دانشگاه علوم پزشکی و خدمات بهداشتی درمانی ایران

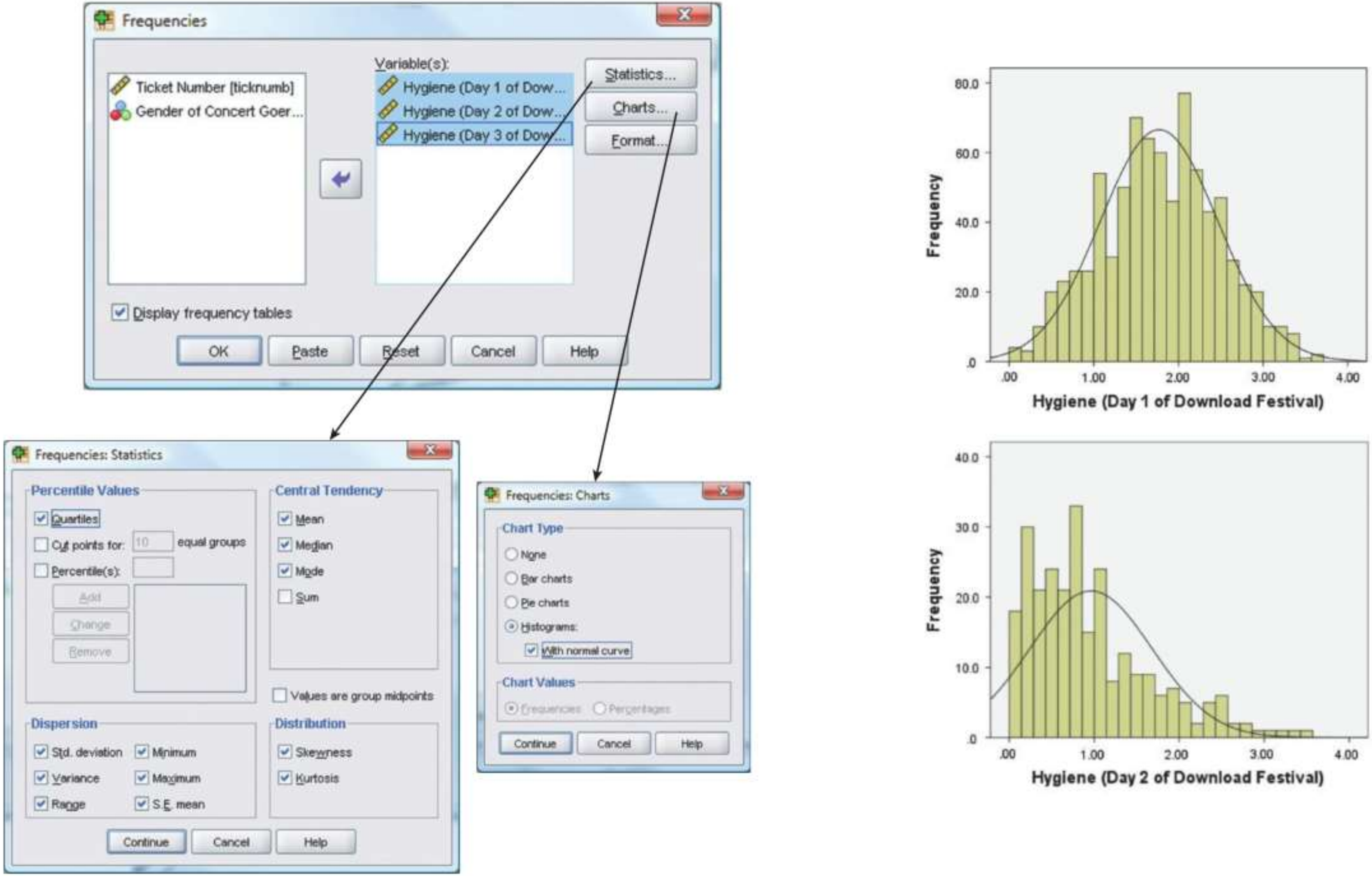
Commonly used statistics in medical research part II

Nader Sadigh MD.

Ali Asghar Children's Hospital

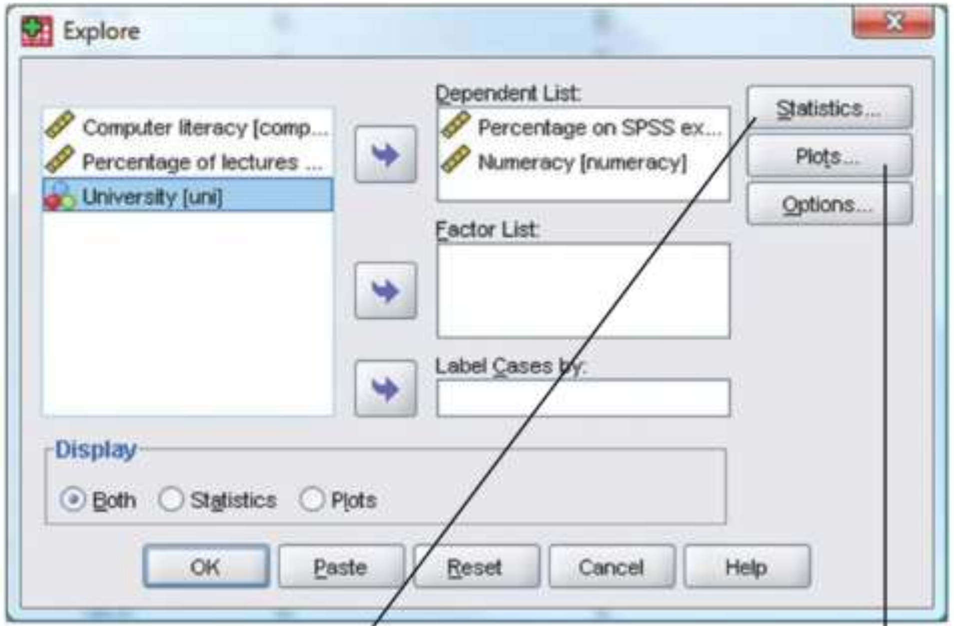
EXPLORING ASSUMPTIONS

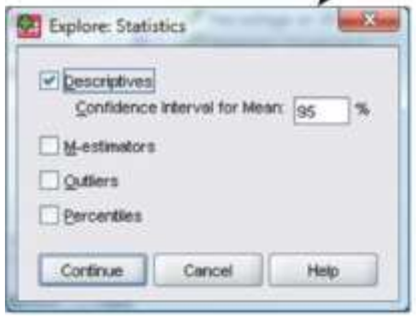
- **1 Normally distributed data**:
- the rationale behind hypothesis testing relies on having something that is normally distributed
  - sampling distribution
  - errors in the model

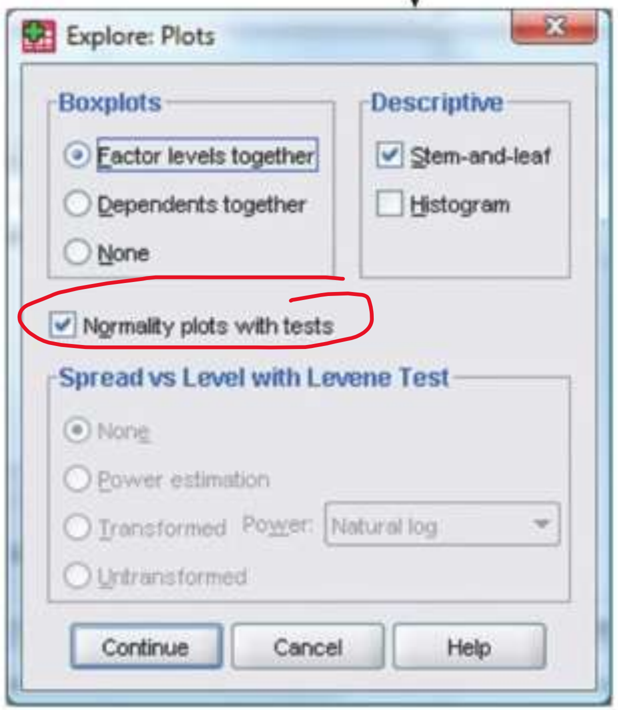

Frequencies
Variable(s):
Ticket Number [ticknumb]
Gender of Concert Goer...
Hygiene (Day 1 of Dow...
Hygiene (Day 2 of Dow...
Hygiene (Day 3 of Dow...
Statistics...
Charts...
Format...
Display frequency tables
OK
Paste
Reset
Cancel
Help
Frequencies: Statistics
Percentile Values
Quartiles
Cut points for: 10 equal groups
Percentile(s):
Add
Change
Remove
Central Tendency
Mean
Median
Mode
Sum
Values are group midpoints
Dispersion
Std. deviation
Minimum
Variance
Maximum
Range
S.E. mean
Distribution
Skewness
Kurtosis
Continue
Cancel
Help
Frequencies: Charts
Chart Type
None
Bar charts
Pie charts
Histograms:
With normal curve
Chart Values
Frequencies
Percentages
Continue
Cancel
Help
Frequency
80.0
60.0
40.0
20.0
.0
.00
1.00
2.00
3.00
4.00
Hygiene (Day 1 of Download Festival)
Frequency
40.0
30.0
20.0
10.0
.0
.00
1.00
2.00
3.00
4.00
Hygiene (Day 2 of Download Festival)

Testing whether a distribution is normal

- to see whether the distribution as a whole deviates from a comparable normal distribution.
- The **Kolmogorov–Smirnov test** and **Shapiro–Wilk test**
- compare the scores in the sample to a normally distributed set of scores with the same mean and standard deviation.
- If test is non-significant ($p > .05$) it tells us that the distribution of the sample is not significantly different from a normal distribution (i.e. it is probably normal).







