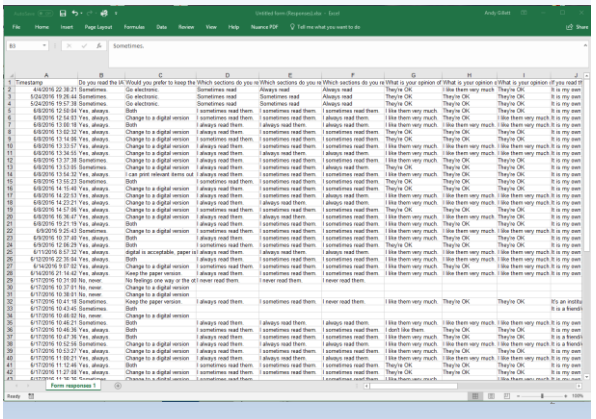
Research Methods

Primary Data Collection: Quantitative Data Analysis

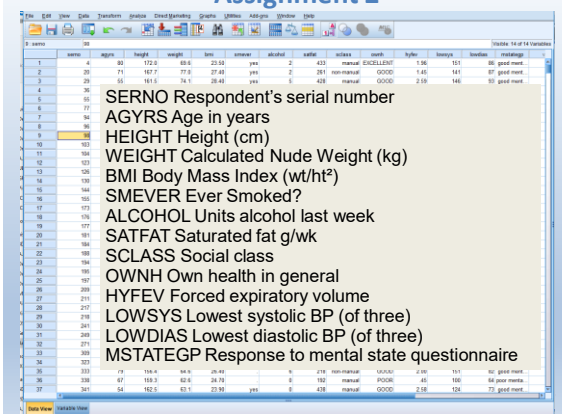
Andy Gillett

1



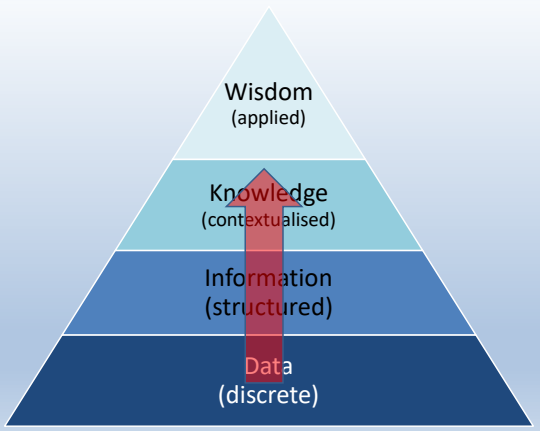
2

Assignment 2



SERNO Respondent's serial number
AGYRS Age in years
HEIGHT Height (cm)
WEIGHT Calculated Nude Weight (kg)
BMI Body Mass Index (wt/ht²)
SMEVER Ever Smoked?
ALCOHOL Units alcohol last week
SATFAT Saturated fat g/wk
SCLASS Social class
OWNH Own health in general
HYFEV Forced expiratory volume
LOWSYS Lowest systolic BP (of three)
LOWDIAS Lowest diastolic BP (of three)
MSTATEGP Response to mental state questionnaire

3



4

Types of Data

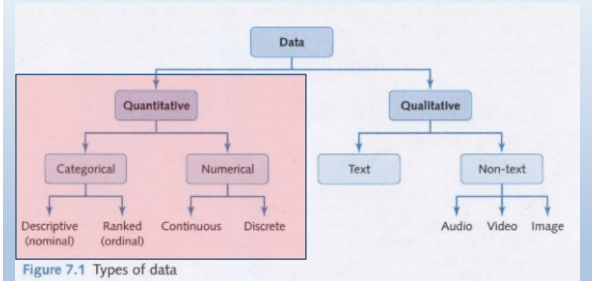


Figure 7.1 Types of data

(Saunders & Lewis, 2012: 165)

5

IBM SPSS Statistics

6

SPSS

7

7

8

8

Quantitative Data

Two main types:
Categorical and Numerical data.

Categorical data –
- grouped into a descriptive set (nominal);
- put in rank order (ordinal).

9

9

Categorical Data

Nominal

In which faculty are you studying?	Arts and Humanities Business and Law Physical Sciences Social Sciences	□1 □2 □3 □4

Ordinal

How likely do you believe it is that you will pay off your overdraft within a year of graduation?	Very likely Quite likely Unsure Quite unlikely Very unlikely	□1 □2 □3 □4 □5

(Saunders & Lewis, 2012: 166)

10

10

Quantitative Data

Numerical data - measured using numbers.

Either:
- discrete data – discrete units
- continuous data – any value

11

11

Numerical Data

Discrete

On how many days did you visit the library last week? *(For example, for 2 write: [2])*	[ ]

Continuous

How much time did you spend studying last week ?	[ ] [ ] hours [ ] [ ] mins
(For example, for 2 hours, 15 minutes write: [ ] [2] hours [1] [5] mins)	

(Saunders & Lewis, 2012: 166-167)

12

12

Definitions

- continuous data: numerical data whose values are measured numerically as quantities and can theoretically take any value, depending on the level of accuracy with which they are measured.
- discrete data: numerical data whose values are measured numerically as quantities in discrete units and can therefore only take a finite number of values.

scale

- ranked or ordinal data: categorical data that are put into a definite order.

ordinal

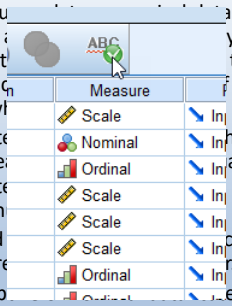
- descriptive or nominal data: categorical data that are grouped into sets (categories) that have no obvious rank order.

nominal

13

13

Definitions



- continu… …a whose values … …y as quantit… …take any value, … …accuracy with wh…
- discrete… …hose values are me… …antities in discrete… …only take a finite n…
- ranked … …cal data that are… …r.
- descrip… …egorical data that are grouped into sets (categories) that have no obvious rank order.

scale

ordinal

nominal

14

14

Definitions

- continuous data: numerical data whose values are measured numerically as quantities and can theoretically take any value, depending on the level of accuracy with which they are measured.
- discrete data: numerical data whose values are measured numerically as quantities in discrete units and can therefore only take a finite number of values.

parametric

- ranked or ordinal data: categorical data that are put into a definite order.
- descriptive or nominal data: categorical data that are grouped into sets (categories) that have no obvious rank order.

non-parametric

15

15

Types of Data

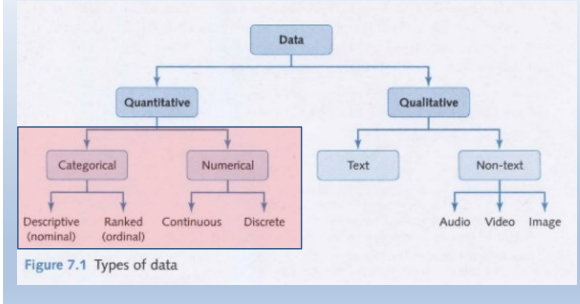


Figure 7.1 Types of data

(Saunders & Lewis, 2012: 165)

16

16

Preparing Data: Coding

CODING CLOSED QUESTIONS

Each response to a question needs an individual code that allows for data analysis. For example, Male or Female?

So the coding of the question could be:

1 = male

2 = female

17

17

Preparing Data

CODING CLOSED QUESTIONS

Each response to a question needs an individual code that allows for data analysis. For example, Male or Female?

So the full coding of the question could be:

1 = male
2 = female
3 = would prefer not to respond
4 = is unwilling to be characterised in this manner
5 = this information is confidential
9 = missing

18

18

Preparing Data

The analytical socio-economic classes are:

1 Higher managerial and professional occupations
2 Lower managerial and professional occupations
3 Intermediate occupations
4 Small employers and own-account workers
5 Lower supervisory and technical occupations
6 Semi-routine occupations
7 Routine occupations
8 Never worked, and long-term unemployed.

19

19

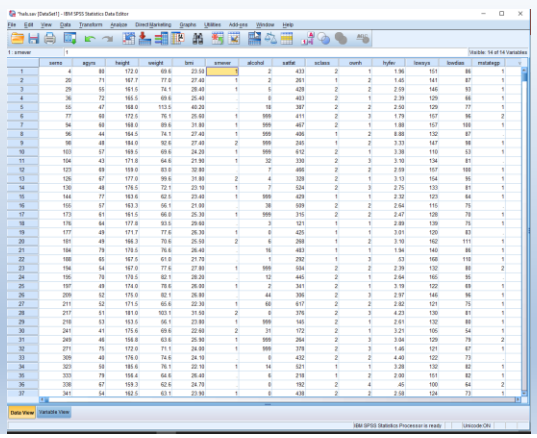
Preparing Data

For a scaled question such as 'How enjoyable did you find your university studies? Answer on the scale 1-6', answer codes might be:

- Very enjoyable 1
- Quite enjoyable 2
- Mostly enjoyable 3
- Not very enjoyable 4
- Not enjoyable 5
- Dreadful 6

20

20



21

height	weight	bmi	smever	alcohol	satfat	sclass	ownh
172.0	69.6	23.50	1	2	433	2	1
167.7	77.0	27.40	1	2	261	1	2
161.5	74.1	28.40	1	5	428	2	2
165.5	69.6	25.40	.	0	403	2	1
168.0	113.5	40.20	.	18	387	2	2
172.5	76.1	25.60	1	999	411	2	3
168.0	89.6	31.80	1	999	467	2	1
164.5	74.1	27.40	1	999	406	1	2
184.0	92.6	27.40	2	999	245	1	2
169.5	69.6	24.20	1	999	612	2	1
171.8	64.6	21.90	1	32	330	2	3
159.0	83.0	32.80	.	7	466	2	2
177.0	99.6	31.80	2	4	328	2	1
176.5	72.1	23.10	1	7	524	2	3
163.6	62.5	23.40	1	999	429	1	1
163.3	56.1	21.00	.	38	509	2	2
161.5	66.0	25.30	1	999	315	2	2
177.8	93.5	29.60	.	3	121	1	1
171.7	77.6	26.30	1	0	425	1	1
166.3	70.6	25.50	2	6	268	1	2
170.5	76.6	26.40	.	16	483	1	1

22

22

Analysing Data Quantitatively

How to prepare your data

Spreadsheet (Excel)

Statistical analysis software (SPSS, R, Minitab)

	A	B	C	D	E
1	Student identifier	Faculty in which studying	Likelihood of paying off overdraft within year of graduation	Days visted library last week	Hours spent studying last week
2	1	2	2	3	28
3	2	2	5	4	34
4	3	2	5	3	57
5	4	3	5	2	34
6	5	1	5	0	8

Figure 7.2 Data matrix in Excel

(Saunders & Lewis, 2012: 168)

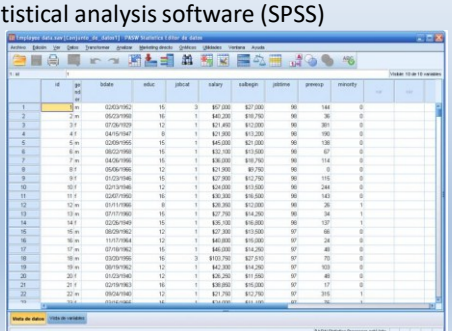
23

23

Analysing Data Quantitatively

How to prepare your data

Statistical analysis software (SPSS)



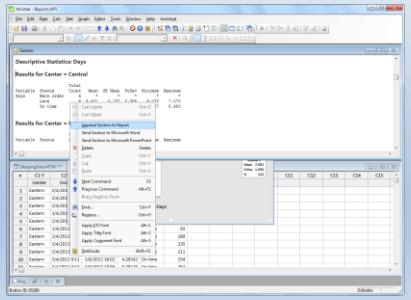
24

24

Analysing Data Quantitatively

How to prepare your data

Statistical analysis software (Minitab)



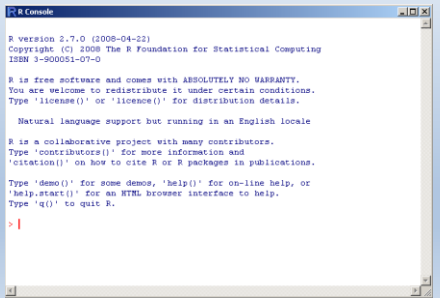
25

25

Analysing Data Quantitatively

How to prepare your data

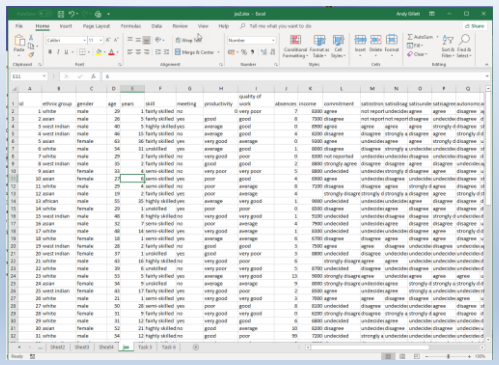
Statistical analysis software (R)



26

26

MS Excel

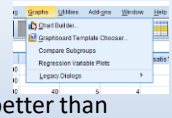


27

27

Presenting your Data

Explore and understand your data – tables, charts and graphs.



