Basic Oncological Sciences
Statistics and Clinical Epidemiology
Medical Statistics

Comparing two or more groups and using measures and tests of association

Dr Margaret MacDougall
Medical Statistician
Public Health Sciences
University of Edinburgh Medical School
Teviot Place
Edinburgh EH8 9AG

Methods covered

* Identifying the type of data you have

* Choosing the right hypothesis test and checking for Normality

* Chi-squared tests for two groups and related corrections

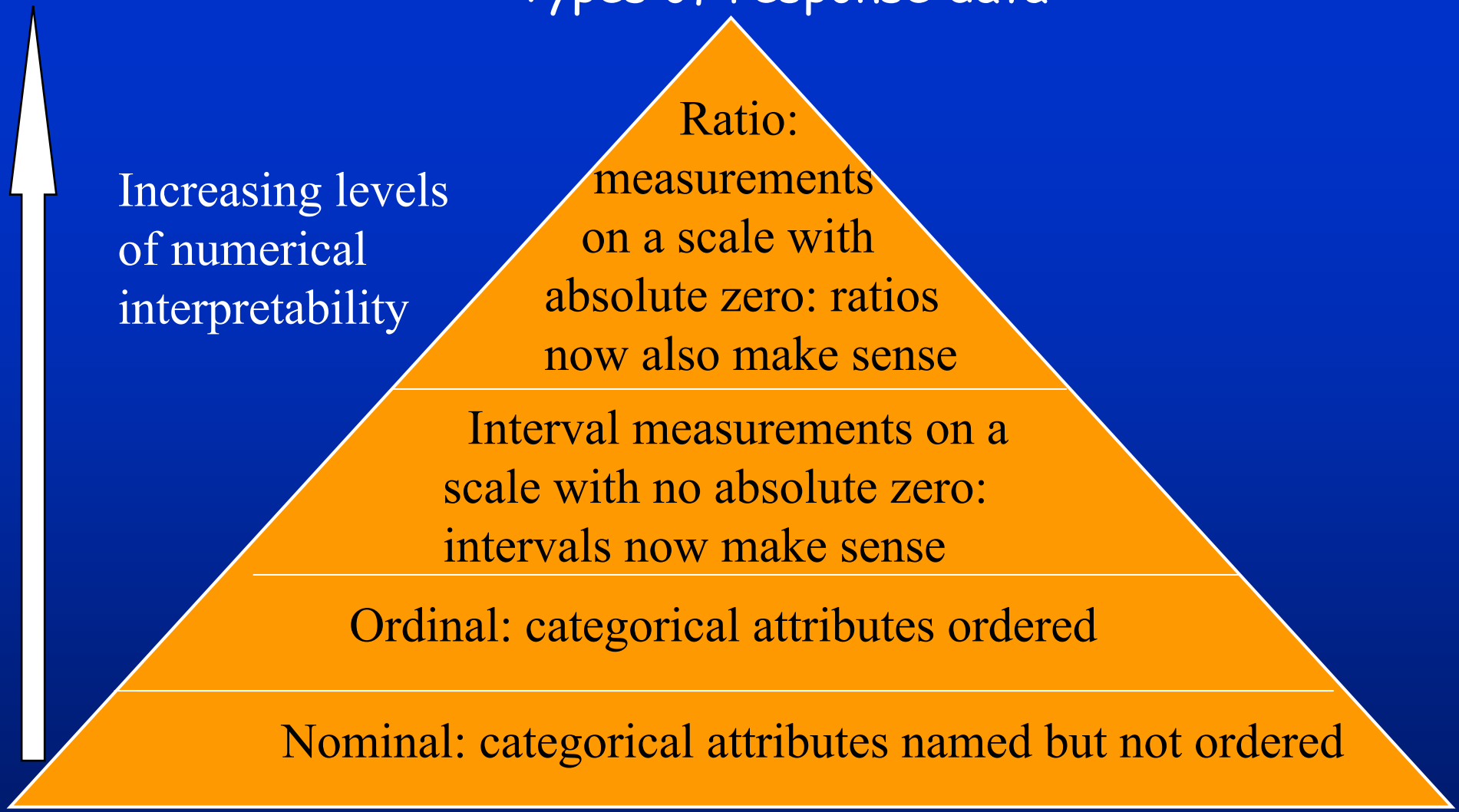
* Correlation coefficients and simple linear regression