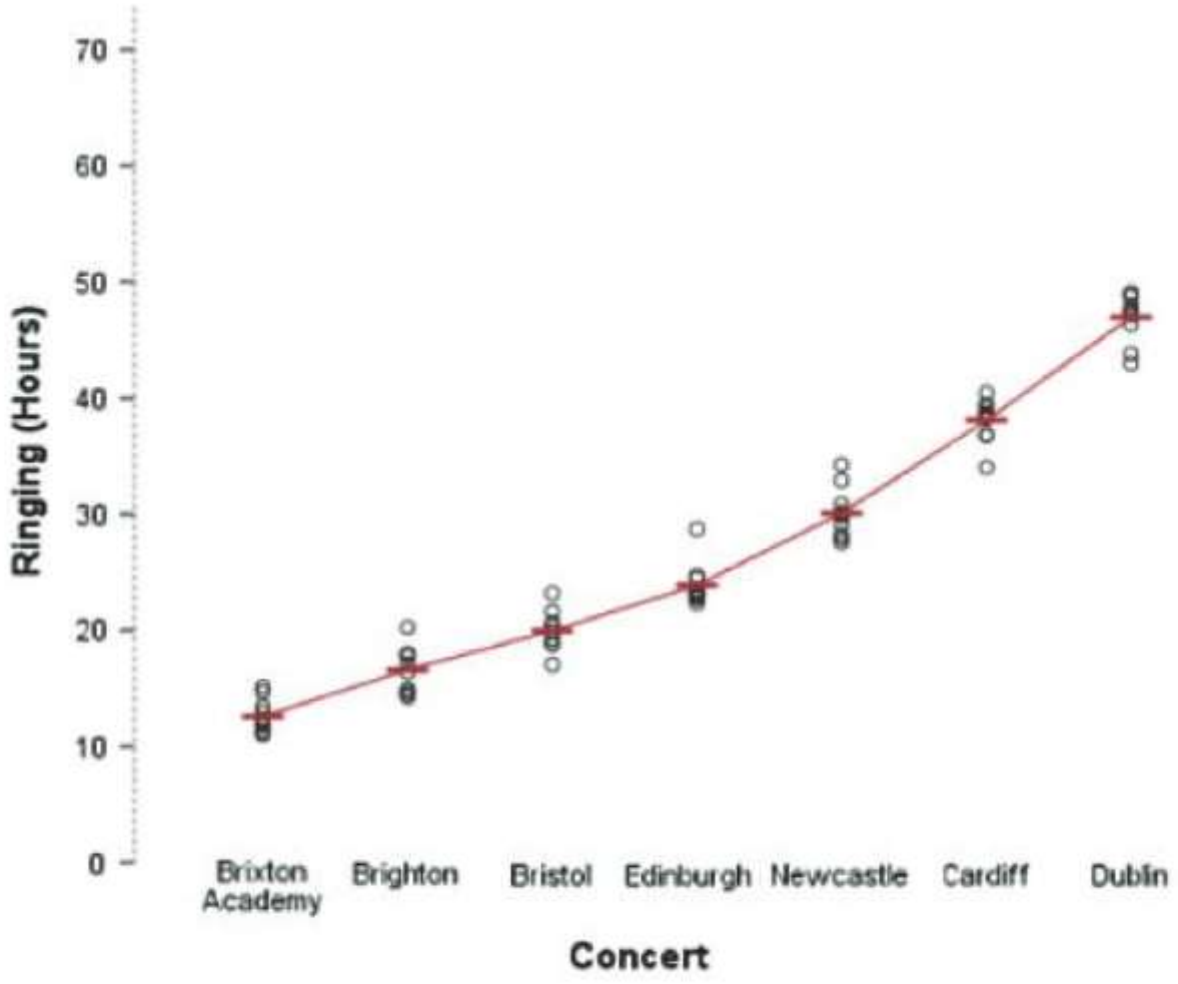
Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Percentage on SPSS exam	.102	100	.012	.961	100	.005
Numeracy	.153	100	.000	.924	100	.000

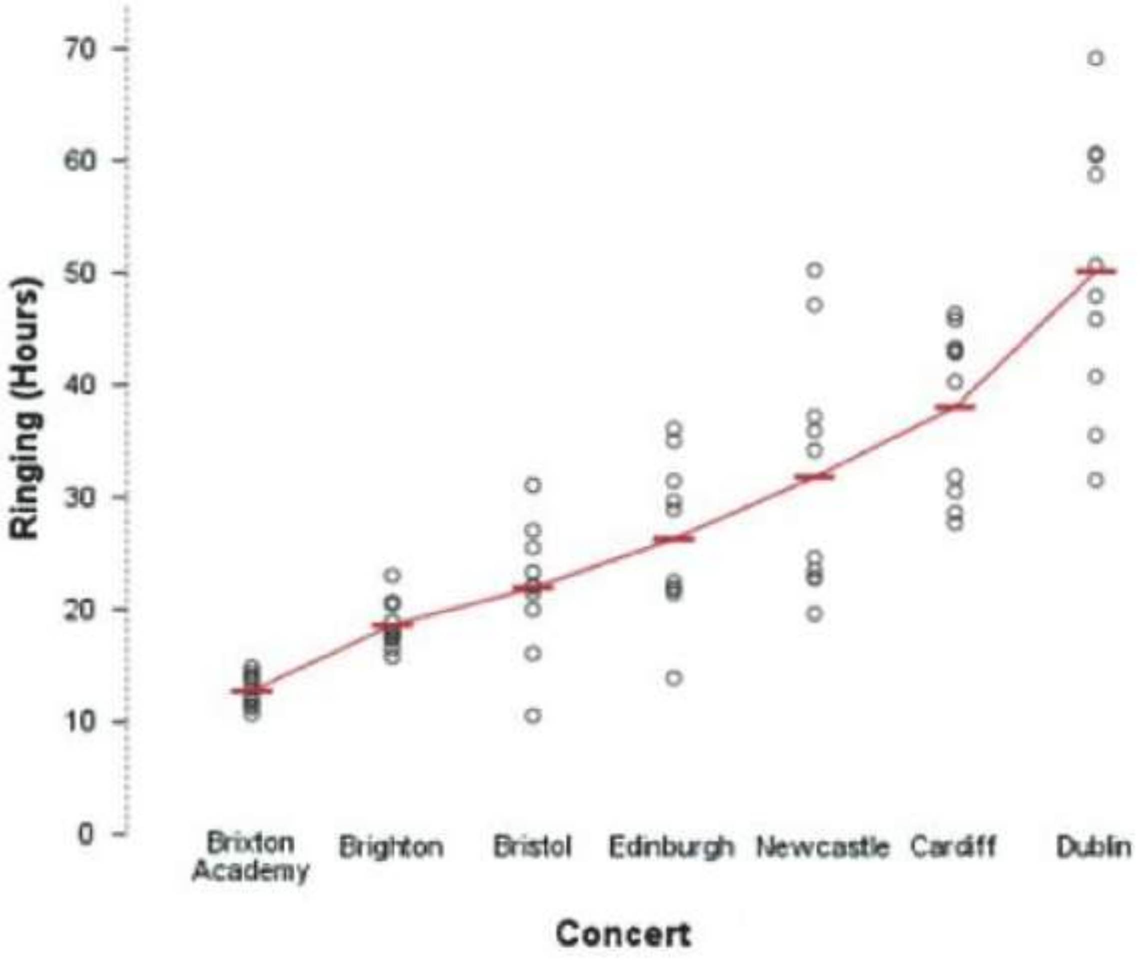
a. Lilliefors Significance Correction

2 Homogeneity of variance:

- variances should be the same throughout the data.
- In designs in which you test several groups of participants this assumption means that each of these samples comes from populations with the same variance.
- In correlational designs, this assumption means that the variance of one variable should be stable at all levels of the other variable.



