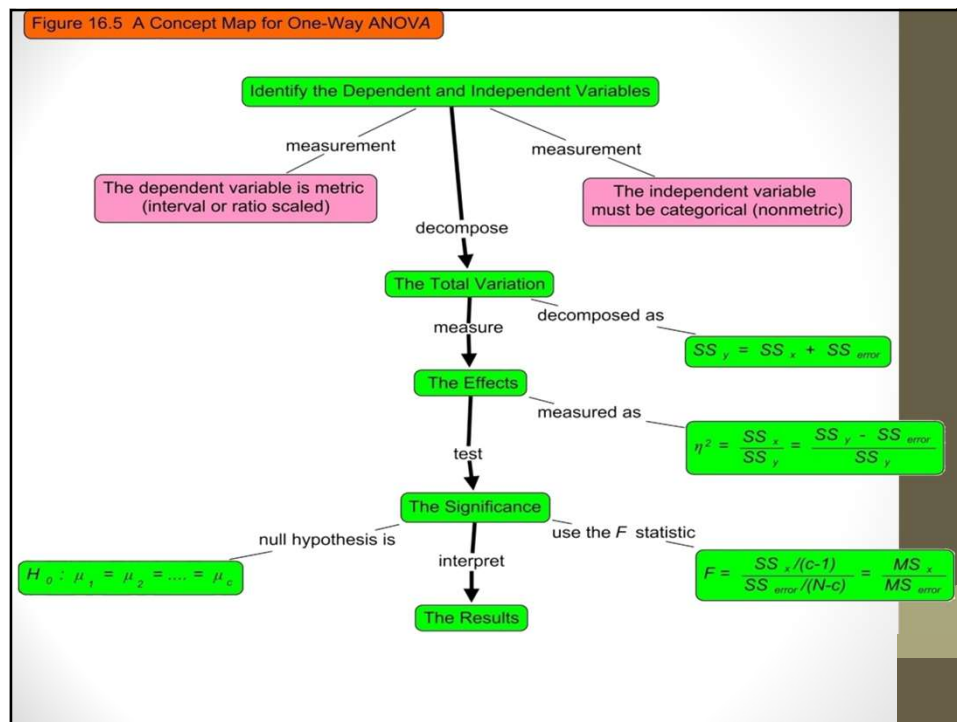
- **Nonmetric analysis of variance** examines the difference in the central tendencies of more than two groups when the dependent variable is measured on an ordinal scale.
- One such procedure is the **k-sample median test**. As its name implies, this is an extension of the median test for two groups, which was considered in Chapter 15.

Nonmetric Analysis of Variance

- A more powerful test is the **Kruskal-Wallis one-way analysis of variance**. This is an extension of the Mann-Whitney test (Chapter 15). This test also examines the difference in medians. All cases from the k groups are ordered in a single ranking. If the k populations are the same, the groups should be similar in terms of ranks within each group. The rank sum is calculated for each group. From these, the Kruskal-Wallis H statistic, which has a chi-square distribution, is computed.
- The Kruskal-Wallis test is more powerful than the k -sample median test as it uses the rank value of each case, not merely its location relative to the median. However, if there are a large number of tied rankings in the data, the k -sample median test may be a better choice.

Multivariate Analysis of Variance

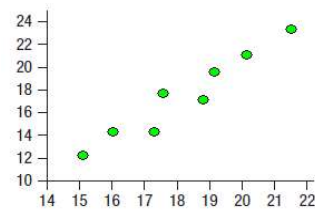
- **Multivariate analysis of variance (MANOVA)** is similar to analysis of variance (ANOVA), except that instead of one metric dependent variable, we have two or more.
- In MANOVA, the null hypothesis is that the vectors of means on multiple dependent variables are equal across groups.
- Multivariate analysis of variance is appropriate when there are two or more dependent variables that are correlated.



Statistical Tests: Correlational Analysis

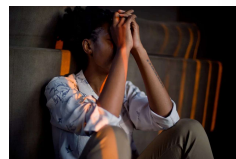
What are correlations?

- Used to determine how strongly variables are related
- Knowledge of one variable may help to predict the other
- For example: Height
 - Weight
 - Height of mother and father
 - Age



Using correlations in research

- Correlations form the first step
- When two variables are correlated, it is interesting to examine the origin of this correlation
- **NOT CAUSATION**



What can you do with correlations?

- Search for correlations
- Determine whether correlations are **significant**
- Determine how **strong** the correlations are
- **Pearson product-moment correlation**
most common with ratio/interval scales



Product Moment Correlation

- The **product moment correlation**, r , summarizes the strength of association between two metric (interval or ratio scaled) variables, say X and Y .
- It is an index used to determine whether a linear or straight-line relationship exists between X and Y .
- As it was originally proposed by Karl Pearson, it is also known as the *Pearson correlation coefficient*. It is also referred to as *simple correlation*, *bivariate correlation*, or merely the *correlation coefficient*.



Product Moment Correlation

From a sample of n observations, X and Y , the product moment correlation, r , can be calculated as:

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

Division of the numerator and denominator by $(n-1)$ gives

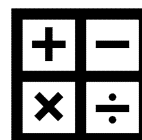
$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}}$$

$$= \frac{COV_{xy}}{S_x S_y}$$

In essence, it is the standardized **covariance** of x and y

Pearson correlation coefficient

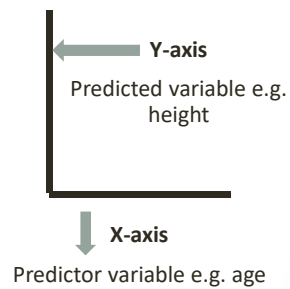
- + 1.00 (positive co-variation, each variable predicts the other)
- - 1.00 (negative co-variation, full prediction)
- +/- indicates positive or negative co-variation
- +1.00 = full, positive correlation
- 0.00 = no correlation
- -1.00 = full, negative correlation



Visualising a correlation: Scatterplots

- Scatterplot: One variable plotted against the other

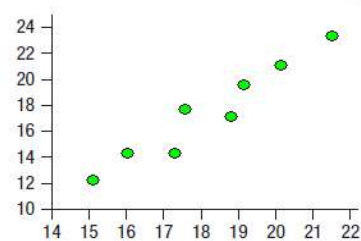
- Predictor variable – X-axis
- Predicted variable – Y axis