* T-tests for two groups

* Sensitivity, specificity and positive and negative predictive values

* Past examination questions

Types of response data



Examples of types of response data

Type of data	Examples
Nominal	Presence ('Yes') or absence ('No') of post-op complications; tumour morphology after resection: 'diffuse', 'nodular', 'scattered', 'central scar', 'mixed', 'complete response'
Interval (less common)	Atmospheric temperature (in Fahrenheit): there is no absolute zero in the scale and therefore it does not makes sense to say that an atmospheric temperature of 38°F at one time point is twice as high as that of 19°F at another time point. (The arithmetic difference 38°F- 19°F does, however, make sense as a difference of 19°F.)

Examples of types of response data (continued)

Type of data	Examples
Ordinal	Cancer stage on basis of BCLC* staging system (Stage A (early stage HCC) to Stage D (end stage HCC)); assessment of palliative care provided for relative (rating on scale of 1 to 5 (where '1' denotes 'poor' and '5' denotes 'excellent')
Ratio	Change in Hb (g/dl) from pre- to post-op stages (e.g. -6.6, -1.9, …4.4,…) Parsonnet score (scale from 0 (no risk of mortality) to > 19 (very high risk of mortality); can also be categorized: low risk (0-4), elevated risk (5-9), significantly elevated risk (10-14), high risk (15-19), very high risk (> 19) [resultant data is ordinal]

* BCLC: Barcelona Clinic Liver Cancer

General Procedure for Test

Step 1: Define Null Hypothesis (H_0)

Step 2: Identify appropriate hypothesis test

Step 3: Choose significance level (α): 0.05 (5%) or 0.01 (1%) [a significance level of 0.01 may be used as a cut-off point where it is intended that H_0 be rejected only where a highly significant result is obtained]

Step 4: Calculate relevant **Test Statistic** for chosen test

Step 5: Use look-up tables (see Appendix A of handout) to decide whether test statistic lies within **rejection region for H_0**

Step 6:
apply decision rule

if the p-value is less than the significance level ($p \leq 0.05$, say), this result is statistically significant: reject H_0; otherwise, do not reject H_0.

Tips when writing up

On using a statistical package to obtain a p-value,

always quote the p-value within the conclusion

The significance level, α acts as a useful cut-off point for the p-value in deciding whether or not to reject H_0 but in so doing, introduces some level of subjectivity to the decision rule in that for example, the very similar p-values 0.044 and 0.051 would lead to different decisions for α=0.05. It is therefore best to quote the p-value together with the decision in order that the reader can decide how convincing the conclusion is.

Always mention the sample size

Especially for small sample sizes, increasing the sample size could lead to enough of a decrease in the p-value to change the overall conclusion for such borderline cases.

Ways of stating the possible conclusions for a hypothesis test with $\alpha = 0.05$

Case 1: $p > 0.05$; suppose sample size is 35 and p=0.06 for illustration

For a sample size of 35,there is insufficient evidence at the 5% significance level to reject H_0 (p=0.06). Hence it cannot be concluded that H_A is probably true.