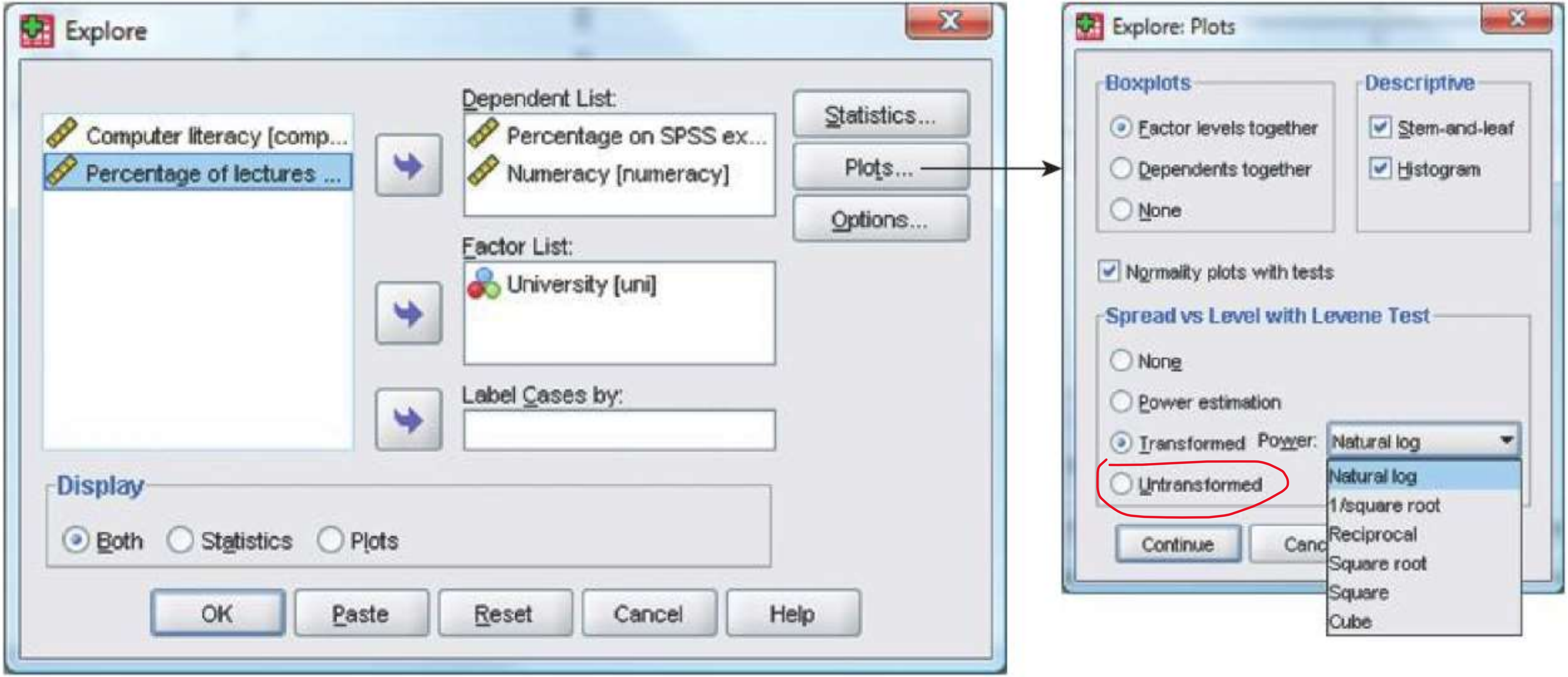
Ringing (Hours)
70
60
50
40
30
20
10
0
Brixton Academy
Brighton
Bristol
Edinburgh
Newcastle
Cardiff
Dublin
Concert



Ringing (Hours)
70
60
50
40
30
20
10
0
Brixton Academy
Brighton
Bristol
Edinburgh
Newcastle
Cardiff
Dublin
Concert

Levene's test

- tests the null hypothesis that the variances in different groups are equal (i.e. the difference between the variances is zero).


Explore
Computer literacy [comp...
Percentage of lectures ...
Dependent List:
Percentage on SPSS ex...
Numeracy [numeracy]
Factor List:
University [uni]
Label Cases by:
Statistics...
Plots...
Options...
Display
Both
Statistics
Plots
OK
Paste
Reset
Cancel
Help
Explore: Plots
Boxplots
Factor levels together
Dependents together
None
Descriptive
Stem-and-leaf
Histogram
Normality plots with tests
Spread vs Level with Levene Test
None
Power estimation
Transformed
Power:
Natural log
Untransformed
Natural log
1/square root
Reciprocal
Square root
Square
Cube
Continue
Canc

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
Percentage on SPSS exam	Based on Mean	2.584	1	98	.111
	Based on Median	2.089	1	98	.152
	Based on Median and with adjusted df	2.089	1	94.024	.152
	Based on trimmed mean	2.523	1	98	.115
Numeracy	Based on Mean	7.368	1	98	.008
	Based on Median	5.366	1	98	.023
	Based on Median and with adjusted df	5.366	1	83.920	.023
	Based on trimmed mean	6.766	1	98	.011

- **3 Interval data:**
- Data should be measured at least at the interval level. This assumption is tested by common sense and so won’t be discussed further

- **4 Independence**:
- like that of normality, is different depending on the test you're using.
- data from different participants are independent
  - the behaviour of one participant does not influence the behaviour of another.
  - In repeated-measures designs (in which participants are measured in more than one experimental condition), we expect scores in the experimental conditions to be non-independent for a given participant, but behaviour between different participants should be independent.

Correcting problems in the data

- Dealing with outliers:
  - 1 *Remove the case*

    2 *Transform the data*:

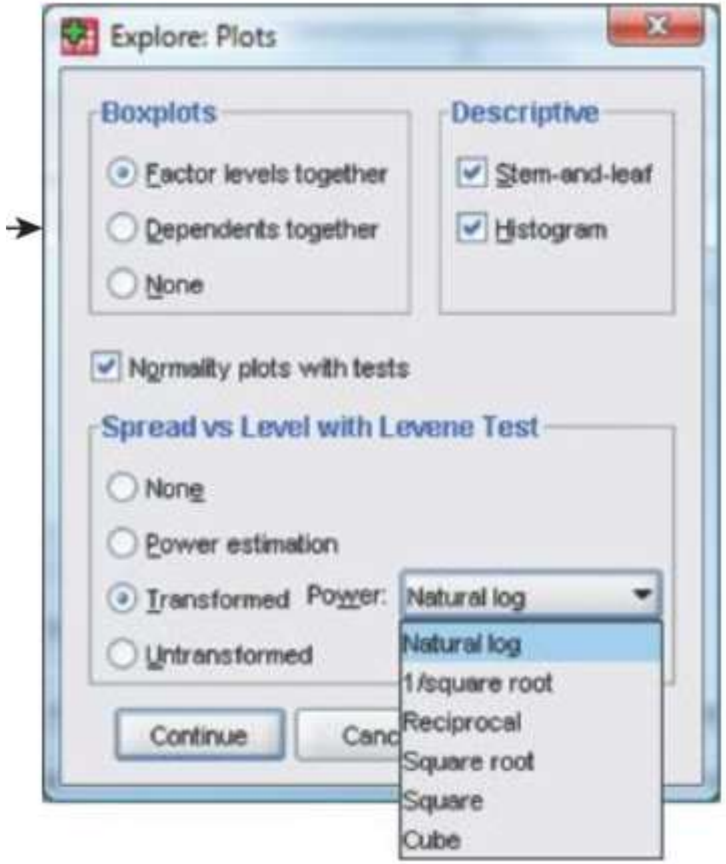
    3 *Change the score*:

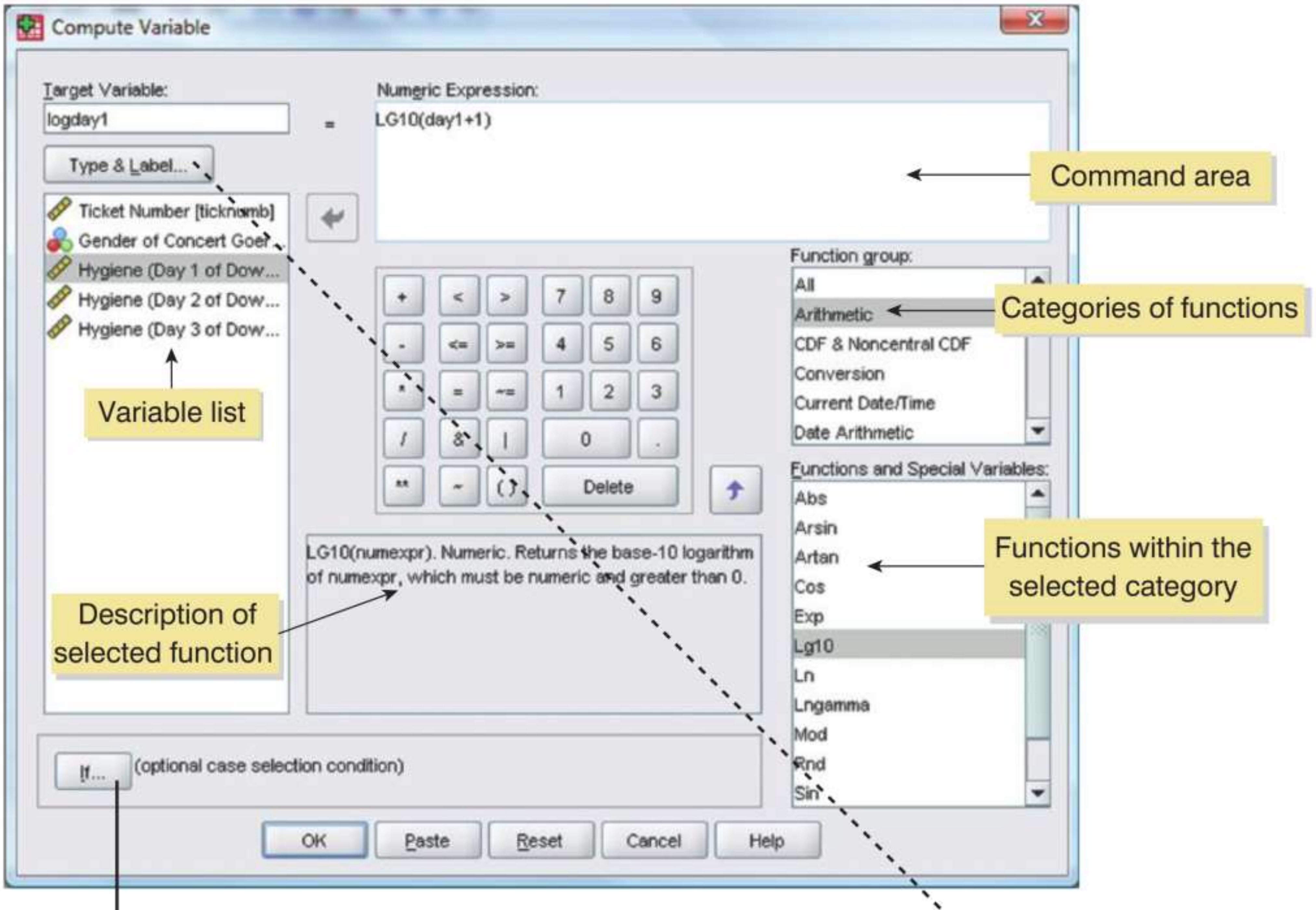
    (a) *The next highest score plus one*