Visualising a correlation: Scatterplots

High, positive correlation

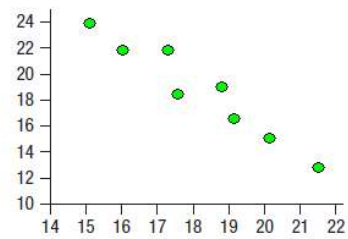
All dots are close the diagonal going from the lower left corner to the upper right corner



Visualising a correlation: Scatterplots

High, negative correlation

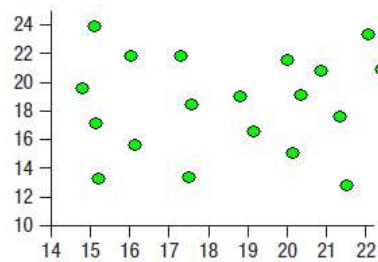
All dots are close the diagonal going from the upper left corner to the lower right corner



Visualising a correlation: Scatterplots

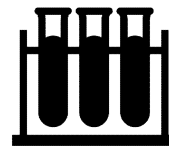
Null / no correlation

The dots are scattered throughout the plot



Hypothesis testing: Pearson correlation coefficient

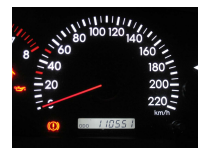
- r = correlation coefficient
- Many correlations in real life are between $-.30$ and $+.30$ so we want to know if they are significant



Hypothesis testing: Pearson correlation coefficient

Significance depends on:

1. The size of the correlation (the further away from 0, the higher chances it's significant)
2. The number of data pairs (the higher n = higher chances of significance)



Pearson correlation coefficient

Correlations			
		Stats	Social
Stats	Pearson Correlation	1	.451
	Sig. (2-tailed)		.191
	N	10	10
Social	Pearson Correlation	.451	1
	Sig. (2-tailed)	.191	
	N	10	10

Cohen (1988)

- .10 = Small
- .30 = Medium/moderate
- .50 = Large/strong

Statistical significance

Correlations			
		Stats	Social
Stats	Pearson Correlation	1	.451
	Sig. (2-tailed)		.191
	N	10	10
Social	Pearson Correlation	.451	1
	Sig. (2-tailed)	.191	
	N	10	10

Reporting results: APA style

The correlation between the exam marks in statistics and social psychology failed to reach significance ($r = .451$, $N = 10$, $p = .191$)

The value of r

The number of
data pairs on
which r is based

Significance

Examples

- Work motivation & work performance
- Organizational commitment & organizational citizenship behaviours
- Gender (2 groups) & neuroticism
- *When one of the two variables is dichotomous, r and t tests are identical.*

Multi-Group Correlational Analyses

- Should not simply put all groups together and do one correlational analysis

Explaining Attitude Toward the City of Residence



Table 17.1

Respondent No	Attitude Toward the City	Duration of Residence	Importance Attached to Weather
1	6	10	3
2	9	12	11
3	8	12	4
4	3	4	1
5	10	12	11
6	4	6	1
7	5	8	7
8	2	2	4
9	11	18	8
10	9	9	10
11	10	17	8
12	2	2	5

Product Moment Correlation

The **correlation coefficient** may be calculated as follows:

$$\bar{X} = (10 + 12 + 12 + 4 + 12 + 6 + 8 + 2 + 18 + 9 + 17 + 2)/12 \\ = 9.333$$

$$\bar{Y} = (6 + 9 + 8 + 3 + 10 + 4 + 5 + 2 + 11 + 9 + 10 + 2)/12 \\ = 6.583$$

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) &= (10 - 9.33)(6 - 6.58) + (12 - 9.33)(9 - 6.58) \\ &+ (12 - 9.33)(8 - 6.58) + (4 - 9.33)(3 - 6.58) \\ &+ (12 - 9.33)(10 - 6.58) + (6 - 9.33)(4 - 6.58) \\ &+ (8 - 9.33)(5 - 6.58) + (2 - 9.33)(2 - 6.58) \\ &+ (18 - 9.33)(11 - 6.58) + (9 - 9.33)(9 - 6.58) \\ &+ (17 - 9.33)(10 - 6.58) + (2 - 9.33)(2 - 6.58) \\ &= -0.3886 + 6.4614 + 3.7914 + 19.0814 \\ &+ 9.1314 + 8.5914 + 2.1014 + 33.5714 \\ &+ 38.3214 - 0.7986 + 26.2314 + 33.5714 \\ &= 179.6668 \end{aligned}$$

Product Moment Correlation

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= (10 - 9.33)^2 + (12 - 9.33)^2 + (12 - 9.33)^2 + (4 - 9.33)^2 \\ &+ (12 - 9.33)^2 + (6 - 9.33)^2 + (8 - 9.33)^2 + (2 - 9.33)^2 \\ &+ (18 - 9.33)^2 + (9 - 9.33)^2 + (17 - 9.33)^2 + (2 - 9.33)^2 \\ &= 0.4489 + 7.1289 + 7.1289 + 28.4089 \\ &+ 7.1289 + 11.0889 + 1.7689 + 53.7289 \\ &+ 75.1689 + 0.1089 + 58.8289 + 53.7289 \\ &= 304.6668 \end{aligned}$$

$$\begin{aligned} \sum_{i=1}^n (Y_i - \bar{Y})^2 &= (6 - 6.58)^2 + (9 - 6.58)^2 + (8 - 6.58)^2 + (3 - 6.58)^2 \\ &+ (10 - 6.58)^2 + (4 - 6.58)^2 + (5 - 6.58)^2 + (2 - 6.58)^2 \\ &+ (11 - 6.58)^2 + (9 - 6.58)^2 + (10 - 6.58)^2 + (2 - 6.58)^2 \\ &= 0.3364 + 5.8564 + 2.0164 + 12.8164 \\ &+ 11.6964 + 6.6564 + 2.4964 + 20.9764 \\ &+ 19.5364 + 5.8564 + 11.6964 + 20.9764 \\ &= 120.9168 \end{aligned}$$

$$\text{Thus, } r = \frac{179.6668}{\sqrt{(304.6668)(120.9168)}} = 0.9361$$

Decomposition of the Total Variation

$$\begin{aligned}
 r^2 &= \frac{\text{Explained variation}}{\text{Total variation}} \\
 &= \frac{SS_x}{SS_y} \\
 &= \frac{\text{Total variation} - \text{Error variation}}{\text{Total variation}} \\
 &= \frac{SS_y - SS_{error}}{SS_y}
 \end{aligned}$$

Decomposition of the Total Variation

- When it is computed for a population rather than a sample, the product moment correlation is denoted by ρ , the Greek letter rho. The coefficient r is an estimator of ρ .
- The statistical significance of the relationship between two variables measured by using r can be conveniently tested. The hypotheses are:

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Decomposition of the Total Variation

The test statistic is:

$$t = r \left[\frac{n-2}{1-r^2} \right]^{1/2}$$

which has a t distribution with $n - 2$ degrees of freedom.