- Particular tables or graphs are better than others for highlighting certain aspects of your data
- Categorical and numerical data often require different tables and graphs.
- All tables and graphs must be labelled clearly, explicitly referred to and designed so as not to distort the data.

28

28

Presenting your Data

Table 7.1 Presenting data using tables and graphs

For...	use a...	to present...	which allows the...
categorical data	table	a summary	individual values to be read
categorical data	bar graph	values for each category	highest and lowest values for a variable to be seen[2]
numerical data[1]	histogram	values for data grouped in categories	highest and lowest values for a variable to be seen
categorical data[1]	pie chart	proportions in each category	relative proportions in each of a variable's categories to be seen
numerical data	line graph	values for a variable over time	trend over time to be seen[2]
numerical data	scatter graph	a relationship	relationship between two variables to be seen

[1] Numerical data will need to be grouped into categories.
[2] To also compare variables, use a multiple bar chart or a multiple line graph.

(Saunders & Lewis, 2012: 170)

29

29

Tables

Table 12.6 First-year examination and coursework scores (2)

Student number	Examination score	Coursework score
1	30	70
2	35	65
3	40	60
4	45	55
5	50	50
6	55	45
7	60	40
8	65	35
9	70	30

30

30

	Organic area (ha) (1) 2006	% changes 2007–2008
EU-27	7 764 722	7.4
BE	36 153	10.8
BG	16 663	22.1
CZ	320 311	9.1
DK	150 104	8.7
DE	907 786	4.9
EE	87 346	9.8
IE	42 816	4.1
EL	317 824	13.6
ES	1 317 539	33.3
FR	583 799	4.8
IT	1 002 414	−12.9
CY	2 323	:
LV	161 624	9.1
LT	122 200	1.5
LU	3 535	5.0
HU	122 817	15.0
MT	20	:
NL	50 434	7.3
AT	447 678	0.6
PL	313 944	8.5
PT	233 475	:
RO	140 132	6.6
SI	29 836	1.8
SK	140 755	19.4
FI	150 374	1.1
SE	336 439	9.1
UK	726 381	10.0
NO	52 248	6.9

(1) MT data 2006, CY, PT data 2007

(Saunders & Lewis, 2012: 171)

31

31

Satisfaction with Course

		Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree	Total
Gender	Male	6	13	19	7	5	50
	Female	0	3	9	26	12	50
Total		6	16	28	33	17	100

Satisfaction with Course

		Strongly Disagree	Disagree	Undecided	Agree	Strongly Agree	Total
Ethnic Group	White	2	7	16	11	6	42
	Asian	3	6	10	16	6	41
	Caribbean	0	2	2	2	4	10
	African	1	1	0	4	1	7
Total		6	16	28	33	17	100

32

32

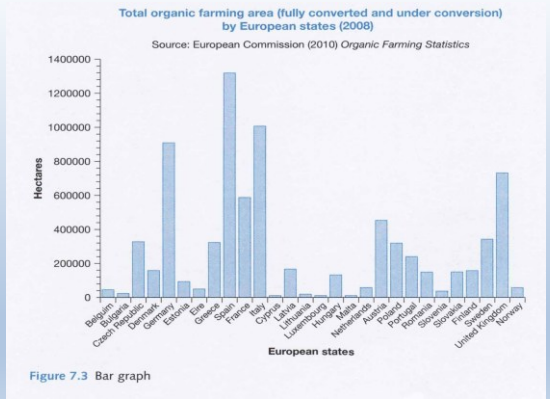
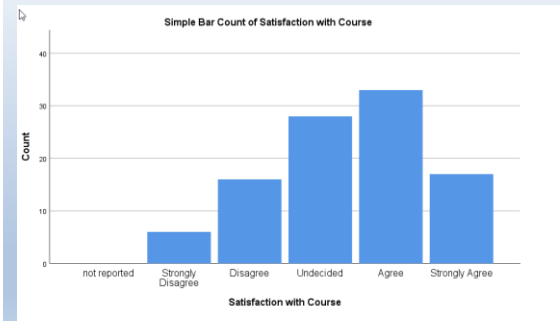


Figure 7.3 Bar graph

(Saunders & Lewis, 2012: 172)

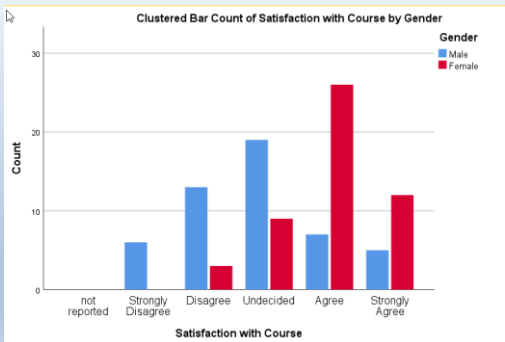
33

33



34

34



35

35

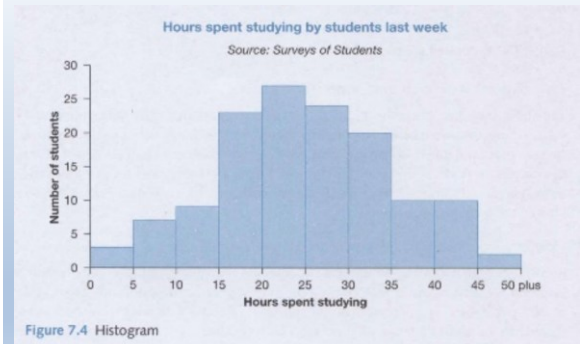


Figure 7.4 Histogram

(Saunders & Lewis, 2012: 173)

36

36

How much time did you spend studying last week? [ ][ ] hours [ ][ ] mins
(For example, for 2 hours, 15 minutes write: [ ][2] hours [1][5] mins)

To draw a histogram, you could place these data into a series of equal width categories such as:

less than 5 hours	25 to less than 30 hours
5 to less than 10 hours	30 to less than 35 hours
10 to less than 15 hours	35 to less than 40 hours
15 to less than 20 hours	40 to less than 45 hours
20 to less than 25 hours	45 hours or more

(Saunders & Lewis, 2012: 173) "bins -> binning" - "unbinning"

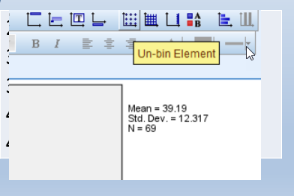
37

37

How much time did you spend studying last week? [ ][ ] hours [ ][ ] mins
(For example, for 2 hours, 15 minutes write: [ ][2] hours [1][5] mins)

To draw a histogram, you could place these data into a series of equal width categories such as:

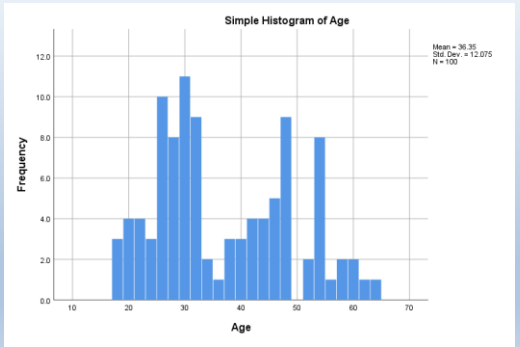
less than 5 hours
5 to less than 10 hours
10 to less than 15 hours
15 to less than 20 hours
20 to less than 25 hours



(Saunders & Lewis, 2012: 173) "bins -> binning" - "unbinning"

38

38



39

39

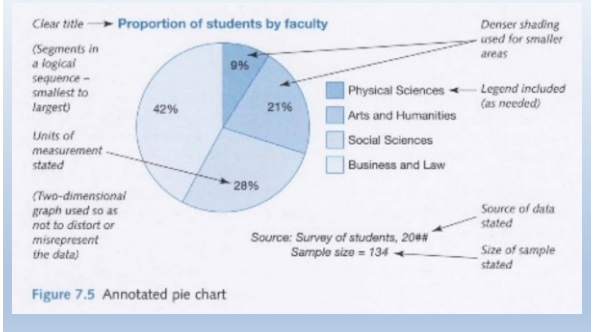
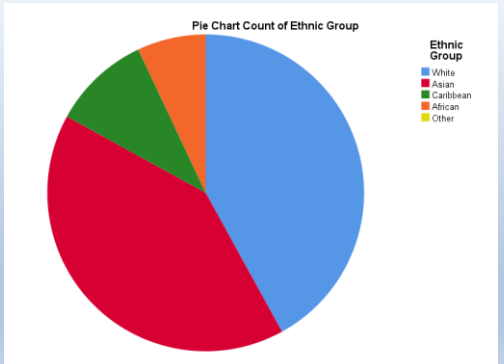


Figure 7.5 Annotated pie chart

(Saunders & Lewis, 2012: 174)

40

40



41

41

(Saunders & Lewis, 2012: 175)

42

42

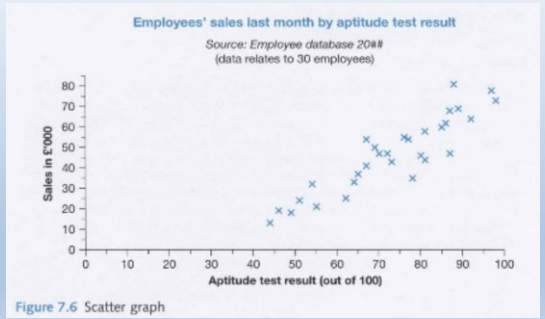


Figure 7.6 Scatter graph

(Saunders & Lewis, 2012: 176)

43

Layout

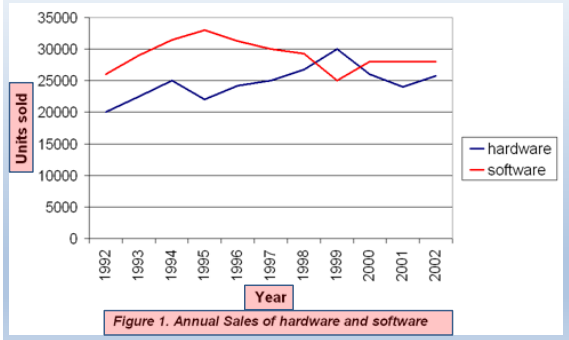


Figure 1. Annual Sales of hardware and software

44

Layout

Table 1
Means and Standard Deviations for Current Samples and APA Normative Sample

	Freudians (n = 30)		Jungians (n = 30)		Adlerians (n = 30)		Normative sample (n = 866)	
Variable	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Factor a	9.33	2.34	8.77	2.53	8.73	2.85	9.15	3.00
Factor b	10.47	3.26	7.97	2.43	8.13	2.86	11.67	4.01
Factor c	8.33	2.28	7.07	2.03	7.83	2.36	10.30	3.36
Factor d	8.97	2.24	8.40	1.77	10.17	3.40	9.94	2.87
Factor e	11.53	3.09	15.20	2.07	9.10	2.86	11.11	3.39
Factor f	9.30	2.58	6.83	2.57	12.73	3.63	11.82	3.75
Factor g	8.27	2.12	7.33	2.15	8.57	2.27	10.11	3.14
Factor h	9.00	1.95	8.47	2.52	9.80	2.95	10.59	3.20
Factor I	45.03	5.96	39.90	7.28	44.50	7.77	52.61	11.35
Factor II	52.03	4.32	58.37	3.62	46.37	5.14	49.29	6.11

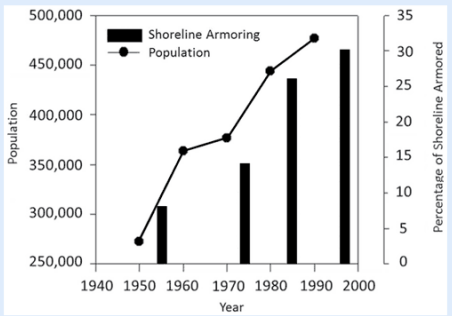
Note. APA = American Psychological Association. For a description of each factor, see text (introduction).