    (b) *Convert back from a z-score*

- Dealing with non-normality and unequal variances
  - **Choosing a transformation**

    trial and error




Compute Variable
Target Variable:
logday1
Type & Label...
Ticket Number [ticknumb]
Gender of Concert Goer...
Hygiene (Day 1 of Dow...
Hygiene (Day 2 of Dow...
Hygiene (Day 3 of Dow...
Variable list
Numeric Expression:
LG10(day1+1)
Command area
Function group:
All
Arithmetic
CDF & Noncentral CDF
Conversion
Current Date/Time
Date Arithmetic
Categories of functions
Functions and Special Variables:
Abs
Arsin
Artan
Cos
Exp
Lg10
Ln
Lngamma
Mod
Rnd
Sin
Functions within the selected category
Delete
LG10(numexpr). Numeric. Returns the base-10 logarithm of numexpr, which must be numeric and greater than 0.
Description of selected function
If...
(optional case selection condition)
OK
Paste
Reset
Cancel
Help

How do we measure relationships?

- The correlation coefficient has to lie between -1 and +1.
- A coefficient of +1 indicates a perfect positive relationship, a coefficient of -1 indicates a perfect negative relationship, a
- coefficient of 0 indicates no linear relationship at all.
- The correlation coefficient is a commonly used measure of the size of an effect:
  - values of ±.1 represent a small effect,
  - ±.3 is a medium effect
  - ±.5 is a large effect.

- warning about interpretation: causality

Pearson’s correlation coefficient

- **assumptions of Pearson’s r**
  Pearson’s correlation requires only that data are interval
- if you want to establish whether the correlation coefficient is significant, then more assumptions are required:
  - sampling distribution has to be normally distributed: both variables to be normally distributed

Correlations

		Exam performance (%)	Exam Anxiety	Time spent revising
Exam performance (%)	Pearson Correlation	1.000	-.441**	.397**
	Sig. (1-tailed)	.	.000	.000
	N	103	103	103
Exam Anxiety	Pearson Correlation	-.441**	1.000	-.709**
	Sig. (1-tailed)	.000	.	.000
	N	103	103	103
Time spent revising	Pearson Correlation	.397**	-.709**	1.000
	Sig. (1-tailed)	.000	.000	.
	N	103	103	103

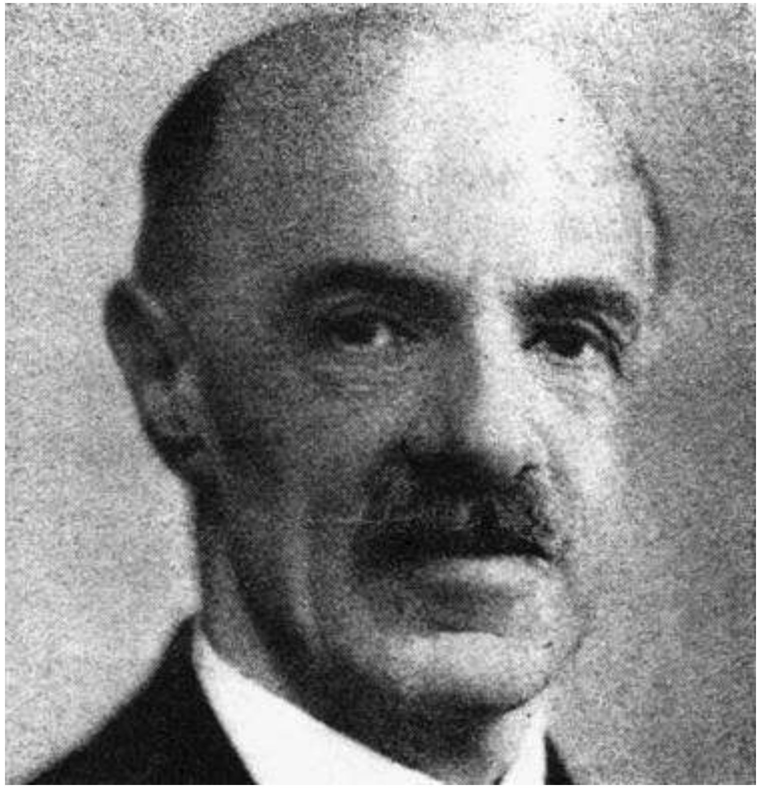
**. Correlation is significant at the 0.01 level (1-tailed).

using R^2 for interpretation

- The correlation coefficient squared (known as the **coefficient of determination**, R^2) is a measure of the amount of variability in one variable that is shared by the other.
- These two variables had a correlation of -0.4410 and so the value of R^2 will be $(-0.4410)^2 = 0.194$. This value tells us how much of the variability in exam performance is shared by exam anxiety.

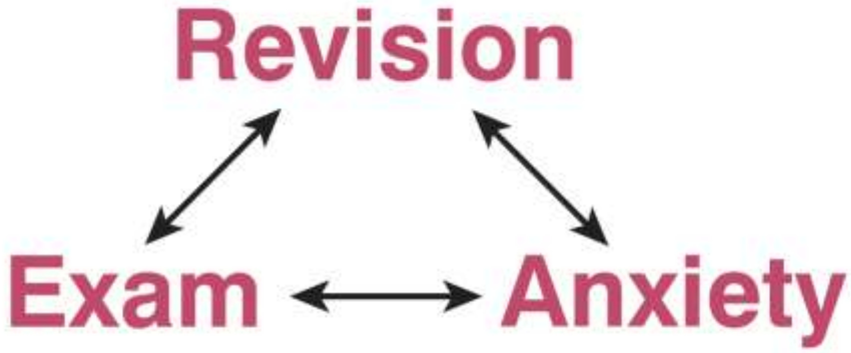
Spearman's correlation coefficient

- r_s, is a non-parametric statistic and so can be used when the data have violated parametric assumptions such as nonnormally distributed data
- Spearman's test works by first ranking the data



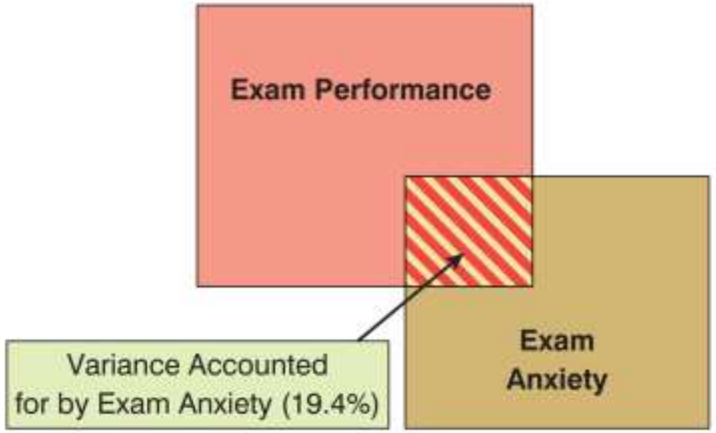
Partial correlation

- A correlation between two variables in which the effects of other variables are held constant is known as a partial correlation.