For the correlation coefficient calculated based on the data given in Table 17.1,

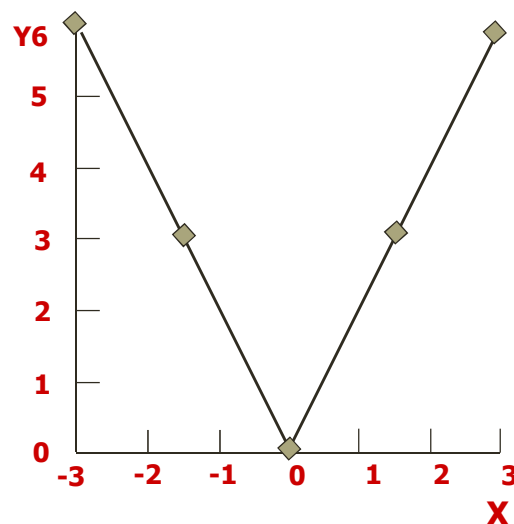
$$t = 0.9361 \left[\frac{12-2}{1 - (0.9361)^2} \right]^{1/2}$$

$$= 8.414$$

and the degrees of freedom = $12 - 2 = 10$. From the t distribution table (Table 4 in the Statistical Appendix), the critical value of t for a two-tailed test and $\alpha = 0.05$ is 2.228. Hence, the null hypothesis of no relationship between X and Y is rejected.

A Nonlinear Relationship for Which $r = 0$

Fig. 17.1



Partial Correlation

A **partial correlation coefficient** measures the association between two variables after controlling for, or adjusting for, the effects of one or more additional variables.

$$r_{xy \cdot z} = \frac{r_{xy} - (r_{xz})(r_{yz})}{\sqrt{1 - r_{xz}^2} \sqrt{1 - r_{yz}^2}}$$

- Partial correlations have an *order* associated with them. The order indicates how many variables are being adjusted or controlled.
- The simple correlation coefficient, r , has a zero-order, as it does not control for any additional variables while measuring the association between two variables.

Partial Correlation

- The coefficient $r_{xy \cdot z}$ is a **first-order partial correlation coefficient**, as it controls for the effect of one additional variable, Z .
- A **second-order partial correlation coefficient** controls for the effects of two variables, a third-order for the effects of three variables, and so on.
- The special case when a partial correlation is larger than its respective zero-order correlation involves a **suppressor effect**.

Part Correlation Coefficient

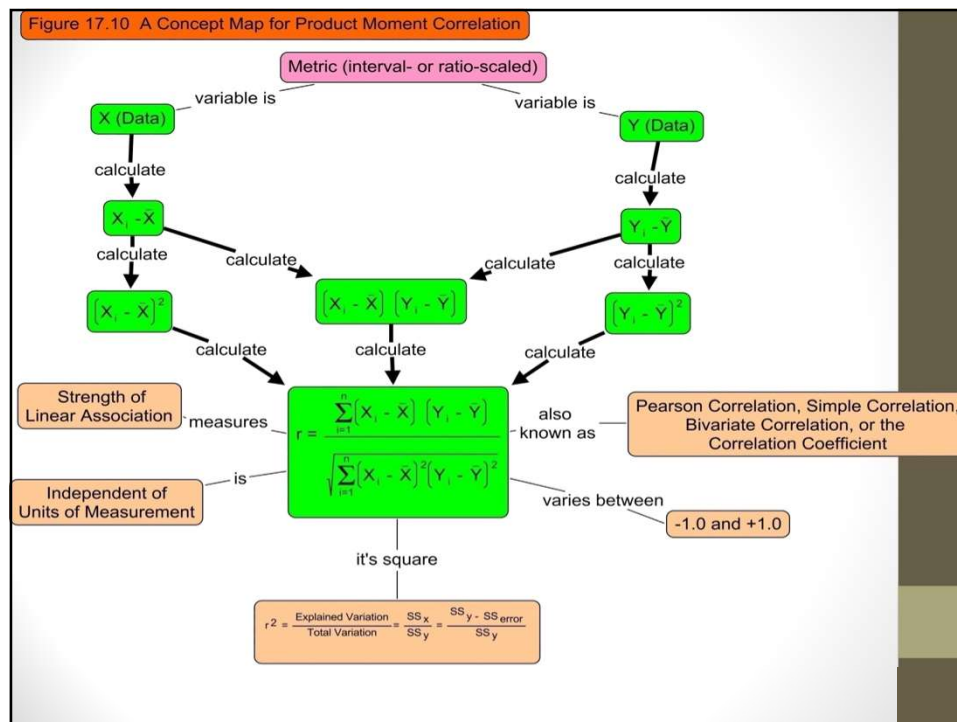
The **part correlation coefficient** represents the correlation between Y and X when the linear effects of the other independent variables have been removed from X but not from Y. The part correlation coefficient, $r_{y(x.z)}$ is calculated as follows:

$$r_{y(x.z)} = \frac{r_{xy} - r_{yz}r_{xz}}{\sqrt{1 - r_{xz}^2}}$$

The partial correlation coefficient is generally viewed as more important than the part correlation coefficient.

Nonmetric Correlation

- If the nonmetric variables are ordinal and numeric, Spearman's rho, ρ_s , and Kendall's tau, τ , are two measures of **nonmetric correlation**, which can be used to examine the correlation between them.
- Both these measures use rankings rather than the absolute values of the variables, and the basic concepts underlying them are quite similar. Both vary from -1.0 to +1.0 (see Chapter 15).
- In the absence of ties, **Spearman's** ρ_s yields a closer approximation to the Pearson product moment correlation coefficient, ρ , than Kendall's τ . In these cases, the absolute magnitude of τ tends to be smaller than Pearson's ρ .
- On the other hand, when the data contain a large number of tied ranks, **Kendall's** τ seems more appropriate.