Table/Figure

45

Layout

Figure 1. Population Growth and Shoreline Armoring in Mobile Bay, Alabama.



46

Layout

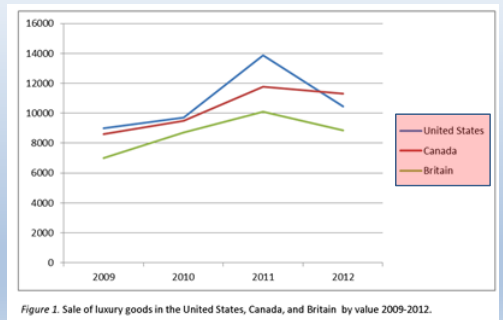


Figure 1. Sale of luxury goods in the United States, Canada, and Britain by value 2009-2012.

47

Layout

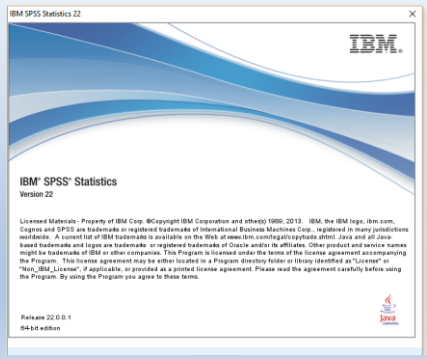
TABLE 2.
UNBLOCKED SHEAR WALLS

Fastener			Panel			Ultimate Loads (plf)					
Size	Spacing[a] (in.)	Length (in.)	Type	Thickness (in.)	No. of Tests	Min.	Max.	Avg.	Stud Spacing (in.)	Target Design Shear (plf)	Load Factor[b]
STRUCTURAL I											
8d	6, 12	2-1/2	Ply	1/2	1			593	16	215	2.8
Nail[c]	6, 6	2-1/2	Ply	1/2	1			955	16	215	4.4
RATED SHEATHING											
	6, 12	2	Ply	5/16	2	550	588[d]	569	16	135	4.2
6d	6, 6	2	Ply	5/16	2	650	750	700	16	135	5.2
Nail[e]	6, 12	2	Ply	3/8	1			300	24	100	3.0
	6, 6	2	Ply	3/8	1			375	24	100	3.8
	6, 12	1-1/2	Ply	15/32	1			346	16	130	2.7
16 ga Staple	6, 12	2	Ply	15/32	1			381	16	130	2.9
	6, 12	1-3/4	OSB	3/8	1			335	16	105	3.2
15 ga Staple	6, 12	1-1/2	Ply	15/32	1			294	24	105	2.8

(a) The first number is the spacing at supported panel edges, the second is the spacing at intermediate framing members. Single nail at each stud along the unblocked panel edge except as noted.
(b) The load factor is determined by dividing the average ultimate load by the target design shear.
(c) Stainless steel common nail.
(d) At the unblocked edge an additional nail was placed along each panel edge and 3/4" away from the regular nailing.
(e) Common nail.

48

IBM SPSS Statistics

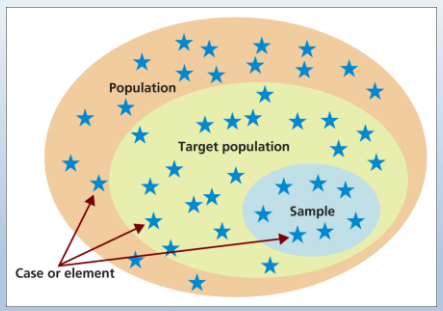


IBM SPSS Statistics

Awesome though it is, SPSS is not a magic oven that can miraculously transform garbage input into haute cuisine output. To get useful output, you need properly collected data and carefully considered statistical analysis.

(Colman & Pulford, 2008: 2)

Data Collection



Your sample must represent your population.

(Saunders et al, 2016: 275)

Describing Data using Statistics

Descriptive	Inferential
Description	Inference
Sample	Population

"When an individual uses descriptive statistics, he talks about the data he has; but with inferential statistics, he talks about data he does not have."

(Popham & Sirotnik, 1973: 40)

Describing Data using Statistics

Description

Inference

- Exploring Relationships
- Comparing Groups

Describing Data using Statistics

Description:

- Organise data into comprehensible form so that trends etc. can be communicated to others.
- Simplify the organisation and presentation of data.
- Reduce the data.

Frequency Distribution

Numerical data: Age Distribution

Student	Age
1	25
2	27
3	25
4	23
5	24
6	26
7	25
8	26
9	28
10	22
11	24
12	23

55

55

Frequency Distribution

Numerical data: Age Distribution

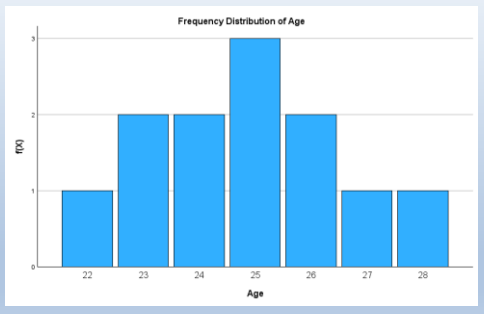
X	*f*(X)	*cf*(X)	*p*(X)	*cp*(X)
22	1	1	.085	.085
23	2	3	.17	.25
24	2	5	.17	.42
25	3	8	.25	.67
26	2	10	.17	.83
27	1	11	.085	.91
28	1	12	.085	1

X = age of students
f(X) = frequency – number of students at that age
cf(X) = cumulative frequency - number of students at that age or younger
p(X) = proportion – proportion of students at that age (relative frequency)
cp(X) = cumulative proportion - proportion of students at that age or younger

56

56

Frequency Distribution



57

57

Describing Data using Statistics

Description:

- The central tendency, which is the common, middle or average value.
- The dispersion, which is how the data values are spread, or dispersed, around the central tendency.

58

58

Describing Central Tendency

Categorical data:

The Mode

The category that occurs most often

59

59

Describing Central Tendency

Categorical data: The Mode

Participant	Country
1	France
2	Poland
3	France
4	Portugal
5	Spain
6	Germany
7	France
8	Germany
9	Finland
10	Italy
11	Spain
12	Portugal

60

60

Describing Central Tendency

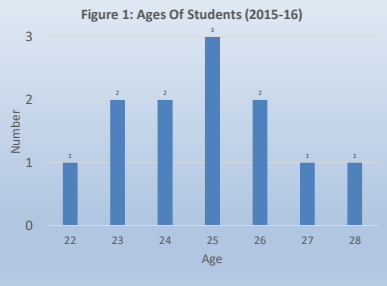
Numerical data: The Mode

Participant	Age
1	25
2	27
3	25
4	23
5	24
6	26
7	25
8	26
9	28
10	22
11	24
12	23

61

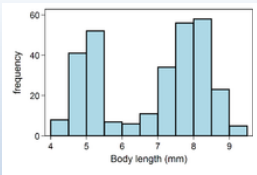
61

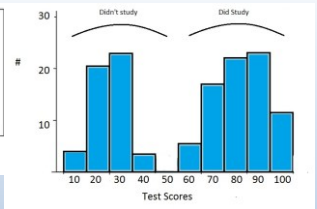
Describing Central Tendency

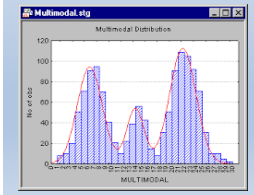


62

62







Bi-modal

Multi-modal

63

63

Describing Central Tendency

Categorical data: The Median

22, 23, 23, 24, 24, 25, 25, 25, 26, 26, 27, 28

Quartiles (1, 2, 3)

Percentiles (25, 50, 75)

64

64

Describing Central Tendency

Categorical data: The Mean

22+23+23+24+24+25+25+25+26+26+27+28

= 298 /12 = 24.83

65

65

Describing Central Tendency

Statistics

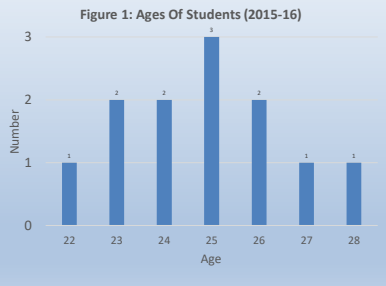
Age

N	Valid	12
	Missing	0
Mean		24.83
Median		25.00
Mode		25
Percentiles	25	23.25
	50	25.00
	75	26.00

66

66

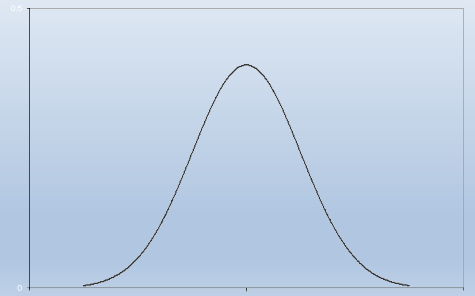
Describing Central Tendency



67

67

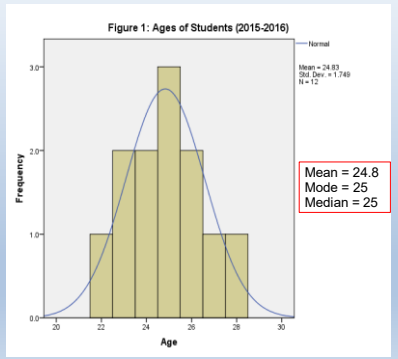
Normal Distribution



68

68

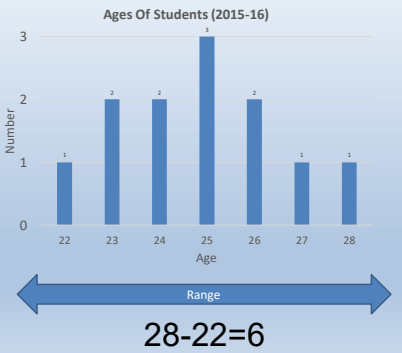
Describing Central Tendency



69

69

Dispersion

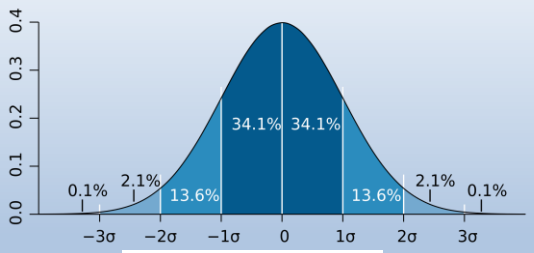


28-22=6

70

70

Standard Deviation - σ

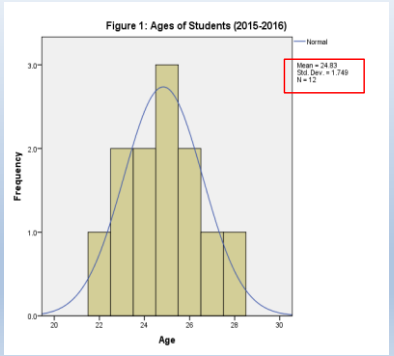


$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

71

71

Dispersion



72

72

IQR

Dispersion

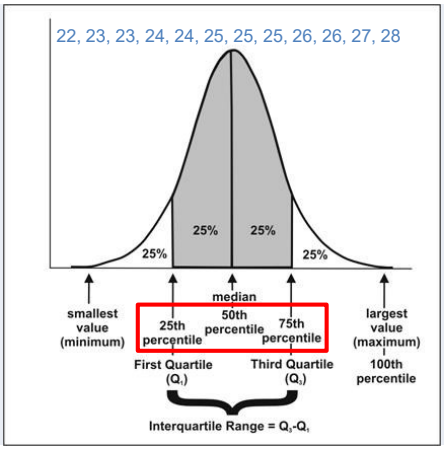
Categorical data: The Median

22, 23, 23, 24, 24, 25, 25, 25, 26, 26, 27, 28

Quartiles

73

73



74

74

Describing Central Tendency & Dispersion

Mode – most common

Median – middle value of rank order

Mean – average

Dispersion/Range – highest minus lowest

Standard Deviation – deviation from mean

IQR – middle 50% range

75

75

Describing Central Tendency & Dispersion

Descriptives

			Statistic	Std. Error
Age	Mean		24.83	.505
	95% Confidence Interval for Mean	Lower Bound	23.72	
		Upper Bound	25.94	
	5% Trimmed Mean		24.81	
	Median		25.00	
	Variance		3.061	
	Std. Deviation		1.749	
	Minimum		22	
	Maximum		28	
	Range		6	
	Interquartile Range		3	
	Skewness		.181	.637
	Kurtosis		-.428	1.232

76

76

Table 7.2 Describing a variable using statistics

For...	use the...	to describe the...	which represents the...
categorical data	mode	central tendency	category that occurs most often
numerical data	mode *or the*		value that occurs most often
	median *or the*		middle value when all the data values are put in rank order
	mean		average value
numerical data	range *or the*	dispersion	difference between the highest and lowest data values when they are put in rank order
	inter-quartile range *or the*		difference within the middle 50 per cent of data values when they are put in rank order
	standard deviation[1]		extent the data values differ from the mean (95 per cent of data will lie with plus or minus 1.96 standard deviations of the mean)
numerical data	index number[2]	trend	relative change in a series of data values over time

[1] only use if the data are normally distributed.
[2] only use if the data have been collected over time.