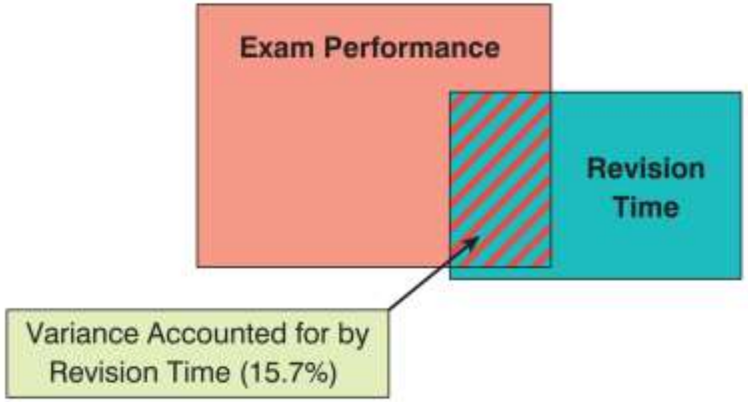


Partial Correlation

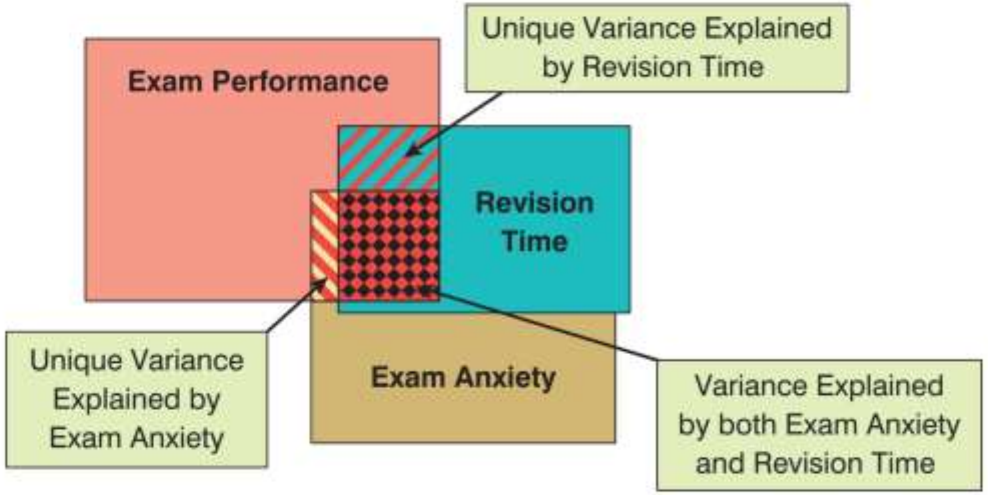
1


Exam Performance
Exam
Anxiety
Variance Accounted
for by Exam Anxiety (19.4%)

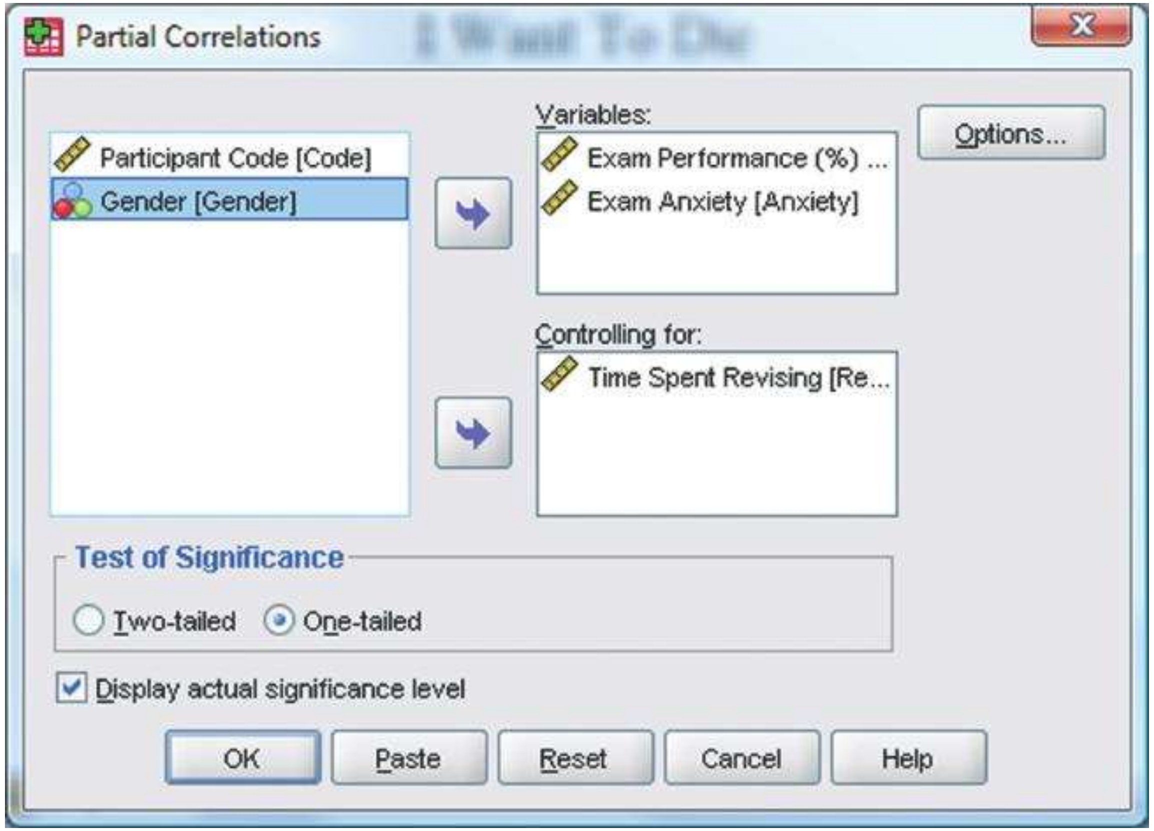
2

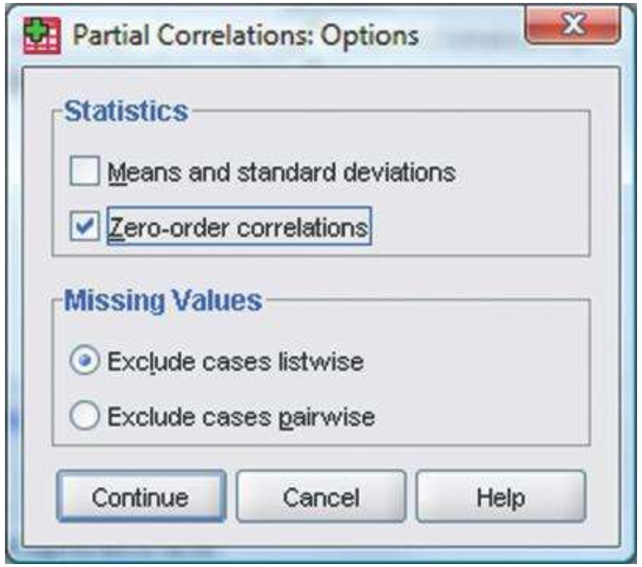

Exam Performance
Revision
Time
Variance Accounted for by
Revision Time (15.7%)

3


Unique Variance Explained
by Revision Time
Exam Performance
Revision
Time
Exam Anxiety
Unique Variance
Explained by
Exam Anxiety
Variance Explained
by both Exam Anxiety
and Revision Time

- Partial correlation between exam anxiety and exam performance while ‘controlling’ for the effect of revision time..





- Running this analysis has shown us that exam anxiety alone does explain some of the variation in exam scores, but there is a complex relationship between anxiety, revision and exam performance that might otherwise have been ignored.

Correlations

Control Variables			Exam Performance (%)	Exam Anxiety	Time Spent Revising
-none-[a]	Exam Performance (%)	Correlation	1.000	-.441	.397
		Significance (1-tailed)	.	.000	.000
		df	0	101	101
	Exam Anxiety	Correlation	-.441	1.000	-.709
		Significance (1-tailed)	.000	.	.000
		df	101	0	101
	Time Spent Revising	Correlation	.397	-.709	1.000
		Significance (1-tailed)	.000	.000	.
		df	101	101	0
Time Spent Revising	Exam Performance (%)	Correlation	1.000	-.247	
		Significance (1-tailed)	.	.006	
		df	0	100	
	Exam Anxiety	Correlation	-.247	1.000	
		Significance (1-tailed)	.006	.	
		df	100	0	

a. Cells contain zero-order (Pearson) correlations.

How to report correlation coefficents

- There was a significant relationship between the number of adverts watched and the number of packets of sweets purchased, $r = .87$, p (one-tailed) $< .05$.