Statistical Tests: Regression Analysis

Regression Analysis

Regression analysis examines associative relationships between a metric dependent variable and one or more independent variables in the following ways:

- Determine whether the independent variables explain a significant variation in the dependent variable: whether a relationship exists.
- Determine how much of the variation in the dependent variable can be explained by the independent variables: strength of the relationship.
- Determine the structure or form of the relationship: the mathematical equation relating the independent and dependent variables.
- Predict the values of the dependent variable.
- Control for other independent variables when evaluating the contributions of a specific variable or set of variables.
- Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality.

Statistics Associated with Bivariate Regression Analysis

- **Bivariate regression model.** The basic regression equation is $Y_i = \beta_0 + \beta_1 X_i + e_i$, where Y = dependent or criterion variable, X = independent or predictor variable, β_0 = intercept of the line, β_1 = slope of the line, and e_i is the error term associated with the i th observation.
- **Coefficient of determination.** The strength of association is measured by the coefficient of determination, r^2 . It varies between 0 and 1 and signifies the proportion of the total variation in Y that is accounted for by the variation in X .
- **Estimated or predicted value.** The estimated or predicted value of Y_i is $\hat{Y}_i = a + b x$, where $\hat{Y}_i$ is the predicted value of Y_i , and a and b are estimators of β_0 and β_1 , respectively.

Statistics Associated with Bivariate Regression Analysis

- **Regression coefficient.** The estimated parameter b is usually referred to as the non-standardized regression coefficient.
- **Scattergram.** A scatter diagram, or scattergram, is a plot of the values of two variables for all the cases or observations.
- **Standard error of estimate.** This statistic, SEE, is the standard deviation of the actual Y values from the predicted $\hat{Y}$ values.
- **Standard error.** The standard deviation of b , SE_b , is called the standard error.

Statistics Associated with Bivariate Regression Analysis

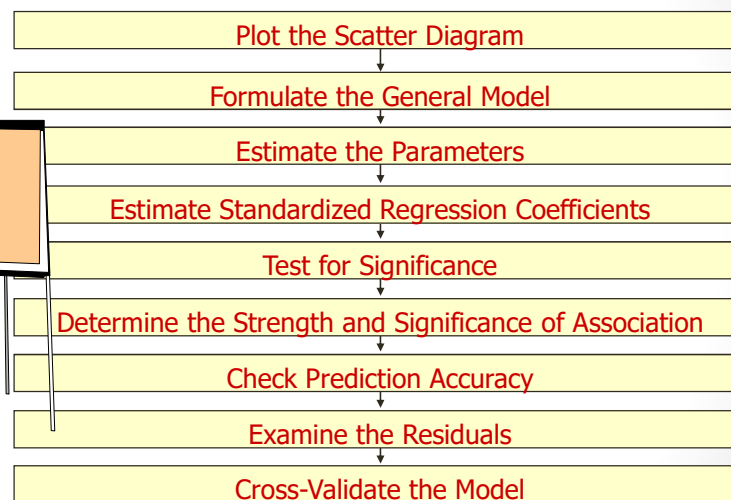
- **Standardized regression coefficient.** Also termed the beta coefficient or beta weight, this is the slope obtained by the regression of Y on X when the data are standardized.
- **Sum of squared errors.** The distances of all the points from the regression line are squared and added together to arrive at the sum of squared errors, which is a measure of total error, $\sum e_j^2$
- **t statistic.** A t statistic with $n - 2$ degrees of freedom can be used to test the null hypothesis that no linear relationship exists between X and Y , or $H_0: \beta = 0$, where $t = b / SE_b$

Conducting Bivariate Regression Analysis

Plot the Scatter Diagram

- A **scatter diagram**, or **scattergram**, is a plot of the values of two variables for all the cases or observations.
- The most commonly used technique for fitting a straight line to a scattergram is the **least-squares procedure**. In fitting the line, the least-squares procedure minimizes the sum of squared errors, $\sum e_j^2$.

Conducting Bivariate Regression Analysis



Conducting Bivariate Regression Analysis

Formulate the Bivariate Regression Model

In the bivariate regression model, the general form of a straight line is: $Y = \beta_0 + \beta_1 X$

where

Y = dependent or criterion variable

X = independent or predictor variable

β_0 = intercept of the line

β_1 = slope of the line

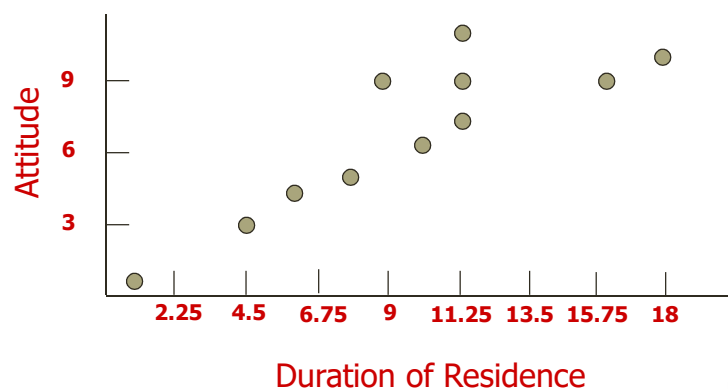
The regression procedure adds an error term to account for the probabilistic or stochastic nature of the relationship:

$$Y_i = \beta_0 + \beta_1 X_i + e_i$$

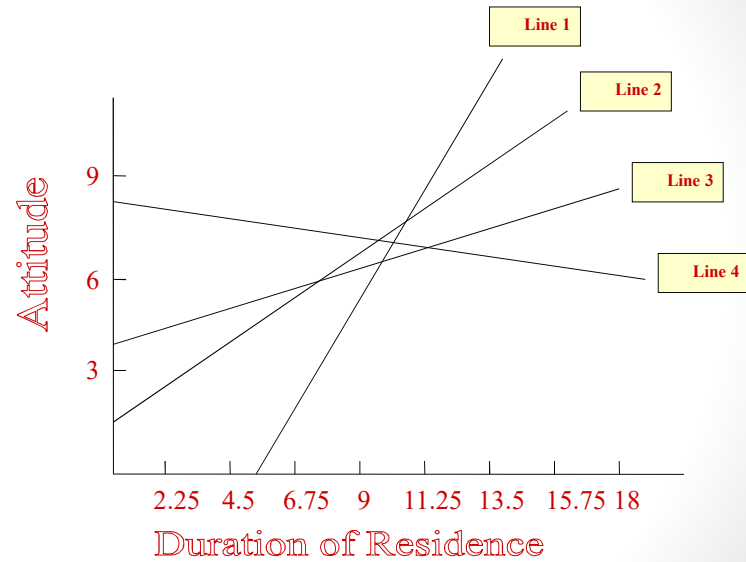
where e_i is the error term associated with the i th observation.