(Saunders & Lewis, 2012: 172)

77

77

Example: Descriptive Statistics

Mean sales for the organisation's 30 employees were £46,600. As the mean, median and mode are virtually the same, this suggests these data are normally distributed. Consequently the standard deviation of 18.46 indicates that 95 per cent of sales fell within the range £10,318 to £82,682, the complete range being £68,000.

(Saunders & Lewis, 2012: 178)

78

78

Describing Data using Statistics

Description

- Central Tendency
- Dispersion

Inference

- Exploring Relationships
- Comparing Groups

79

79

Describing Data using Statistics

Description

Inference

- Exploring Relationships
- Comparing Groups

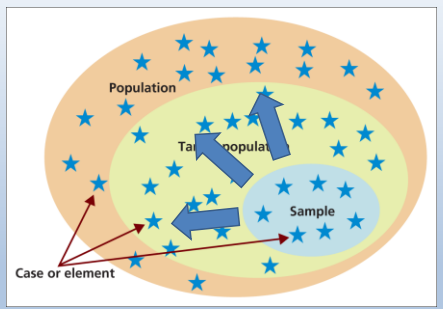
"When an individual uses descriptive statistics, he talks about the data he has; but with inferential statistics, he talks about data he does not have."

(Popham & Sirotnik, 1973: 40)

80

80

Sample – Population



Inferring

(Saunders et al, 2016: 275)

81

81

Describing Data using Statistics

Description

Inference

- Exploring Relationships
- Comparing Groups

"When an individual uses descriptive statistics, he talks about the data he has; but with inferential statistics, he talks about data he does not have."

(Popham & Sirotnik, 1973: 40)

82

82

Describing Data using Statistics

Relationships:

- The association between two variables.
- The correlation between two variables.

Correlation/correlate; Association/associate; Relationship/relate; Covariance/covariation/covary

83

83

Table 7.3 Examining relationships between variables using statistics

For . . .	use . . . (symbol in brackets)	to examine . . .	which represents the . . .
categorical data	chi-square test (χ^2)	association	probability of the association between two variables occurring by chance
categorical ranked data	Spearman's rank correlation coefficient (ρ) or	correlation	strength of the relationship between two variables and the probability of this occurring by chance
	Kendal's rank correlation coefficient (τ)	correlation (when data contains tied ranks)	strength of the relationship between two variables and the probability of this occurring by chance
numerical data split into two groups using a categorical variable	independent groups t-test (t)	difference	probability of the differences between the values in the two groups occurring by chance
numerical data split into three-plus groups using a categorical variable	Analysis of variance (ANOVA) (F)	difference	probability of the differences between the values in the groups occurring by chance
pairs of numerical data for two variables measuring the same feature under different conditions	paired t-test (t)	difference	probability of the differences between each half of the pair (the two variables) occurring by chance

(Saunders & Lewis, 2012: 180)

84

84

Table 7.3 Examining relationships between variables using statistics

For . . .	use . . . (symbol in brackets)	to examine . . .	which represents the . . .
categorical data	chi-square test (χ^2)	association	probability of the association between two variables occurring by chance
categorical ranked data	Spearman's rank correlation coefficient (ρ) or	correlation	strength of the relationship between two variables and the probability of this occurring by chance
	Kendal's rank correlation coefficient (τ)	correlation (when data contains tied ranks)	strength of the relationship between two variables and the probability of this occurring by chance

(Saunders & Lewis, 2012: 180)

85

85

Table 7.3 Examining relationships between variables using statistics (*continued*)

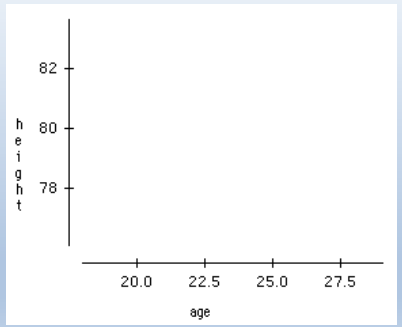
For . . .	use . . . (symbol in brackets)	to examine . . .	which represents the . . .
numerical data	Pearson's product moment correlation coefficient (r)	correlation	strength of the relationship between two variables and the probability of this occurring by chance
	Regression coefficient (coefficient of determination) (r^2)	explanation	strength of a cause and effect relationship between a dependent and one or more independent variables and the probability of this occurring by chance
	Regression equation ($y = a + b\,x$)	prediction	formula to predict the values of a dependent variable, given the values of one or more independent variables

(Saunders & Lewis, 2012: 181)

86

86

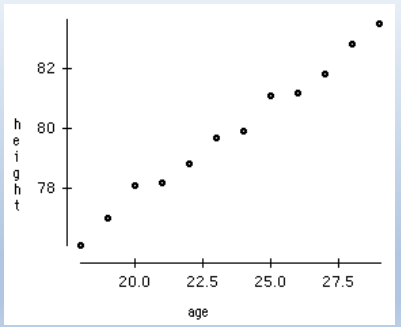
Describing Data using Statistics



87

87

Correlation



88

88

Correlation

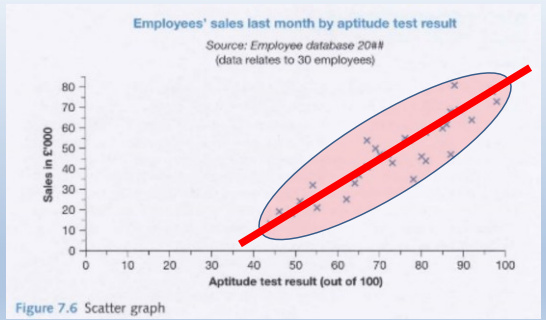


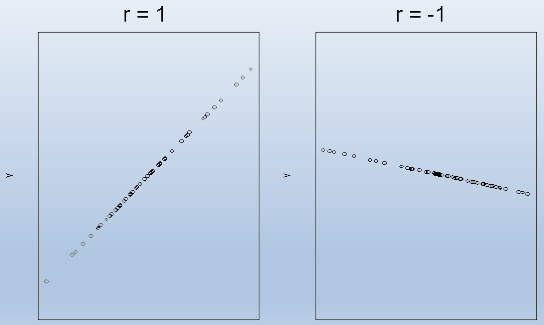
Figure 7.6 Scatter graph

(Saunders & Lewis, 2012: 176)

89

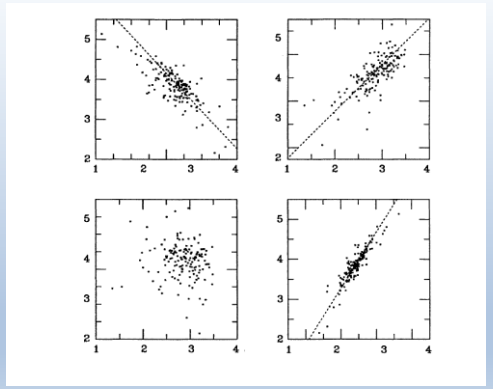
89

Correlation Coefficient (r)



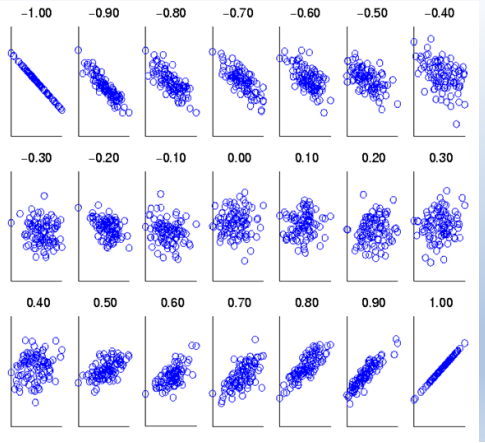
90

90



91

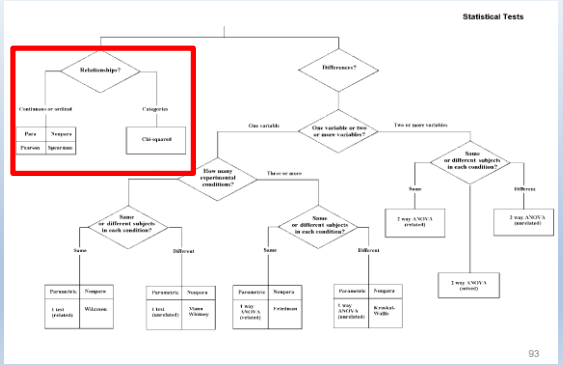
91



92

92

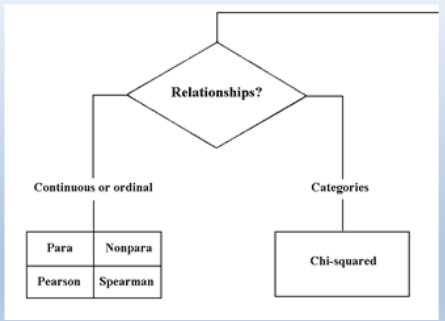
Relationship between Groups



93

93

Relationship between Groups



94

94

Describing Data using Statistics

Examining associations

Sex: male or female

Salary grade: 1-9

Both categorical data - not ranked

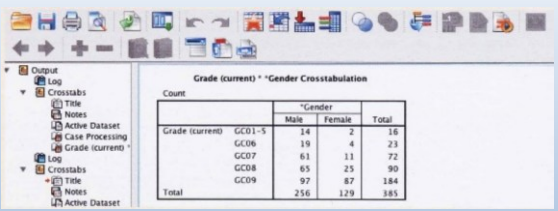
The chi-squared test.