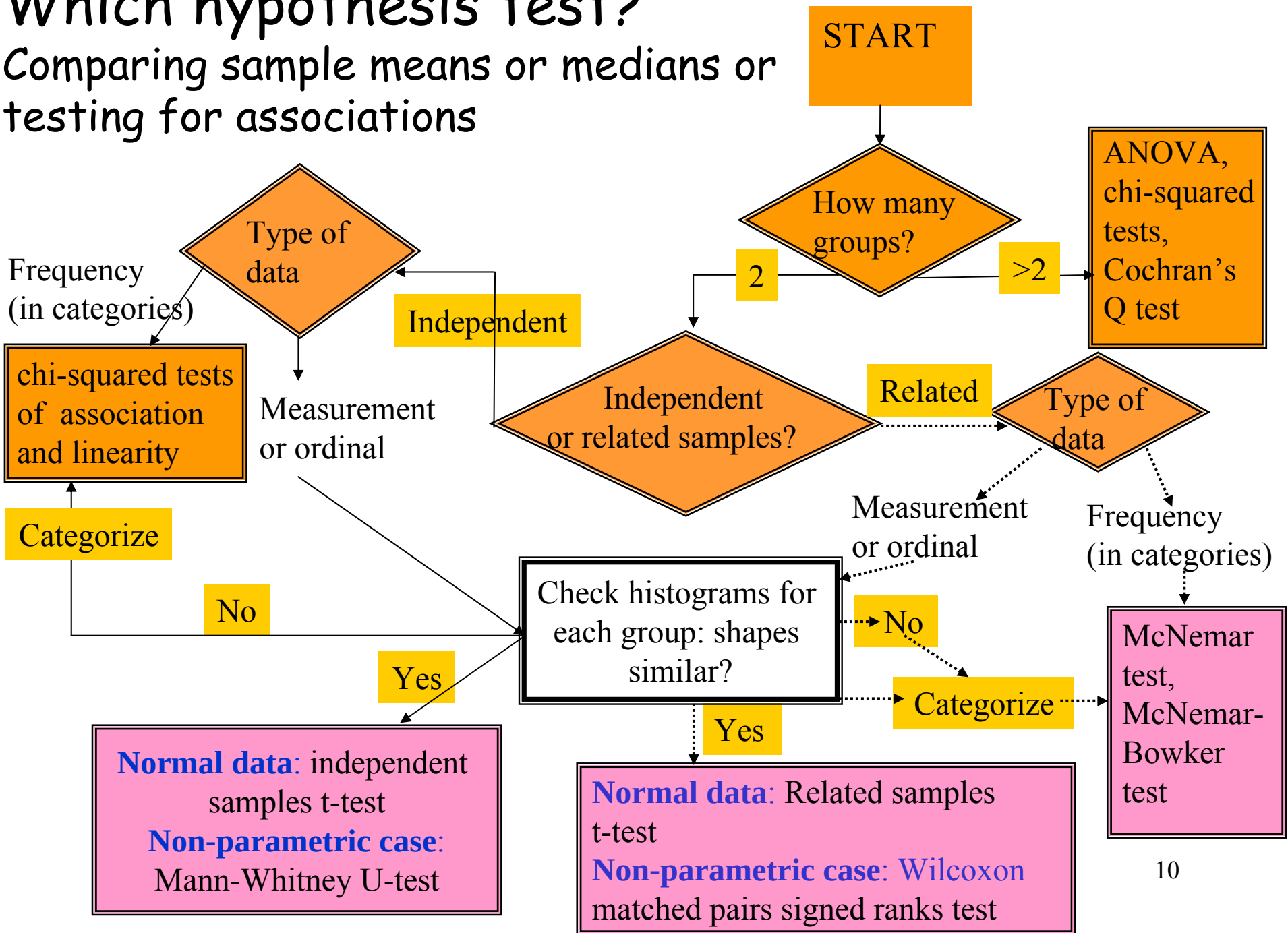
Case 2: $p \leq 0.05$; suppose sample size is 35 and p=0.04 for illustration

For a sample size of 35, there is sufficient evidence at the 5% significance level to reject H_0 (p=0.04). Hence H_A is probably true.

NB. In such cases, specify H_0 and H_A in terms of your problem!

Which hypothesis test?