Plot of Attitude with Duration

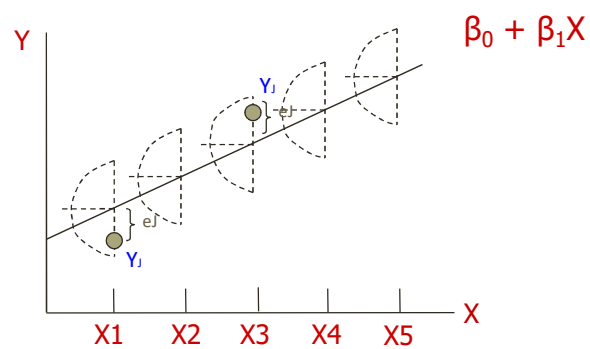
Fig. 17.3



Which Straight Line Is Best?



Bivariate Regression



Conducting Bivariate Regression Analysis

Estimate the Parameters

In most cases, β_0 and β_1 are unknown and are estimated from the sample observations using the equation

$$\hat{Y}_i = a + b x_i$$

where $\hat{Y}_i$ is the estimated or predicted value of Y_i , and a and b are estimators of β_0 and β_1 , respectively.

$$b = \frac{COV_{xy}}{S_x^2}$$

$$= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2}$$

Conducting Bivariate Regression Analysis

Estimate the Parameters

The intercept, a , may then be calculated using:

$$a = \bar{Y} - b \bar{X}$$

For the data in Table 17.1, the estimation of parameters may be illustrated as follows:

$$\sum_{i=1}^{12} X_i Y_i$$

$$= (10)(6) + (12)(9) + (12)(8) + (4)(3) + (12)(10) + (6)(4) \\ + (8)(5) + (2)(2) + (18)(11) + (9)(9) + (17)(10) + (2)(2) \\ = 917$$

$$\sum_{i=1}^{12} X_i^2 = 102 + 122 + 122 + 42 + 122 + 62 \\ + 82 + 22 + 182 + 92 + 172 + 22 \\ = 1350$$

Conducting Bivariate Regression Analysis

Estimate the Parameters

It may be recalled from earlier calculations of the simple correlation that:

$$\bar{X} = 9.333$$

$$\bar{Y} = 6.583$$

Given $n = 12$, b can be calculated as:

$$b = \frac{917 - (12)(9.333)(6.583)}{1350 - (12)(9.333)^2}$$

$$= 0.5897$$

$$a = \bar{Y} - b\bar{X}$$

$$= 6.583 - (0.5897)(9.333)$$

$$= 1.0793$$

Conducting Bivariate Regression Analysis

Estimate the Standardized Regression Coefficient

- **Standardization** is the process by which the raw data are transformed into new variables that have a mean of 0 and a variance of 1 (Chapter 14).
- When the data are standardized, the intercept assumes a value of 0.
- The term **beta coefficient** or **beta weight** is used to denote the standardized regression coefficient.

$$B_{yx} = B_{xy} = r_{xy}$$

- There is a simple relationship between the standardized and non-standardized regression coefficients:

$$B_{yx} = b_{yx} (S_x / S_y)$$

Conducting Bivariate Regression Analysis Test for Significance

The statistical significance of the linear relationship between X and Y may be tested by examining the hypotheses:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

A t statistic with $n - 2$ degrees of freedom can be used, where

$$t = \frac{b}{SE_b}$$

SE_b denotes the standard deviation of b and is called the **standard error**.

Conducting Bivariate Regression Analysis Test for Significance

Using a computer program, the regression of attitude on duration of residence, using the data shown in Table 17.1, yielded the results shown in Table 17.2. The intercept, a , equals 1.0793, and the slope, b , equals 0.5897. Therefore, the estimated equation is:

$$\text{Attitude } (\hat{Y}) = 1.0793 + 0.5897 (\text{Duration of residence})$$

The standard error, or standard deviation of b is estimated as 0.07008, and the value of the t statistic as $t = 0.5897/0.0700 = 8.414$, with $n - 2 = 10$ degrees of freedom.

From Table 4 in the Statistical Appendix, we see that the critical value of t with 10 degrees of freedom and $\alpha = 0.05$ is 2.228 for a two-tailed test. Since the calculated value of t is larger than the critical value, the null hypothesis is rejected.

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

The total variation, SS_y , may be decomposed into the variation accounted for by the regression line, SS_{reg} , and the error or residual variation, SS_{error} or SS_{res} , as follows:

$$SS_y = SS_{reg} + SS_{res}$$

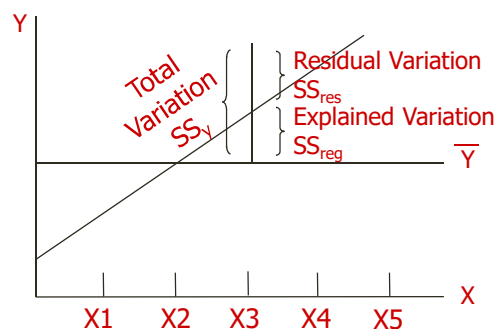
where

$$SS_y = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SS_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Decomposition of the Total Variation in Bivariate Regression



Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

The **strength of association** may then be calculated as follows:

$$r^2 = \frac{SS_{reg}}{SS_y}$$

$$= \frac{SS_y - SS_{res}}{SS_y}$$

To illustrate the calculations of r^2 , let us consider again the effect of attitude toward the city on the duration of residence. It may be recalled from earlier calculations of the simple correlation coefficient that:

$$SS_y = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$= 120.9168$$

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

The predicted values ($\hat{Y}$) can be calculated using the regression equation:

$$\text{Attitude } (\hat{Y}) = 1.0793 + 0.5897 (\text{Duration of residence})$$

For the first observation in Table 17.1, this value is:

$$(\hat{Y}) = 1.0793 + 0.5897 \times 10 = 6.9763.$$

For each successive observation, the predicted values are, in order, 8.1557, 8.1557, 3.4381, 8.1557, 4.6175, 5.7969, 2.2587, 11.6939, 6.3866, 11.1042, and 2.2587.