You use this statistic to test whether there is an association between someone's salary grade and their sex: chi-squared $\chi 2$

95

95

Crosstabulation



(Saunders, Lewis & Thornhill, 2012: 516)

96

96

Crosstabulation

Grade (current) * *Gender Crosstabulation

Count

		*Gender		
		Male	Female	Total
Grade (current)	GC01-5	14	2	16
	GC06	19	4	23
	GC07	61	11	72
	GC08	65	25	90
	GC09	97	87	184
Total		256	129	385

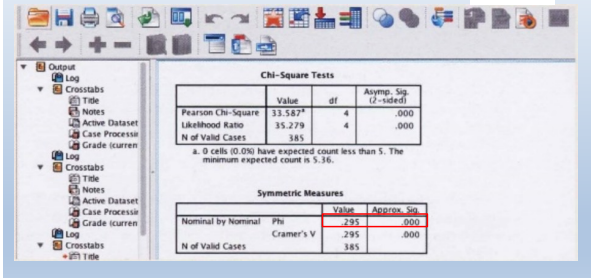
(Saunders, Lewis & Thornhill, 2012: 516)

97

Chi-Squared Test

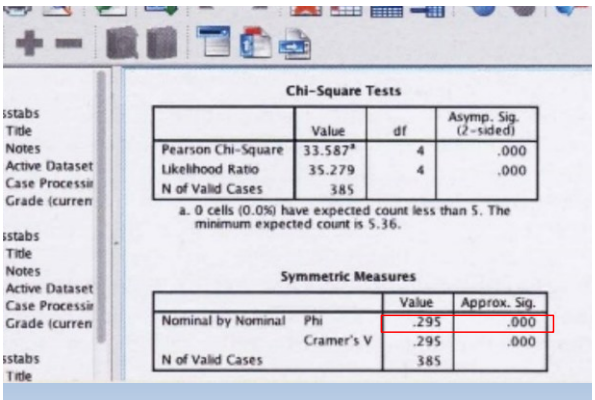
$$\chi_c^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

χ^2



(Saunders, Lewis & Thornhill, 2012: 516)

98



(Saunders, Lewis & Thornhill, 2012: 516)

99

Chi-Squared Test

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.295	.000
	Cramer's V	.295	.000
N of Valid Cases		385	

(Saunders, Lewis & Thornhill, 2012: 516)

100

Correlation Coefficients

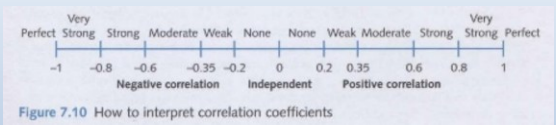


Figure 7.10 How to interpret correlation coefficients

(Saunders & Lewis, 2012: 183)

101

Relationships

Categorical data

- Chi-squared coefficient

Ranked data.

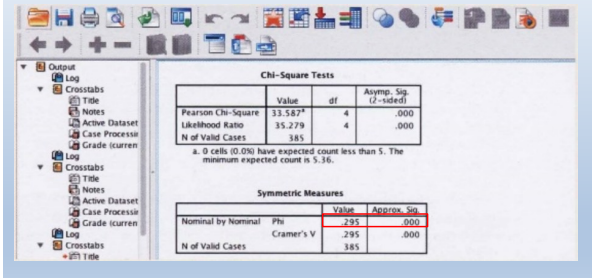
- Kendall's rank correlation coefficient
- Spearman's rank correlation coefficient.

Numerical data

- Pearson's product moment correlation coefficient

102

Chi-Squared Test



(Saunders, Lewis & Thornhill, 2012: 516)

103

103

Chi-Squared Test

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.295	.000
	Cramer's V	.295	.000
N of Valid Cases		385	

(Saunders, Lewis & Thornhill, 2012: 516)

104

104

Significance

How likely the relationship between your variables has occurred by chance.

Significance testing.

If this probability is small (usually 0.05 or less), then you can say your relationship is statistically significant.

!!

105

105

Interpreting Correlational Research

Even random numbers show some degree of correlation.

To what extent could your result have happened randomly?

Is it significant?

106

106

Interpreting Correlational Research

Is it significant?
Significance (alpha level) p
Significant at p<.05
There is less than a 5% chance that the result is due to chance.
This means that the finding has a 95% chance of being true.

Sometimes significant at p<.01
However, see Shaveleson (1988: 248)

107

107

Chi-Squared Test

Symmetric Measures

		Value	Approx. Sig.
Nominal by Nominal	Phi	.295	.000
	Cramer's V	.295	.000
N of Valid Cases		385	

There is a statistically significant weak positive relationship between the sex of the employee and their salary grade (r =.295, p < 0.001).

(Saunders, Lewis & Thornhill, 2012: 516)

108

108

Relationships

How to examine relationships

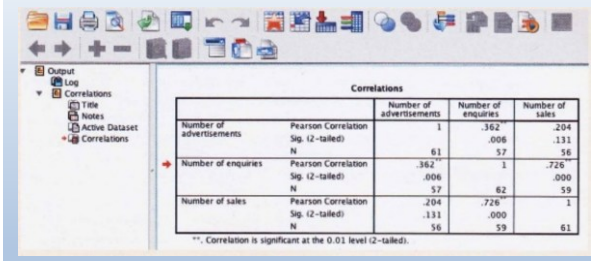
Numerical/Continuous data.

Pearson's Product Moment Correlation

- Number of television advertisements.
- Number of enquiries.
- Number of sales.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\,n\Sigma x^2 - (\Sigma x)^2\,]\,[\,n\Sigma y^2 - (\Sigma y)^2\,]}}$$

109



(Saunders, Lewis & Thornhill, 2012: 522)

110

110

Advertisements/Enquiries/Sales

Correlations

		Number of advertisements	Number of enquiries	Number of sales
Number of advertisements	Pearson Correlation	1	.362**	.204
	Sig. (2-tailed)		.006	.131
	N	61	57	56
Number of enquiries	Pearson Correlation	.362**	1	.726**
	Sig. (2-tailed)	.006		.000
	N	57	62	59
Number of sales	Pearson Correlation	.204	.726**	1
	Sig. (2-tailed)	.131	.000	
	N	56	59	61

**. Correlation is significant at the 0.01 level (2-tailed).

(Saunders, Lewis & Thornhill, 2012: 522)

111

111

Advertisements/Enquiries/Sales

Correlations

		Number of advertisements	Number of enquiries	Number of sales
Number of advertisements	Pearson Correlation	1	.362**	.204
	Sig. (2-tailed)		.006	.131
	N	61	57	56
Number of enquiries	Pearson Correlation	.362**	1	.726**
	Sig. (2-tailed)	.006		.000
	N	57	62	59
Number of sales	Pearson Correlation	.204	.726**	1
	Sig. (2-tailed)	.131	.000	
	N	56	59	61

**. Correlation is significant at the 0.01 level (2-tailed).

(Saunders, Lewis & Thornhill, 2012: 522)

112

112

Advertisements/Enquiries/Sales

Correlations

		Number of advertisements	Number of enquiries	Number of sales
Number of advertisements	Pearson Correlation	1	.362**	.204
	Sig. (2-tailed)		.006	.131
	N	61	57	56
Number of enquiries	Pearson Correlation	.362**	1	.726**
	Sig. (2-tailed)	.006		.000
	N	57	62	59
Number of sales	Pearson Correlation	.204	.726**	1
	Sig. (2-tailed)	.131	.000	
	N	56	59	61

**. Correlation is significant at the 0.01 level (2-tailed).

(Saunders, Lewis & Thornhill, 2012: 522)

113

113

Example

There is a statistically significant strong positive relationship between the number of enquiries and the number of sales ($r = .726$, $p < 0.001$) and a statistically significant but weak to moderate relationship between the number of television advertisements and the number of enquiries ($r = .362$, $p = 0.006$). However, there is no statistically significant relationship between the number of television advertisements and the number of sales ($r = .204$, $p = 0.131$).

(Saunders, Lewis & Thornhill, 2012: 522)

114

114

Interpreting Correlational Research

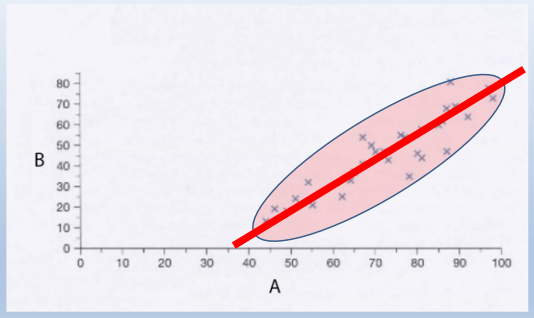
Statistical Significance ≠ Meaningful or Important

Correlation ≠ Causality

115

115

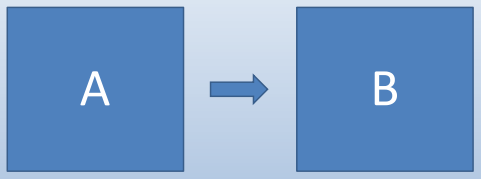
We see a relationship between A & B



116

116

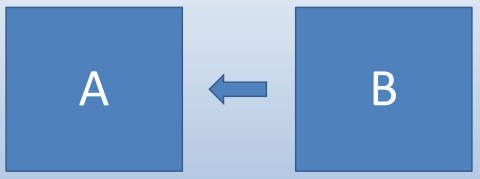
It might be:



117

117

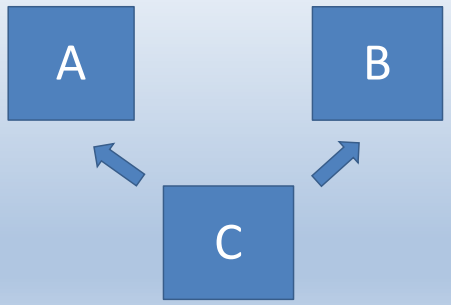
But it could be be:



118

118

Or:



Third variable/Confounding variable

119

119

Dog owners have lower mortality, study finds

Katie Silver
Health reporter, BBC News

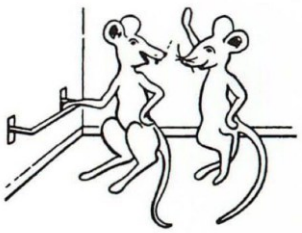
22 hours ago | Health

Dog owners have a lower risk of death from cardiovascular disease or other causes, a study of 3.4 million Swedes has found.

The team analysed national registries for people aged 40 to 80, and compared them to dog ownership registers.

120

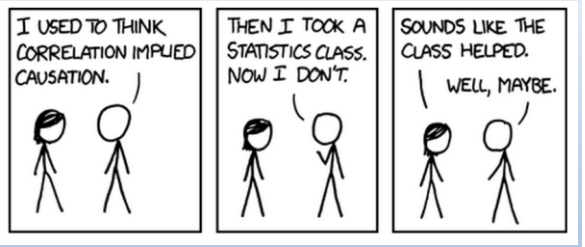
120



"Boy, do we have this guy conditioned. Every time I press the bar down he drops a pellet in."

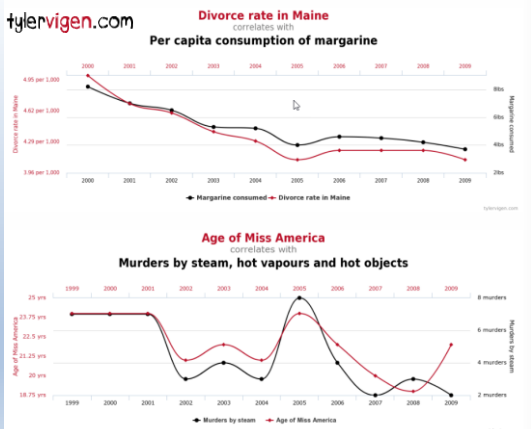
121

121



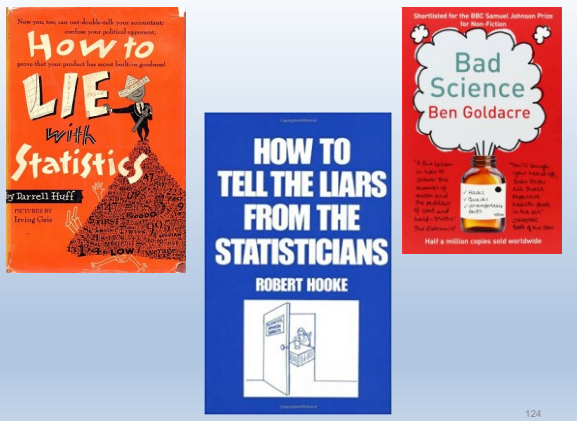
122

122



123

123



124

124

Regression

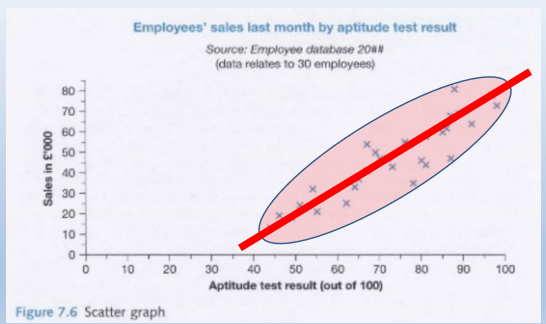


Figure 7.6 Scatter graph

(Saunders & Lewis, 2012: 176)

125

125

Regression

A study looked at the possibility of predicting students' future reading ability (READING) from their intelligence score (IQ).
Ten children were included in the investigation, selected at random from a primary school class. The children's IQ scores were obtained when they entered the school and the corresponding reading score was obtained at the end of the first year. The reading score was the result out of 100 on a test.