Comparing sample means or medians or testing for associations

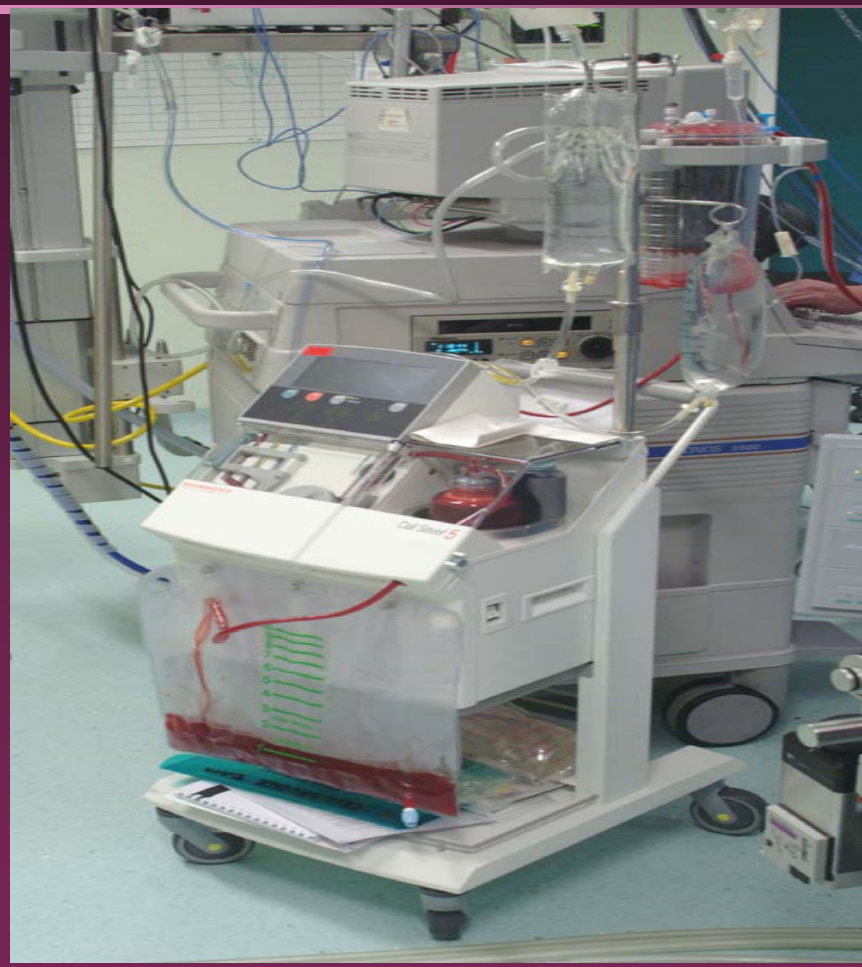


Cell Saver

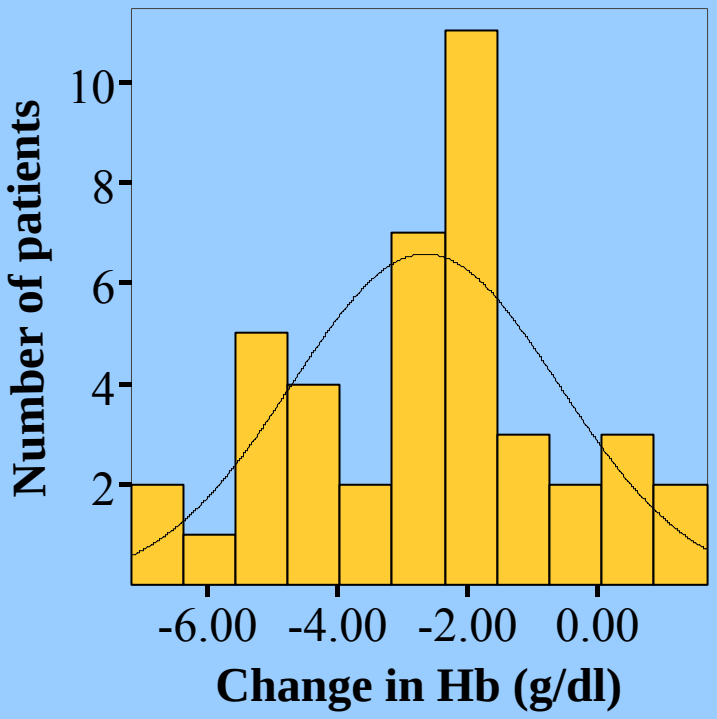
- invented 30 years ago
- salvaged blood collected in sterile reservoir
- processed in centrifuge bowl
- returned to patient (haematocrit about 60%)