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

Therefore,

$$\begin{aligned}
 SS_{reg} &= \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 \\
 &= (6.9763 - 6.5833)^2 + (8.1557 - 6.5833)^2 \\
 &\quad + (8.1557 - 6.5833)^2 + (3.4381 - 6.5833)^2 \\
 &\quad + (8.1557 - 6.5833)^2 + (4.6175 - 6.5833)^2 \\
 &\quad + (5.7969 - 6.5833)^2 + (2.2587 - 6.5833)^2 \\
 &\quad + (11.6939 - 6.5833)^2 + (6.3866 - 6.5833)^2 \\
 &\quad + (11.1042 - 6.5833)^2 + (2.2587 - 6.5833)^2 \\
 &= 0.1544 + 2.4724 + 2.4724 + 9.8922 + 2.4724 \\
 &\quad + 3.8643 + 0.6184 + 18.7021 + 26.1182 \\
 &\quad + 0.0387 + 20.4385 + 18.7021 \\
 &= 105.9524
 \end{aligned}$$

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

$$\begin{aligned}
 SS_{res} &= \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = (6 - 6.9763)^2 + (9 - 8.1557)^2 + (8 - 8.1557)^2 \\
 &\quad + (3 - 3.4381)^2 + (10 - 8.1557)^2 + (4 - 4.6175)^2 \\
 &\quad + (5 - 5.7969)^2 + (2 - 2.2587)^2 + (11 - 11.6939)^2 \\
 &\quad + (9 - 6.3866)^2 + (10 - 11.1042)^2 + (2 - 2.2587)^2 \\
 &= 14.9644
 \end{aligned}$$

It can be seen that $SS_y = SS_{reg} + SS_{res}$. Furthermore,

$$\begin{aligned}
 r^2 &= SS_{reg} / SS_y \\
 &= 105.9524 / 120.9168 \\
 &= 0.8762
 \end{aligned}$$

Conducting Bivariate Regression Analysis Determine the Strength and Significance of Association

Another, equivalent test for examining the significance of the linear relationship between X and Y (significance of b) is the test for the significance of the coefficient of determination. The hypotheses in this case are:

$$H_0: R^2_{pop} = 0$$

$$H_1: R^2_{pop} > 0$$

Conducting Bivariate Regression Analysis Determine the Strength and Significance of Association

The appropriate test statistic is the F statistic:

$$F = \frac{SS_{reg}}{SS_{res}/(n-2)}$$

which has an F distribution with 1 and $n - 2$ degrees of freedom. The F test is a generalized form of the t test (see Chapter 15). If a random variable is t distributed with n degrees of freedom, then t^2 is F distributed with 1 and n degrees of freedom. Hence, the F test for testing the significance of the coefficient of determination is equivalent to testing the following hypotheses:

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

or

$$H_0: \rho = 0$$

$$H_1: \rho \neq 0$$

Conducting Bivariate Regression Analysis

Determine the Strength and Significance of Association

From Table 17.2, it can be seen that:

$$r^2 = 105.9522 / (105.9522 + 14.9644)$$

$$= 0.8762$$

Which is the same as the value calculated earlier. The value of the F statistic is:

$$F = 105.9522 / (14.9644 / 10)$$

$$= 70.8027$$

with 1 and 10 degrees of freedom. The calculated F statistic exceeds the critical value of 4.96 determined from Table 5 in the Statistical Appendix. Therefore, the relationship is significant at $\alpha = 0.05$, corroborating the results of the t test.

α

Bivariate Regression



Multiple R	0.93608
R^2	0.87624
Adjusted R^2	0.86387
Standard Error	1.22329

ANALYSIS OF VARIANCE

	df	Sum of Squares	Mean Square
Regression	1	105.95222	105.95222
Residual	10	14.96444	1.49644
$F = 70.80266$		Significance of $F = 0.0000$	

VARIABLES IN THE EQUATION

Variable	b	SE_b	Beta (β)	T of T	Significance
Duration	0.58972	0.07008	0.93608	8.414	0.0000
(Constant)	1.07932	0.74335		1.452	0.1772

Conducting Bivariate Regression Analysis

Check Prediction Accuracy

To estimate the accuracy of predicted values, $\hat{Y}$, it is useful to calculate the standard error of estimate, SEE.

$$SEE = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}}$$

or

$$SEE = \sqrt{\frac{SS_{res}}{n-2}}$$

or more generally, if there are k independent variables,

$$SEE = \sqrt{\frac{SS_{res}}{n-k-1}}$$

For the data given in Table 17.2, the SEE is estimated as follows:

$$\begin{aligned} SEE &= \sqrt{14.9644/(12-2)} \\ &= 1.22329 \end{aligned}$$

Assumptions

- The error term is **normally** distributed. For each fixed value of X , the distribution of Y is normal.
- The means of all these normal distributions of Y , given X , lie on a **straight line** with slope b .
- The mean of the error term is **0**.
- The variance of the error term is **constant**. This variance does not depend on the values assumed by X .
- The error terms are **uncorrelated**. In other words, the observations have been drawn independently.

Multiple Regression

The general form of the **multiple regression model** is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + e$$

which is estimated by the following equation:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \dots + b_k X_k$$

As before, the coefficient a represents the intercept, but the b 's are now the partial regression coefficients.

Statistics Associated with Multiple Regression

- **Adjusted R^2 .** R^2 , coefficient of multiple determination, is adjusted for the number of independent variables and the sample size to account for the diminishing returns. After the first few variables, the additional independent variables do not make much contribution.
- **Coefficient of multiple determination.** The strength of association in multiple regression is measured by the square of the multiple correlation coefficient, R^2 , which is also called the coefficient of multiple determination.
- **F test.** The F test is used to test the null hypothesis that the coefficient of multiple determination in the population, R^2_{pop} , is zero. This is equivalent to testing the null hypothesis. The test statistic has an F distribution with k and $(n - k - 1)$ degrees of freedom.