Statistical Services and Consultancy Unit (2016: 36-42)

126

126

A study looke g students' future readin r intelligence s Ten children on, selected at random fr children's IQ scores we he school and the corre ained at the end of the fir he result out of 100 on

Statistical Services a

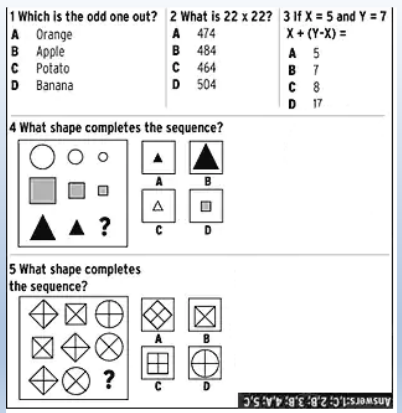
STATISTICS WITH SPSS
(Version 23)

UNIVERSITY OF HERTFORDSHIRE

STATISTICAL SERVICES & CONSULTANCY UNIT

127

IQ



128

IQ : Reading

Pupil No.	1	2	3	4	5	6	7	8	9	10
IQ Score	98	115	124	97	135	105	117	101	110	127
Reading	54	74	81	59	88	70	78	73	84	86

129

Hypotheses

H_0: READING score cannot be predicted from IQ score

H_1: READING score can be predicted from IQ score

H_0 = ***null hypothesis*** - no relationship between two measured phenomena.

Rejecting or disproving the **null hypothesis** is a central task in modern scientific practice.

Statistically significant means that the null hypothesis has been rejected.

130

Hypotheses

H_0 = ***null hypothesis*** - no relationship between two measured phenomena.

Rejecting or disproving the **null hypothesis** is a central task in modern scientific practice.

Statistically significant means that the null hypothesis has been rejected.

Not Statistically Significant means that the null hypothesis cannot be rejected.

131

Regression

A study looked at the possibility of predicting students' future reading ability (READING) from their intelligence score (IQ).
Ten children were included in the investigation, selected at random from a primary school class. The children's IQ scores were obtained when they entered the school and the corresponding reading score was obtained at the end of the first year. The reading score was the result out of 100 on a test.

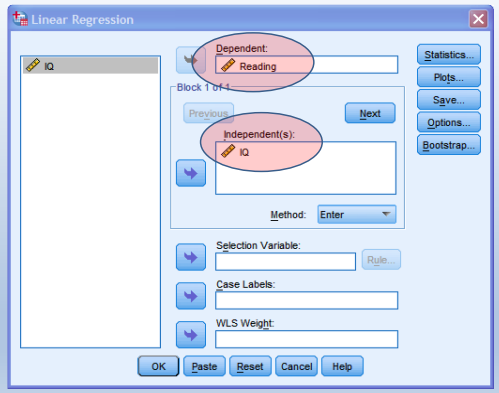
$$x = a + by$$

$$a = \frac{(\Sigma y)(\Sigma x^2) - (\Sigma x)(\Sigma xy)}{n(\Sigma x^2) - (\Sigma x)^2}$$

$$b = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{n(\Sigma x^2) - (\Sigma x)^2}$$

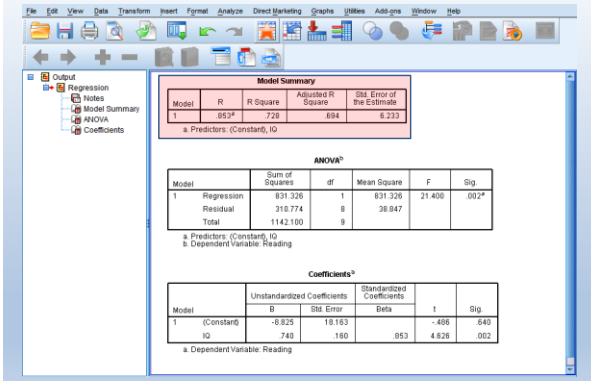
Statistical Services and Consultancy Unit (2016: 36-42)

132



133

133



134

134

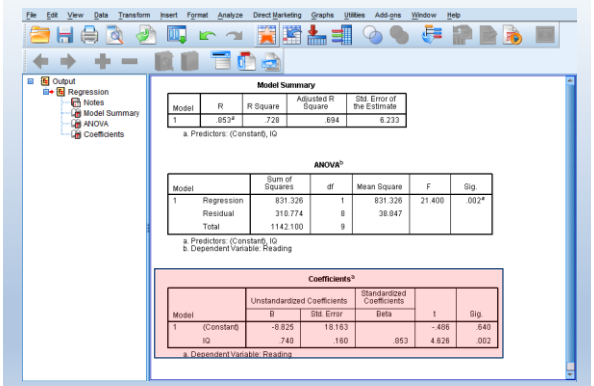
Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.853[a]	.728	.694	6.233

a. Predictors: (Constant), IQ

135

135



136

136

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-8.825	18.163		-.486	.640
	IQ	.740	.160	.853	4.626	.002

a. Dependent Variable: Reading

x = a + by

a = -8.825
b = .740

137

137

Regression

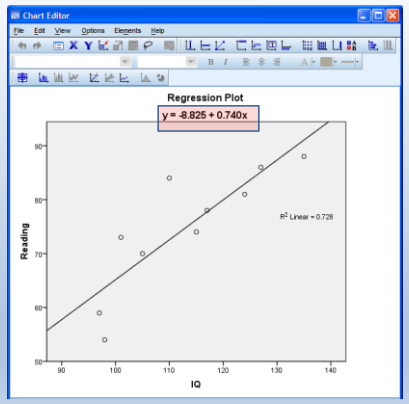
The regression equation

"READING = −8.825 + 0.740 × IQ"

can be obtained by equating the dependent variable **Reading** to the *CONSTANT* value −8.825, plus the independent variable **IQ** multiplied by its coefficient 0.740.

138

138



139

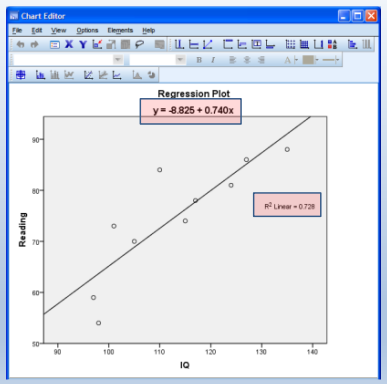
139

Example

The test statistic found from SPSS is *t*, which has the value 4.626 with 8 *df* and associated *p*-value of 0.002. Since *p* is very small, i.e. below 0.05, the gradient is unlikely to be a chance result. Hence it is concluded that there is a relationship between IQ and reading score for primary school children. Hence, we can reject H_0 in favour of H_1. For every primary school child whose IQ is known the reading score can be predicted.

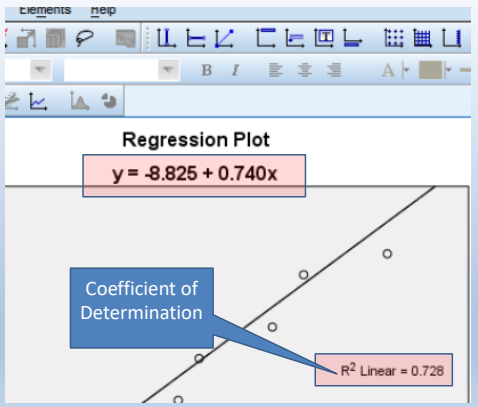
140

140



141

141



142

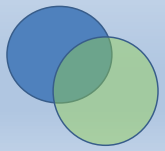
142

Coefficient of Determination

R^2

A measure of how much variance is shared, how much of one variable can be explained by another.

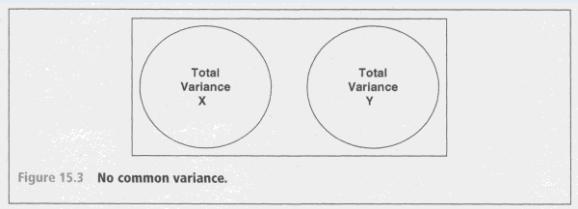
$\equiv r^2$



143

143

Coefficient of Determination



If the correlation (r) between variable X and variable Y = 0, then the coefficient of determination = $0^2 = 0$ or 0%. There is no common variability.

(Burns & Burns, 2008: 349)

144

144

Coefficient of Determination

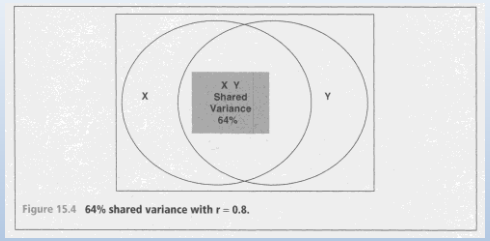


Figure 15.4 **64% shared variance with r = 0.8.**

If the correlation (r) between variable X and variable Y = 0.8, then the coefficient of determination = 0.8^2 = .64 or 64%. 64% of the variance in one variable is predictable from the variance in the other.

(Burns & Burns, 2008: 349)

145

145

Coefficient of Determination

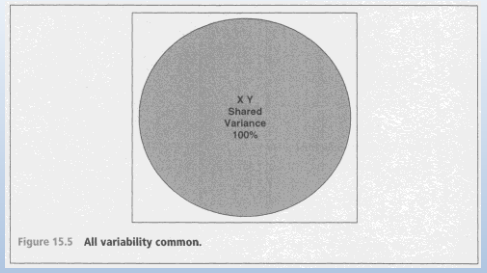


Figure 15.5 **All variability common.**

If the correlation (r) between variable X and variable Y = 1, then the coefficient of determination = 1^2 = 1 x 100 = 100%: 100% of the variability is common to both variables.

(Burns & Burns, 2008: 350)

146

146

Coefficient of Determination

Example
Consider the relationship between the number of customer calls made by a particular salesman and the number of orders obtained during the course of a 10-day journey cycle. Suppose the correlation is +.85. This shows a fairly high positive correlation between the number of calls a salesman makes and the number of orders received. The coefficient of determination is 72.25%, i.e. that 72% of variance in orders is explained by the number of calls made.

(Burns & Burns, 2008: 350)

147

147

Coefficient of Determination

R^2

A measure of how much variance is shared, how much of one variable can be explained by another.

Effect Size

148

148

Describing Data using Statistics

Description

Inference

- Exploring Relationships
- Comparing Groups

149

149

Table 7.3 Examining relationships between variables using statistics

For . . .	*use . . . (symbol in brackets)*	*to examine . . .*	*which represents the . . .*
categorical data	chi-square test (χ^2)	association	probability of the association between two variables occurring by chance
categorical ranked data	Spearman's rank correlation coefficient (ρ) or	correlation	strength of the relationship between two variables and the probability of this occurring by chance
	Kendal's rank correlation coefficient (τ)	correlation (when data contains tied ranks)	strength of the relationship between two variables and the probability of this occurring by chance
numerical data split into two groups using a categorical variable	independent groups *t*-test (*t*)	difference	probability of the differences between the values in the two groups occurring by chance
numerical data split into three-plus groups using a categorical variable	Analysis of variance (ANOVA) (*F*)	difference	probability of the differences between the values in the groups occurring by chance
pairs of numerical data for two variables measuring the same feature under different conditions	paired *t*-test (*t*)	difference	probability of the differences between each half of the pair (the two variables) occurring by chance

(Saunders & Lewis, 2012: 180)

150

150

numerical data split into two groups using a categorical variable	independent groups t-test (*t*)	difference	probability of the differences between the values in the two groups occurring by chance
numerical data split into three-plus groups using a categorical variable	Analysis of variance (ANOVA) (*F*)	difference	probability of the differences between the values in the groups occurring by chance
pairs of numerical data for two variables measuring the same feature under different conditions	paired *t*-test (*t*)	difference	probability of the differences between each half of the pair (the two variables) occurring by chance

(Saunders & Lewis, 2012: 180)

151

151

Comparing Groups: Difference

Table 7.3

Example purchase data

	Product A *M (sd)*	Product B *M (sd)*
Women N = 17	18.5 (2.5)	22.5 (2.0)
Men N = 17	14.5 (3.5)	21.5 (2.5)