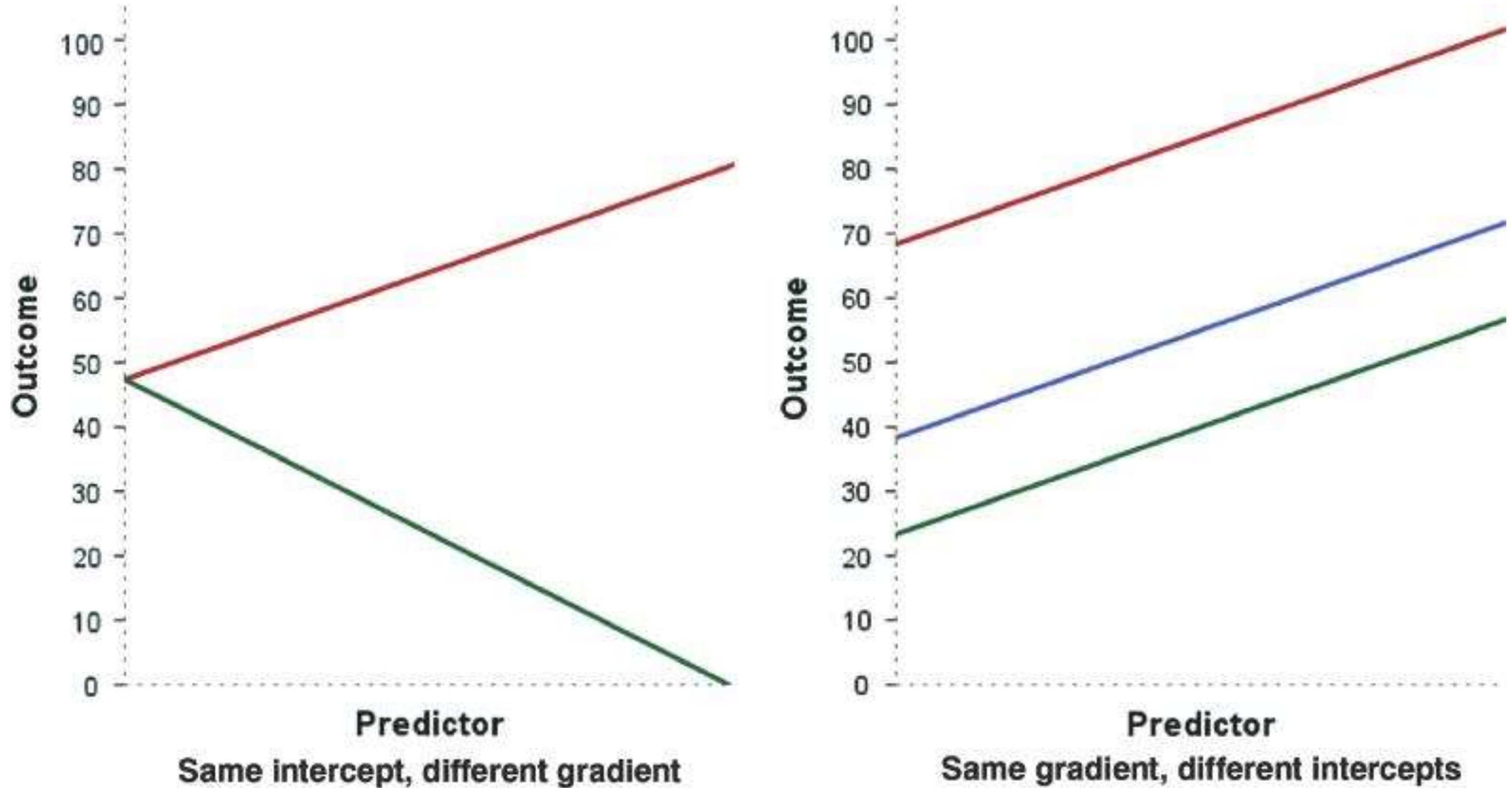
An introduction to regression

- we looked at how to measure relationships between two variables.
- Step further:
  - predict one variable from another.
  - predict levels of stress from the amount of time until you have to give a talk.

- Fit a model to our data and use it to predict values of the dependent variable from one or more independent variables.

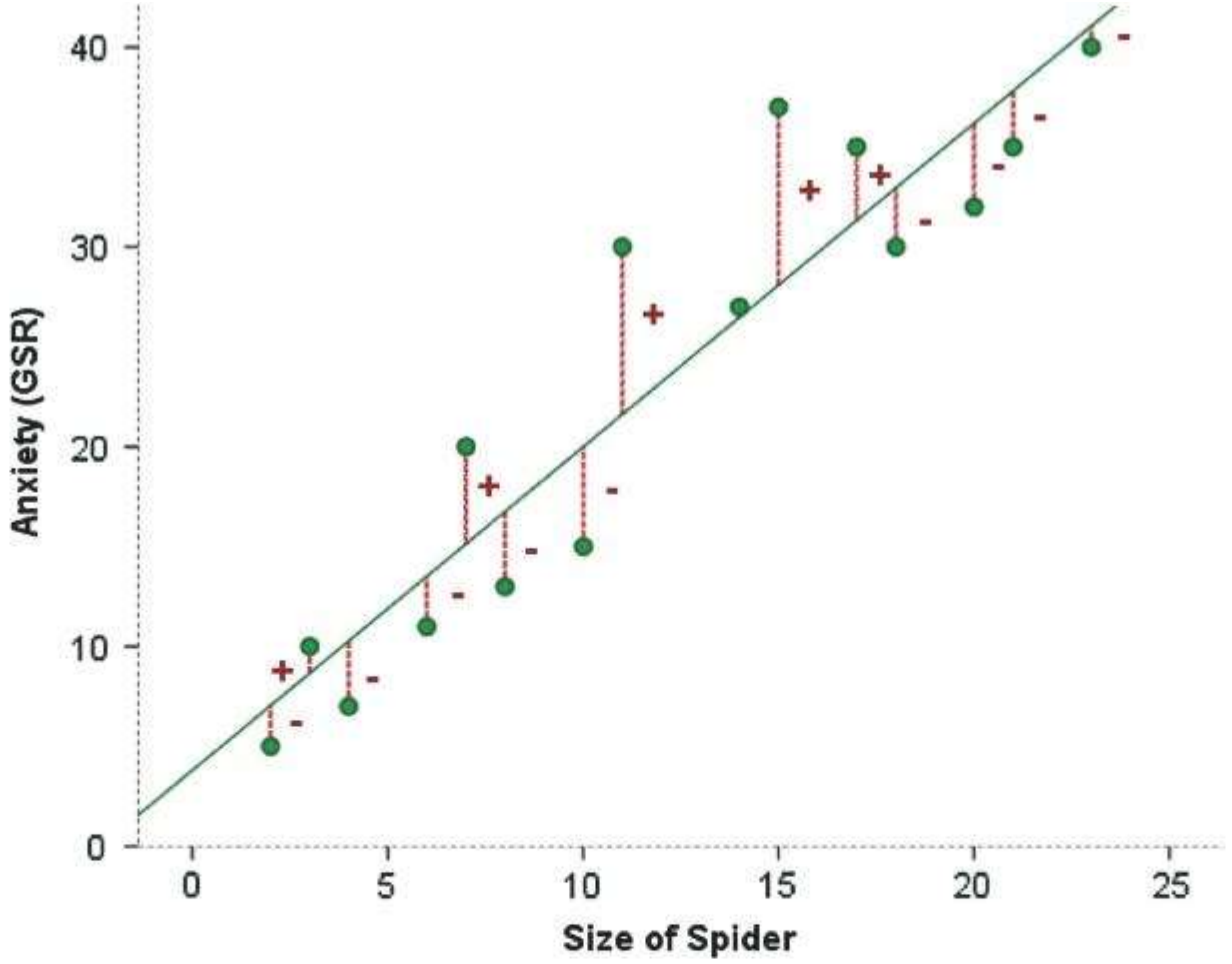
- Predicting an **outcome** from one **predictor** (**simple regression**) or several predictor (**multiple regression**).