Statistics Associated with Multiple Regression

- **Partial F test.** The significance of a partial regression coefficient, b_j , of X_j may be tested using an incremental F statistic. The incremental F statistic is based on the increment in the explained sum of squares resulting from the addition of the independent variable X_j to the regression equation after all the other independent variables have been included.
- **Partial regression coefficient.** The partial regression coefficient, b_1 , denotes the change in the predicted value, $\hat{Y}$, per unit change in X_1 when the other independent variables, X_2 to X_k , are held constant.

$\hat{Y}$

Conducting Multiple Regression Analysis Partial Regression Coefficients

To understand the meaning of a partial regression coefficient, let us consider a case in which there are two independent variables, so that:

$$\hat{Y} = a + b_1X_1 + b_2X_2$$

- First, note that the relative magnitude of the partial regression coefficient of an independent variable is, in general, different from that of its bivariate regression coefficient.
- The interpretation of the partial regression coefficient, b_1 , is that it represents the expected change in Y when X_1 is changed by one unit but X_2 is held constant or otherwise controlled. Likewise, b_2 represents the expected change in Y for a unit change in X_2 , when X_1 is held constant. Thus, calling b_1 and b_2 partial regression coefficients is appropriate.

Conducting Multiple Regression Analysis Partial Regression Coefficients

- It can also be seen that the combined effects of X_1 and X_2 on Y are additive. In other words, if X_1 and X_2 are each changed by one unit, the expected change in Y would be (b_1+b_2) .
- Suppose one was to remove the effect of X_2 from X_1 . This could be done by running a regression of X_1 on X_2 . In other words, one would estimate the equation $\hat{X}_1 = a + b X_2$ and calculate the residual $X_r = (X_1 - \hat{X}_1)$. The **partial regression coefficient**, b_1 , is then equal to the bivariate regression coefficient, b_r , obtained from the equation $\hat{Y} = a + b_r X_r$.

Conducting Multiple Regression Analysis Partial Regression Coefficients

- Extension to the case of k variables is straightforward. The partial regression coefficient, b_1 , represents the expected change in Y when X_1 is changed by one unit and X_2 through X_k are held constant. It can also be interpreted as the bivariate regression coefficient, b , for the regression of Y on the residuals of X_1 , when the effect of X_2 through X_k has been removed from X_1 .

- The relationship of the standardized to the non-standardized coefficients remains the same as before:

$$B_1 = b_1 (S_{x1}/S_y)$$

$$B_k = b_k (S_{xk}/S_y)$$

The estimated regression equation is:

$$(\hat{Y}) = 0.33732 + 0.48108 X_1 + 0.28865 X_2$$

Or

$$\text{Attitude} = 0.33732 + 0.48108 (\text{Duration}) + 0.28865 (\text{Importance})$$

Multiple Regression

Multiple R 0.97210
 R^2 0.94498
 Adjusted R^2 0.93276
 Standard Error 0.85974



ANALYSIS OF VARIANCE

	df	Sum of Squares	Mean Square
Regression	2	114.26425	57.13213
Residual	9	6.65241	0.73916
$F = 77.29364$ Significance of $F = 0.0000$			

VARIABLES IN THE EQUATION

Variable	b	SE_b	Beta (β)	T of T	Significance
IMPORTANCE	0.28865	0.08608	0.31382	3.353	0.0085
DURATION	0.48108	0.05895	0.76363	8.160	0.0000
(Constant)	0.33732	0.56736		0.595	0.5668

Conducting Multiple Regression Analysis Strength of Association

$$SS_y = SS_{reg} + SS_{res}$$

where

$$SS_y = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

$$SS_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Conducting Multiple Regression Analysis Strength of Association

The strength of association is measured by the square of the multiple correlation coefficient, R^2 , which is also called the **coefficient of multiple determination**.

$$R^2 = \frac{SS_{reg}}{SS_y}$$

R^2 is adjusted for the number of independent variables and the sample size by using the following formula:

$$\text{Adjusted } R^2 = R^2 - \frac{k(1 - R^2)}{n - k - 1}$$

Conducting Multiple Regression Analysis Significance Testing

$$H_0 : R^2_{pop} = 0$$

This is equivalent to the following null hypothesis:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$$

The overall test can be conducted by using an F statistic:

$$\begin{aligned} F &= \frac{SS_{reg}/k}{SS_{res}/(n - k - 1)} \\ &= \frac{R^2/k}{(1 - R^2)/(n - k - 1)} \end{aligned}$$

which has an F distribution with k and $(n - k - 1)$ degrees of freedom.

Conducting Multiple Regression Analysis Significance Testing

Testing for the significance of the β_i 's can be done in a manner similar to that in the bivariate case by using t tests. The significance of the partial coefficient for importance attached to weather may be tested by the following equation:

$$t = \frac{b}{SE_b}$$

which has a t distribution with $n - k - 1$ degrees of freedom.

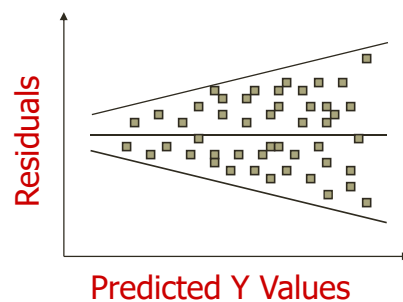
Conducting Multiple Regression Analysis Examination of Residuals

- A **residual** is the difference between the observed value of Y_i and the value predicted by the regression equation $\hat{Y}_i$.
- Scattergrams of the residuals, in which the residuals are plotted against the predicted values, $\hat{Y}_i$, time, or predictor variables, provide useful insights in examining the appropriateness of the underlying assumptions and regression model fit.
- The assumption of a normally distributed error term can be examined by constructing a histogram of the residuals.
- The assumption of constant variance of the error term can be examined by plotting the residuals against the predicted values of the dependent variable, $\hat{Y}_i$.