152

152

Comparing Means

153

153

Comparing Means

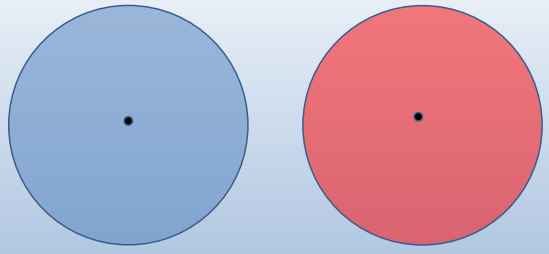
154

154

Comparing Means

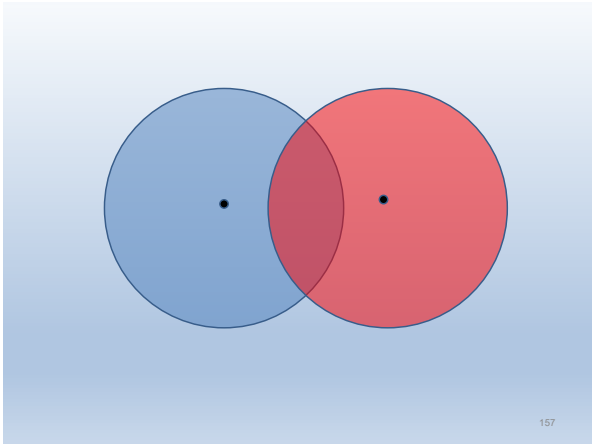
155

155

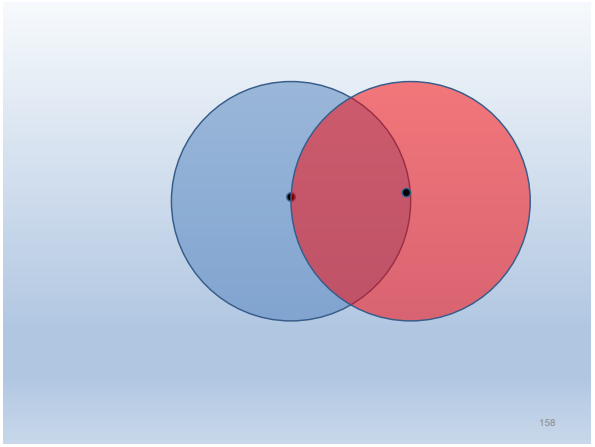


156

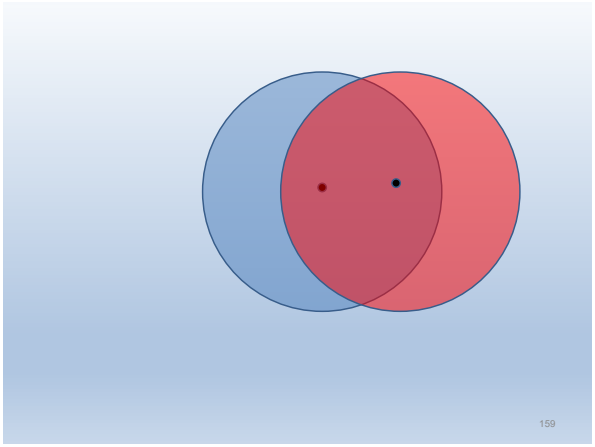
156



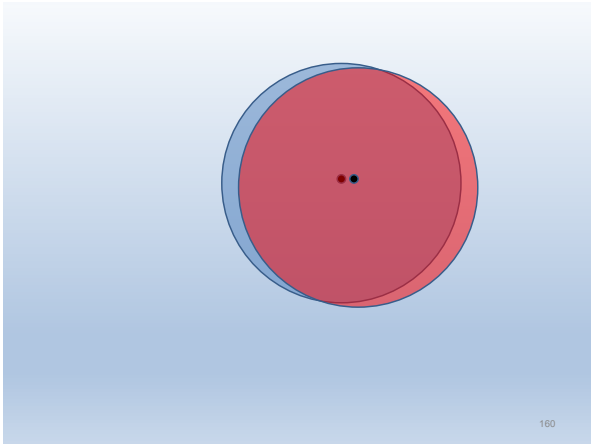
157



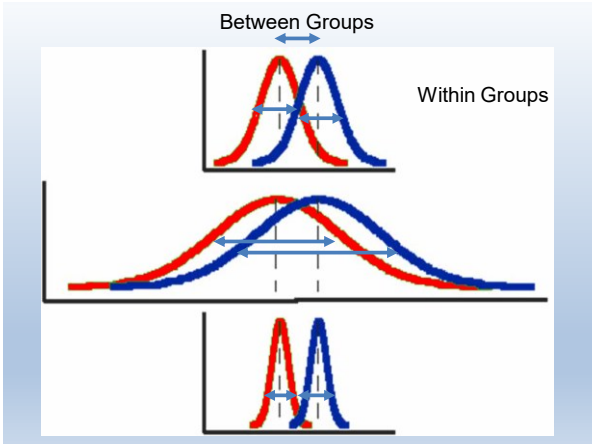
158



159



160



161

t-test

Comparing Means

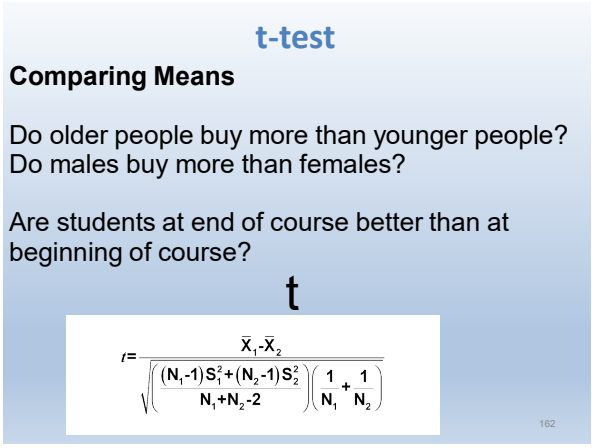
Do older people buy more than younger people?
Do males buy more than females?

Are students at end of course better than at beginning of course?

t

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\left(\frac{(N_1-1)S_1^2+(N_2-1)S_2^2}{N_1+N_2-2}\right)\left(\frac{1}{N_1}+\frac{1}{N_2}\right)}}$$

162



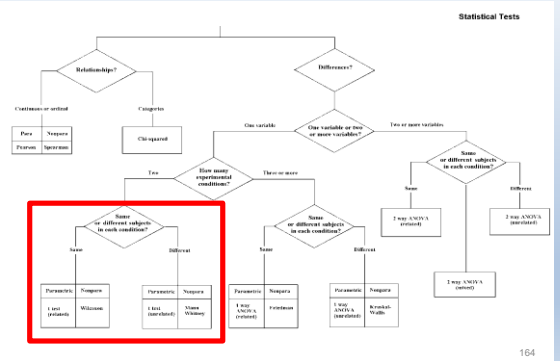
162

Exploring Differences between Groups

t-test (unrelated) – independent samples t-test
Compares 2 unrelated groups (e.g. different towns) on same test. Need one categorical independent variable with only two groups (e.g. M/F) and one continuous dependent variable (e.g. items purchased).

t-test (related) – paired samples / matched t-tests / matched pairs
Compares the same group on 2 occasions (e.g. same product at two times). Need one categorical independent variable (e.g. time 1 & time 2) and one continuous dependent variable (e.g. product).

Comparison of Groups



An experiment is carried out to see if announcing the items on a conveyor belt as they pass helps children remember the items better.
Twenty-two children, all aged 12 years, were randomly allocated; 10 to the announcement group, and the remaining 12 to the no-announcement group.
*The children were shown the items on the conveyor belt **with** or **without** announcing each item as they passed.*

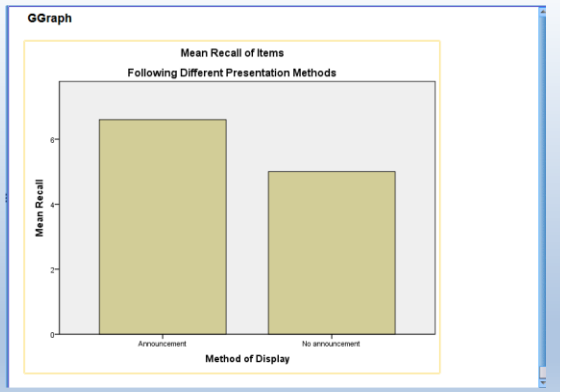
Statistical Services and Consultancy Unit (2016: 15-19)

The number of items recalled out of each of the ten letters in the list was recorded for each child:

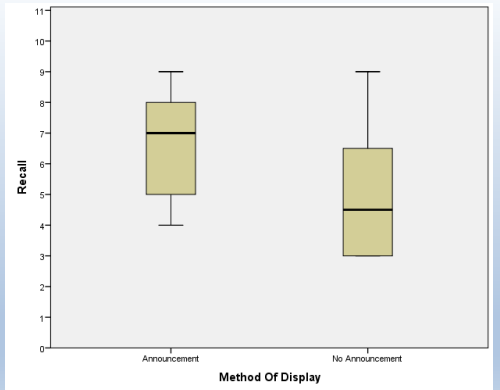
Announcement: 5 7 8 9 5 4 9 8 7 4
No Announcement: 3 8 5 6 3 5 9 3 3 4 7 4

Announcement Mean: 6.6
No Announcement Mean: 5.0

Announcement SD: 1.9
No Announcement SD: 2.0



(Statistical Services and Consultancy Unit, 2016: 17)



Generalisation of result to wider population of twelve-year-old children

H_0: the use of announcements makes no difference
H_1: the use of announcements makes a difference

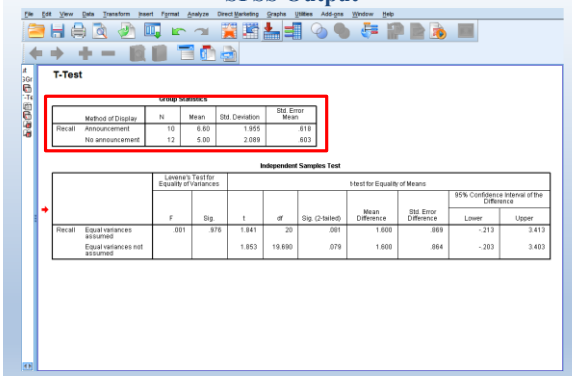
H_0 = **null hypothesis -** no relationship between two measured phenomena.

Rejecting or disproving the null hypothesis is a central task in modern scientific practice.

169

169

SPSS Output



(Statistical Services and Consultancy Unit, 2016: 18)

170

170

SPSS Output

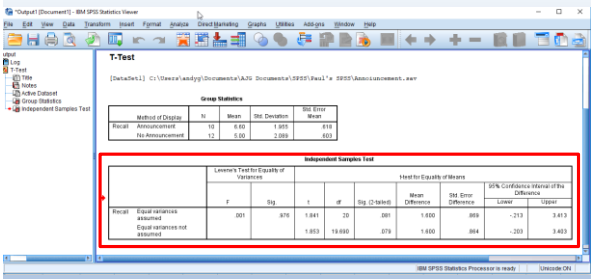
Group Statistics

	Method of Display	N	Mean	Std. Deviation	Std. Error Mean
Recall	Announcement	10	6.60	1.955	.618
	No Announcement	12	5.00	2.089	.603

171

171

SPSS Output: Significance



(Statistical Services and Consultancy Unit, 2016: 18)

172

172

SPSS Output: Significance

Independent Samples Test

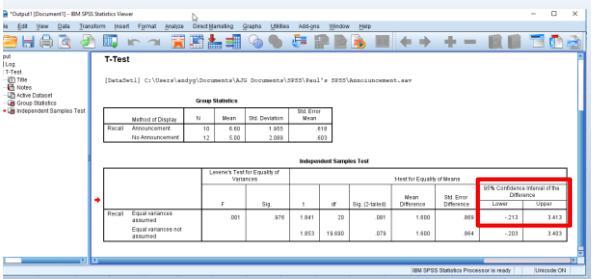
	t-test for Equality of Means				
	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference
	1.841	20	.081	1.600	.869
	1.853	19.690	.079	1.600	.864

(Statistical Services and Consultancy Unit, 2016: 18)

173

173

SPSS Output: Confidence Interval



(Statistical Services and Consultancy Unit, 2016: 18)

174

174

SPSS Output: Confidence Interval

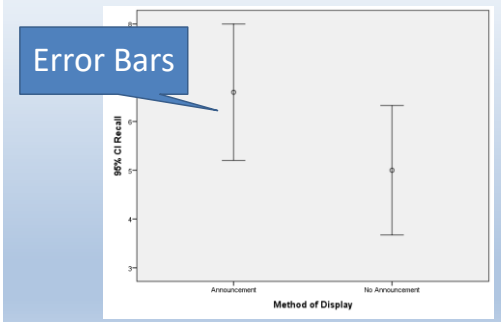
t-test for Equality of Means

Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
		Lower	Upper
1.600	.869	-.213	3.413
1.600	.864	-.203	3.403

175

175

SPSS Output: Confidence Interval



176

176

Effect Size

- Cohen's d = 0.79
- η^2 Eta Squared = 0.14

Benchmarks (Cohen, 1988)

	Small	Medium	Large
d	.20	.50	.80
η^2	.01	.15	.35

(Burns & Burns, 2008, chap. 11)

177

177

Generalisation of result to wider population of twelve-year-old children

H_0 : the use of announcements makes no difference
H_1 : the use of announcements makes a difference

The *p*-value is the probability that the amount of difference found between the means of the two groups could be a chance result. If *p* were quite small, e.g. below 0.05, the difference would be unlikely to be a chance result.