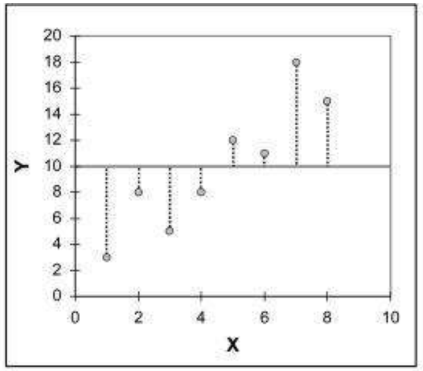
$outcome_i = (model) + error_i$

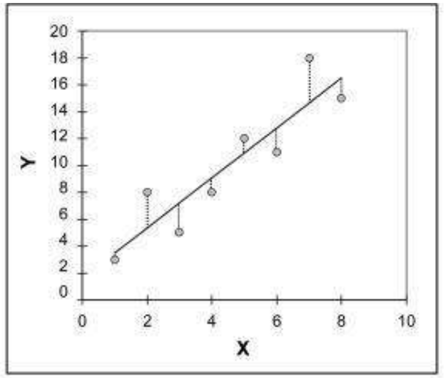


$$Y_i = (b_0 + b_1 X_i) + \varepsilon_i$$

Assessing the goodness of fit: sums of squares, R and R2

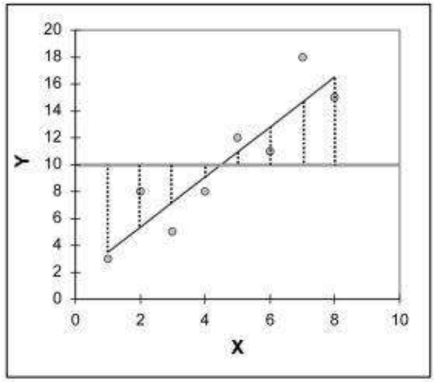




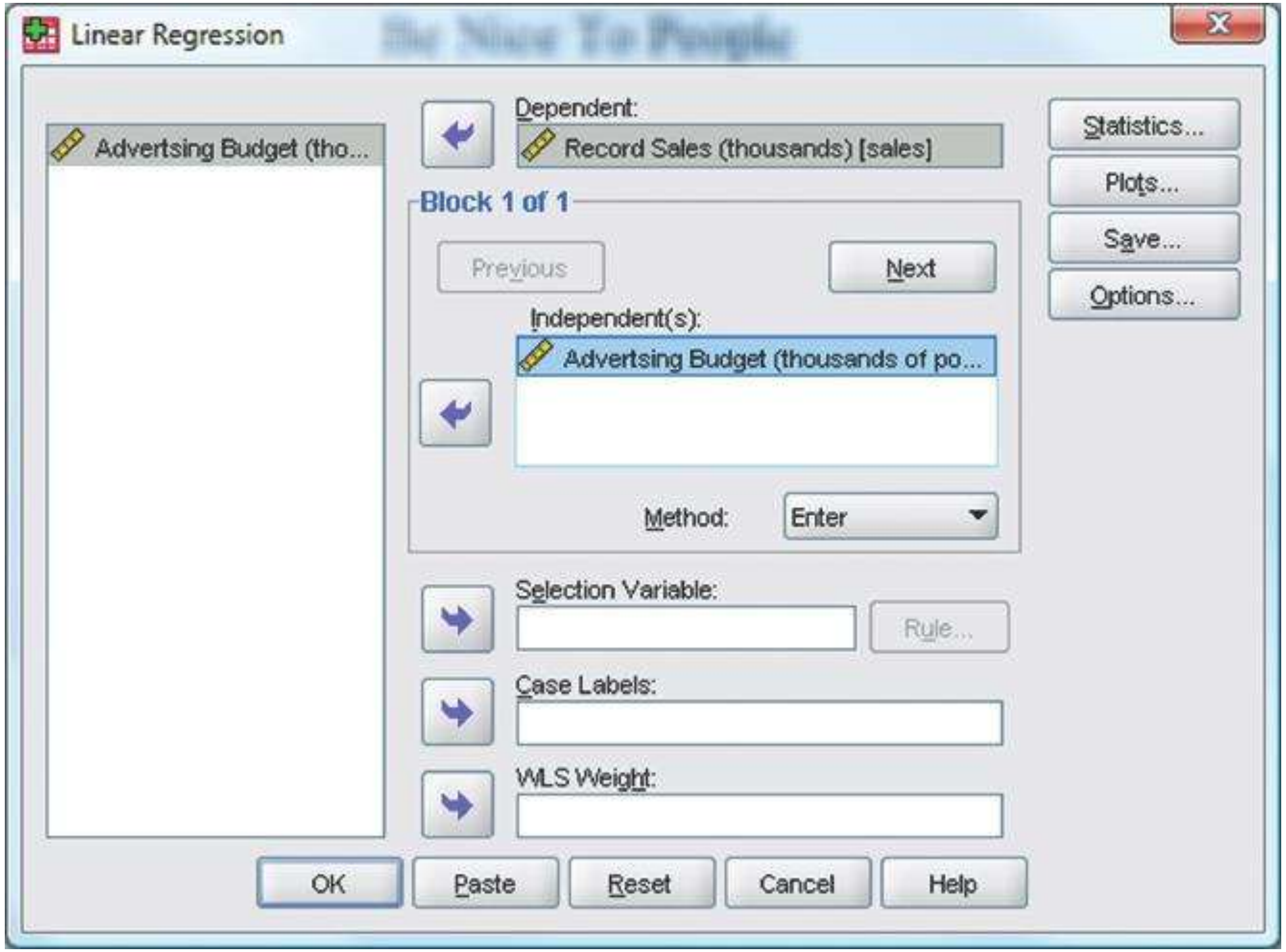


SS_T uses the differences between the observed data and the mean value of *Y*

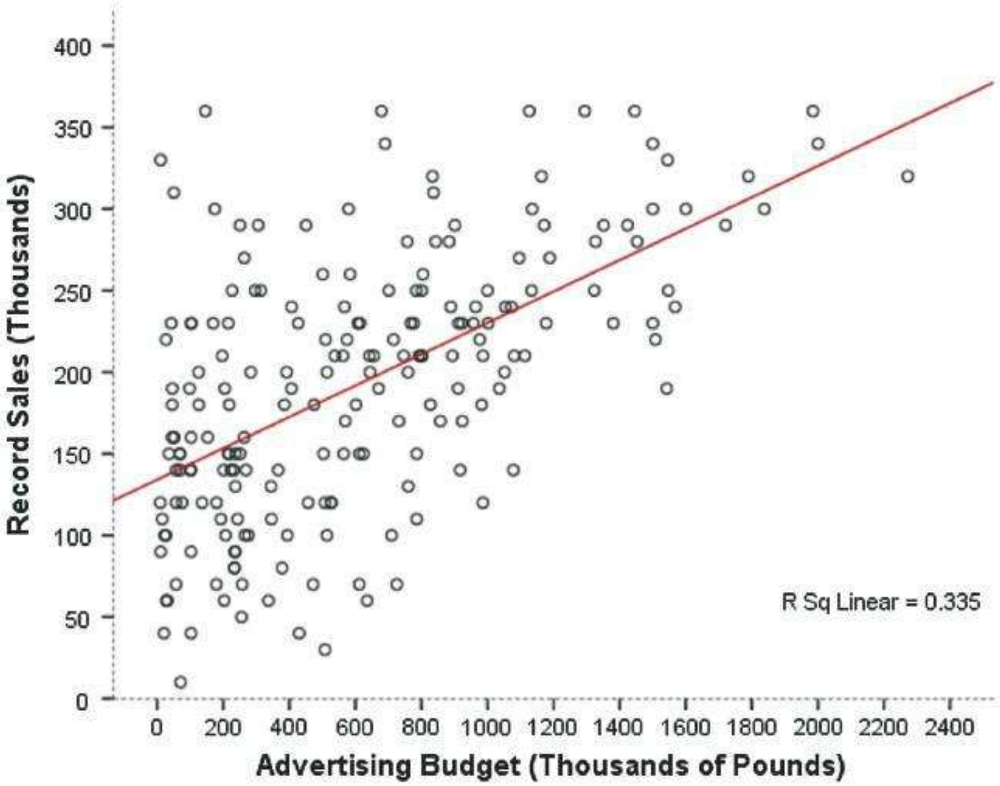
SS_R uses the differences between the observed data and the regression line



SS_M uses the differences between the mean value of *Y* and the regression line



Linear Regression
Advertsing Budget (tho...
Dependent:
Record Sales (thousands) [sales]
Block 1 of 1
Previous
Next
Independent(s):
Advertsing Budget (thousands of po...
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help



Record Sales (Thousands)
400
350
300
250
200
150
100
50
0
0 200 400 600 800 1000 1200 1400 1600 1800 2000 2200 2400
Advertising Budget (Thousands of Pounds)
R Sq Linear = 0.335

- R: because there is only one predictor, this value represents the simple correlation between advertising and record sales
- R2: advertising expenditure can account for 33.5% of the variation in record sales

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.578[a]	.335	.331	65.991

a. Predictors: (Constant), Advertsing Budget (thousands of pounds)

- The ANOVA tells us whether the model, overall, results in a significantly good degree of prediction of the outcome variable.
- Our regression model results in significantly better prediction of record sales than if we used the mean value of record sales. In short, the regression model overall predicts record sales significantly well

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	433687.833	1	433687.833	99.587	.000[a]
	Residual	862264.167	198	4354.870		
	Total	1295952.000	199			

a. Predictors: (Constant), Advertising Budget (thousands of pounds)

b. Dependent Variable: Record Sales (thousands)

SS_M → 433687.833; SS_R → 862264.167; SS_T → 1295952.000; MS_M → 433687.833; MS_R → 4354.870

How do I interpret *b* values?

- b0 was the Y intercept and this value is the value B for the constant
- meaning that when no money is spent on advertising (when $X = 0$), the model predicts that 134,140 records will be sold.
- *b*1 from the table and this value represents the gradient of the regression line.
- *the change in the outcome associated with a unit change in the predictor*.
- *t*-test tells us whether the *b*-value is different from 0.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	134.140	7.537		17.799	.000
	Advertsing Budget (thousands of pounds)	.096	.010	.578	9.979	.000

a. Dependent Variable: Record Sales (thousands)

Using the model

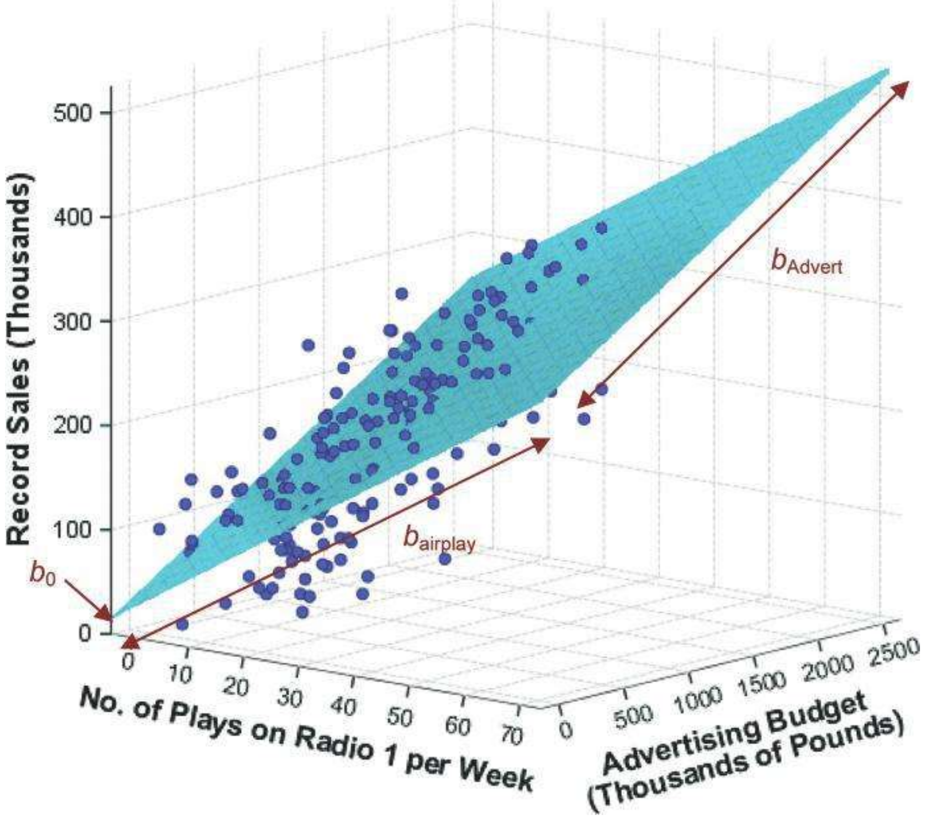
$$\text{record sales}_i = b_0 + b_1 \text{advertising budget}_i$$
$$= 134.14 + (0.096 \times \text{advertising budget}_i)$$

Multiple regression: the basics

- Multiple regression is a logical extension of these principles to situations in which there are several predictors.
- $$Y_i = (b_0 + b_1 X_{i1} + b_2 X_{i2} + \ldots + b_n X_n) + \varepsilon_i$$

An example of a multiple regression model

$$\text{record sales}_i = b_0 + b_1 \text{advertising budget}_i + b_2 airplay_i + \varepsilon_i$$



Comparing two means ..