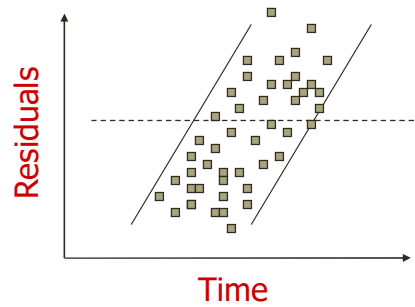
Conducting Multiple Regression Analysis Examination of Residuals

- A plot of residuals against time, or the sequence of observations, will throw some light on the assumption that the error terms are uncorrelated.
- Plotting the residuals against the independent variables provides evidence of the appropriateness or inappropriateness of using a linear model. Again, the plot should result in a random pattern.
- To examine whether any additional variables should be included in the regression equation, one could run a regression of the residuals on the proposed variables.
- If an examination of the residuals indicates that the assumptions underlying linear regression are not met, the researcher can transform the variables in an attempt to satisfy the assumptions.

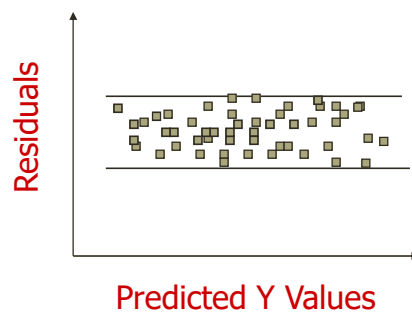
Residual Plot Indicating that Variance Is Not Constant



Residual Plot Indicating a Linear Relationship Between Residuals and Time



Plot of Residuals Indicating that a Fitted Model Is Appropriate



Stepwise Regression

The purpose of **stepwise regression** is to select, from a large number of predictor variables, a small subset of variables that account for most of the variation in the dependent or criterion variable. In this procedure, the predictor variables enter or are removed from the regression equation one at a time. There are several approaches to stepwise regression.

- **Forward inclusion.** Initially, there are no predictor variables in the regression equation. Predictor variables are entered one at a time, only if they meet certain criteria specified in terms of F ratio. The order in which the variables are included is based on the contribution to the explained variance.
- **Backward elimination.** Initially, all the predictor variables are included in the regression equation. Predictors are then removed one at a time based on the F ratio for removal.
- **Stepwise solution.** Forward inclusion is combined with the removal of predictors that no longer meet the specified criterion at each step.

Multicollinearity

- **Multicollinearity** arises when intercorrelations among the predictors are very high.
- Multicollinearity can result in several problems, including:
 - The partial regression coefficients may not be estimated precisely. The standard errors are likely to be high.
 - The magnitudes, as well as the signs of the partial regression coefficients, may change from sample to sample.
 - It becomes difficult to assess the relative importance of the independent variables in explaining the variation in the dependent variable.
 - Predictor variables may be incorrectly included or removed in stepwise regression.

Multicollinearity

- A simple procedure for adjusting for multicollinearity consists of using only one of the variables in a highly correlated set of variables.
- Alternatively, the set of independent variables can be transformed into a new set of predictors that are mutually independent by using techniques such as principal components analysis.
- More specialized techniques, such as ridge regression and latent root regression, can also be used.

Relative Importance of Predictors

Unfortunately, because the predictors are correlated, there is no unambiguous measure of relative importance of the predictors in regression analysis. However, several approaches are commonly used to assess the relative importance of predictor variables.

- **Statistical significance.** If the partial regression coefficient of a variable is not significant, as determined by an incremental F test, that variable is judged to be unimportant. An exception to this rule is made if there are strong theoretical reasons for believing that the variable is important.
- **Square of the simple correlation coefficient.** This measure, r^2 , represents the proportion of the variation in the dependent variable explained by the independent variable in a bivariate relationship.

Relative Importance of Predictors

- **Square of the partial correlation coefficient.** This measure, $R^2_{yxi.xjxk}$, is the coefficient of determination between the dependent variable and the independent variable, controlling for the effects of the other independent variables.
- **Square of the part correlation coefficient.** This coefficient represents an increase in R^2 when a variable is entered into a regression equation that already contains the other independent variables.
- **Measures based on standardized coefficients or beta weights.** The most commonly used measures are the absolute values of the beta weights, $|B_i|$, or the squared values, B_i^2 .
- **Stepwise regression.** The order in which the predictors enter or are removed from the regression equation is used to infer their relative importance.

Cross-Validation

- The regression model is estimated using the entire data set.
- The available data are split into two parts, the **estimation sample** and the **validation sample**. The estimation sample generally contains 50-90% of the total sample.
- The regression model is estimated using the data from the estimation sample only. This model is compared to the model estimated on the entire sample to determine the agreement in terms of the signs and magnitudes of the partial regression coefficients.
- The estimated model is applied to the data in the validation sample to predict the values of the dependent variable, $\hat{Y}_p$, for the observations in the validation sample.
- The observed values Y_p and the predicted values, $\hat{Y}_p$, in the validation sample are correlated to determine the simple r^2 . This measure, r^2 , is compared to R^2 for the total sample and to R^2 for the estimation sample to assess the degree of shrinkage.

Regression with Dummy Variables

Product Usage Category	Original Variable Code	Dummy Variable Code		
		D1	D2	D3
Nonusers.....	1	1	0	0
Light Users.....	2	0	1	0
Medium Users.....	3	0	0	1
Heavy Users.....	4	0	0	0

$$\hat{Y}_i = a + b_1D_1 + b_2D_2 + b_3D_3$$

- In this case, "heavy users" has been selected as a reference category and has not been directly included in the regression equation.
- The coefficient b_1 is the difference in predicted $\hat{Y}_i$ for nonusers, as compared to heavy users.

Analysis of Variance and Covariance with Regression

In regression with dummy variables, the predicted $\hat{Y}$ for each category is the mean of Y for each category.

Product Usage Category	Predicted Value $\hat{Y}$	Mean Value $\bar{Y}$
Nonusers.....	$a + b_1$	$a + b_1$
Light Users.....	$a + b_2$	$a + b_2$
Medium Users.....	$a + b_3$	$a + b_3$
Heavy Users.....	a	a

Analysis of Variance and Covariance with Regression

Given this equivalence, it is easy to see further relationships between dummy variable regression and one-way ANOVA.

Dummy Variable Regression

$$SS_{res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

$$SS_{reg} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

$$R^2$$

Overall F test

One-Way ANOVA

$$= SS_{within} = SS_{error}$$

$$= SS_{between} = SS_x$$

$$= \eta^2$$

= F test