178

178

Generalisation of result to wider population of twelve-year-old children

H_0 : the use of announcements makes no difference
H_1 : the use of announcements makes a difference

Since our *p* was quite large 0.081, > 0.05, we can conclude that the difference may well be a chance result.

The conclusion is that there is no benefit to the wider population from the use of announcements. The apparent benefit found in the sample of 22 children was a chance finding due to the luck of sampling and random allocation. This apparent benefit of announcements cannot be generalised.

179

179

Generalisation of result to wider population of twelve-year-old children

H_0 : the use of announcements makes no difference
H_1 : the use of announcements makes a difference

An independent-samples t-test was conducted to compare the effect of the use announcements on the memory of 22 children. There was no significant difference in scores for the announcement group (M=6.60, SD=1.95) and the no-announcement group (M=5, SD=2.089; t(20) = 1.853, *p* = 0.081, two-tailed) The magnitude of the difference in the means (mean difference = 1.6, 95% *CI*: -.213 to 3.413) was medium (eta squared = 0.14).

180

180

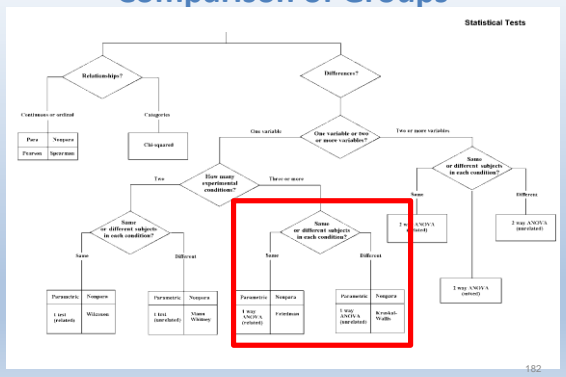
Analysis of Variance

If three or more distinct groups:

Use one-way analysis of variance or ANOVA.

181

Comparison of Groups



182

Data for Analysis of Variance

A sample of adults was selected and randomly allocated to four groups. All were asked to consume amounts of alcohol as follows:

Group 1: NO ALCOHOL

Group 2: ONE UNIT OF ALCOHOL

Group 3: TWO UNITS OF ALCOHOL

Group 4: THREE UNITS OF ALCOHOL

after which they were asked drive a car simulator and perform an emergency stop at 40 miles per hour.

(Statistical Services and Consultancy Unit, 2016: 26-31)

183

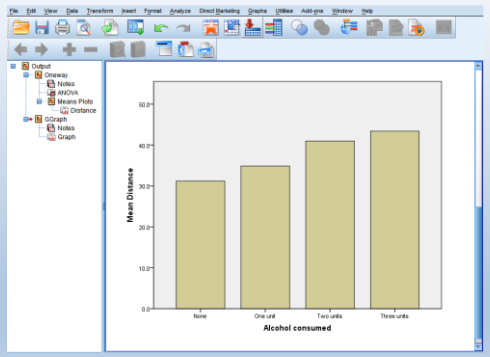
Data for Analysis of Variance

The distance to stopping the simulator in metres by each group member is shown below.

The results were as follow.

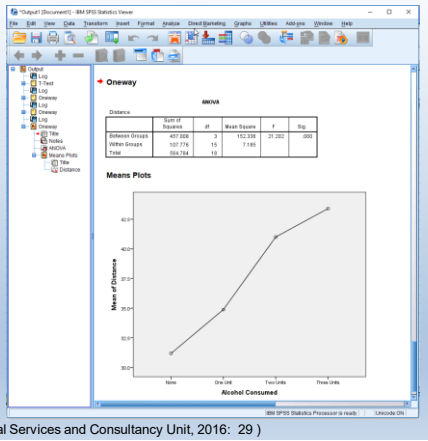
NO ALCOHOL	31.2	27.4	29.8	33.5	34.0
ONE UNIT	31.2	33.7	34.3	39.2	36.0
TWO UNITS	42.1	40.6	43.4	37.9	
THREE UNITS	46.7	45.1	43.2	41.7	40.3

184



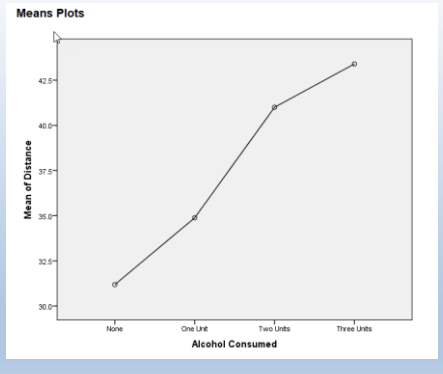
(Statistical Services and Consultancy Unit, 2016: 30)

185



(Statistical Services and Consultancy Unit, 2016: 29)

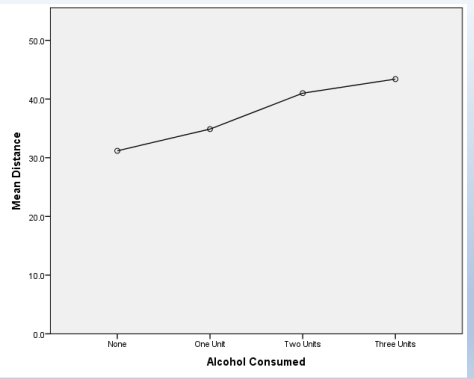
186



(Statistical Services and Consultancy Unit, 2016: 29)

187

187



188

188

$$F^{*} = \frac{\frac{\sum w_k \left(\bar{X}_k - X'\right)^2}{k-1}}{1+\frac{2(k-2)}{k^2-1}\sum\left(\frac{1}{n_k-1}\right)\left(1-\frac{w_k}{\sum w_k}\right)^2}$$

ANOVA

Distance

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	457.008	3	152.336	21.202	.000
Within Groups	107.776	15	7.185		
Total	564.784	18			

(Statistical Services and Consultancy Unit, 2016: 29)

189

189

Post Hoc Tests

Multiple Comparisons

Dependent Variable: Distance

Tukey HSD

(I) Alcohol Consumed	(J) Alcohol Consumed	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
None	One Unit	-3.7000	1.6953	.173	-8.586	1.186
	Two Units	-9.8200*	1.7981	.000	-15.002	-4.638
	Three Units	-12.2200*	1.6953	.000	-17.106	-7.334
One Unit	None	3.7000	1.6953	.173	-1.186	8.586
	Two Units	-6.1200*	1.7981	.018	-11.302	-.938
	Three Units	-8.5200*	1.6953	.001	-13.406	-3.634
Two Units	None	9.8200*	1.7981	.000	4.638	15.002
	One Unit	6.1200*	1.7981	.018	.938	11.302
	Three Units	-2.4000	1.7981	.556	-7.582	2.782
Three Units	None	12.2200*	1.6953	.000	7.334	17.106
	One Unit	8.5200*	1.6953	.001	3.634	13.406
	Two Units	2.4000	1.7981	.556	-2.782	7.582

*. The mean difference is significant at the 0.05 level.

190

190

Describing Data using Statistics

Description

- Central Tendency
- Dispersion

Inference

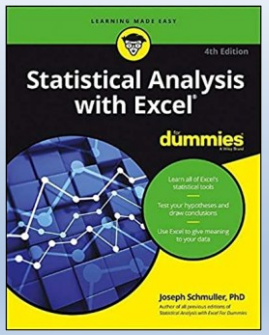
– Exploring Relationships
- Correlation
- Regression

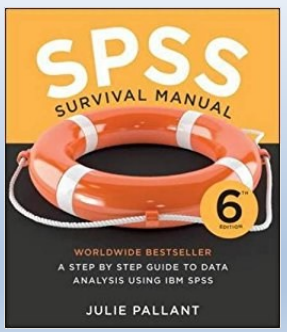
– Comparing Groups
- t-Test
- ANOVA

191

191

Reading

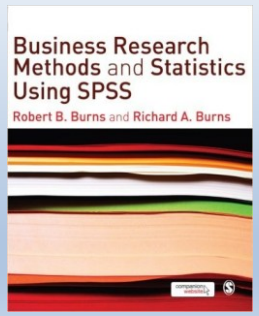


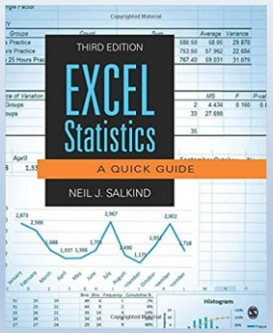


192

192

Reading

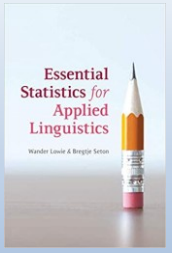


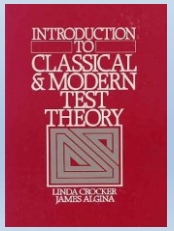


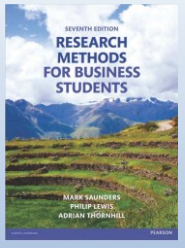
193

193

Reading







194

194

References

Aldrich, J. O. & Rodriguez, H. M. (2013). *Building SPSS graphs to understand data*. London: Sage.

Anderson, D., Sweeney, D. & Williams, T. (2001). *Modern business statistics with Microsoft Excel*. New York: Cengage.

Burns, R. B. & Burns, R. A. (2008). *Business research methods and statistics using SPSS*. London: Sage.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* New York: Routledge.

Colman, A. M. & Pulford, B. D. (2008). *A crash course in SPSS for Windows* (4th ed.). Oxford: Blackwell.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. NewYork: Holt, Rinehart and Winston.

Davis, G. & Pecar, B. (2013). *Business statistics using Excel*. Oxford: Oxford University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press

Hand, D. J. (2008) *Statistics: A very short introduction*. Oxford: Oxford University Press.

Jelen, B. (2013). *Excel 2013 charts and graphs*. Indianapolis: Que.

Lee, N. & Peters, M. (2015). *Business statistics using Excel and SPSS*. London: Sage.

Lowie, W. & Seton, B. (2013). *Essential statistics for applied linguistics*. London: Palgrave Macmillan.

Pallant, J. (2016). *SPSS survival manual* (6th ed.). London: McGraw Hill.

Popham, W. J. , & Sirotnik, K. A. (1973). *Educational statistics: Use and interpretation* (2nd edn). New

Salkind, N. J. (2011*). Statistics for people who (think they) hate statistics* (4th ed.). London: Sage.

Salkind, N. J. (2015). *Excel statistics: A quick guide* (3rd ed.). London: Sage.

Saunders, M. & Lewis, P. (2012). *Doing research in business and management*. London: FT/Prentice Hall.

Saunders, M., Lewis, P. & Thornhill, A. (2016). *Research methods for business students* (7th ed.). London: FT/Prentice Hall.

Schmuller, J. (2016). *Statistical analysis with Excel for dummies* (4th edn). London: Wiley.

Shavelson, R. (1988). *Statistical reasoning for the behavioral sciences*. Boston: Allyn and Bacon.

Statistical Services and Consultancy Unit (2016). *Statistics with SPSS* (Version 23). Hatfield: University of Hertfordshire.

195

195