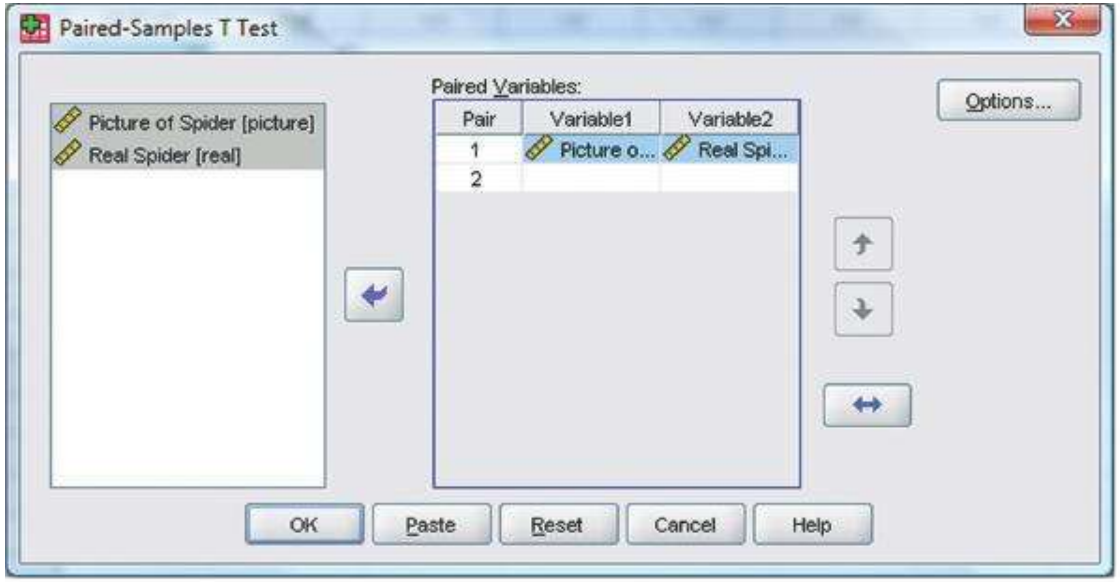
Assumptions of the t-test

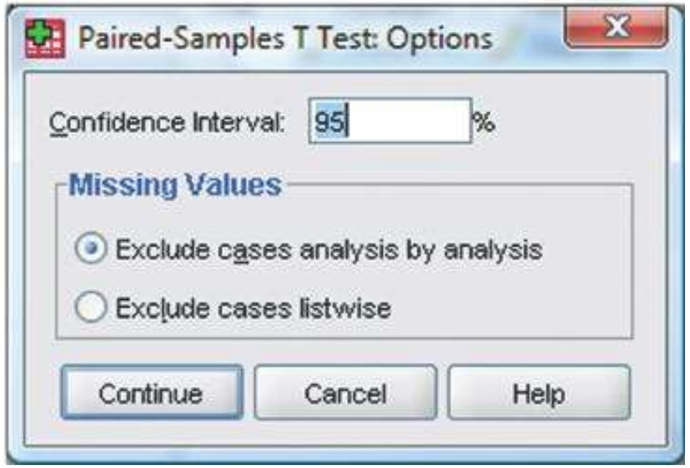
- The sampling distribution is normally distributed. In the dependent *t*-test this means that the sampling distribution of the *differences* between scores should be normal, not the scores themselves

- Data are measured at least at the interval level.

- The independent *t*-test, because it is used to test different groups of people, also assumes:
  - Variances in these populations are roughly equal (*homogeneity of variance*).
  - Scores are independent (because they come from different people)

dependent t-test





Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Picture of Spider	40.00	12	9.293	2.683
	Real Spider	47.00	12	11.029	3.184

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Picture of Spider & Real Spider	12	.545	.067

Paired Samples Test

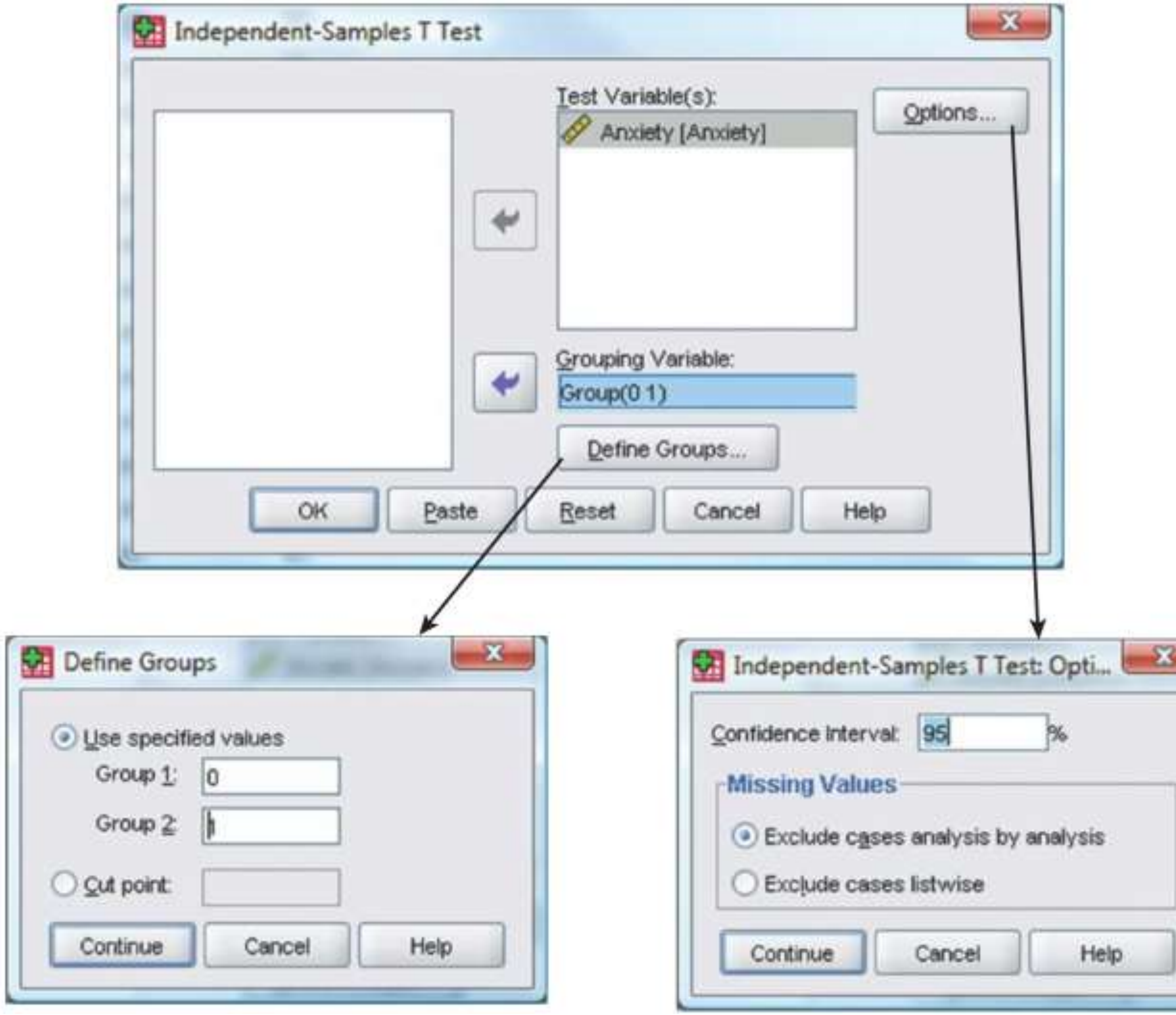
		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Picture of Spider - Real Spider	-7.000	9.807	2.831	-13.231	-.769	-2.473	11	.031

On average, participants experienced significantly greater anxiety to real spiders (*M* = 47.00, *SE* = 3.18) than to pictures of spiders (*M* = 40.00, *SE* = 2.68), $t(11) = -2.47$, $p < .05$

The independent t-test

Group Statistics

	Spider or Picture?	N	Mean	Std. Deviation	Std. Error Mean
Anxiety	Picture	12	40.00	9.293	2.683
	Real Spider	12	47.00	11.029	3.184



Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Anxiety	Equal variances assumed	.782	.386	-1.681	22	.107	-7.000	4.163	-15.634	1.634
	Equal variances not assumed			-1.681	21.385	.107	-7.000	4.163	-15.649	1.649

On average, participants experienced greater anxiety to real spiders (*M* = 47.00, *SE* = 3.18) than to pictures of spiders (*M* = 40.00, *SE* = 2.68). This difference was not significant $t(22) = -1.68$, $p > .05$;