Contra-indications

- malignancy (? irradiation)
- contamination

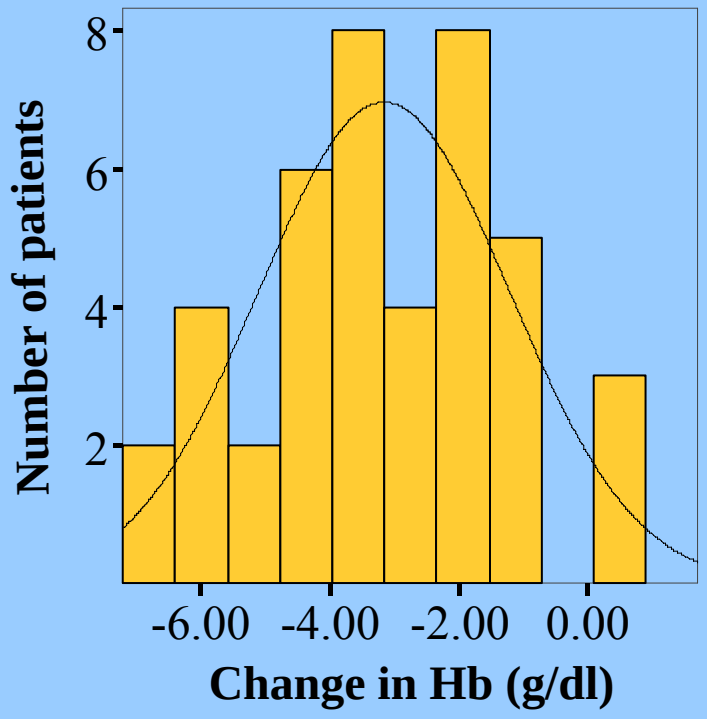


Approximations to Normality in distributions representing change in Hb levels from pre- to post-op stages for a) cell salvage and b) non-cell salvage group



a) cell salvage

Shapiro-Wilks p-value: 0.573

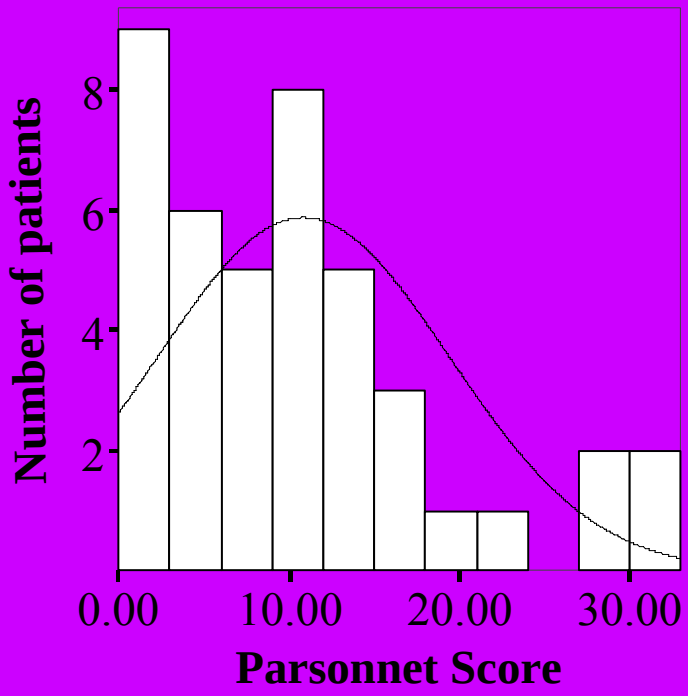


b) no cell salvage

Shapiro Wilks p-value: 0.836

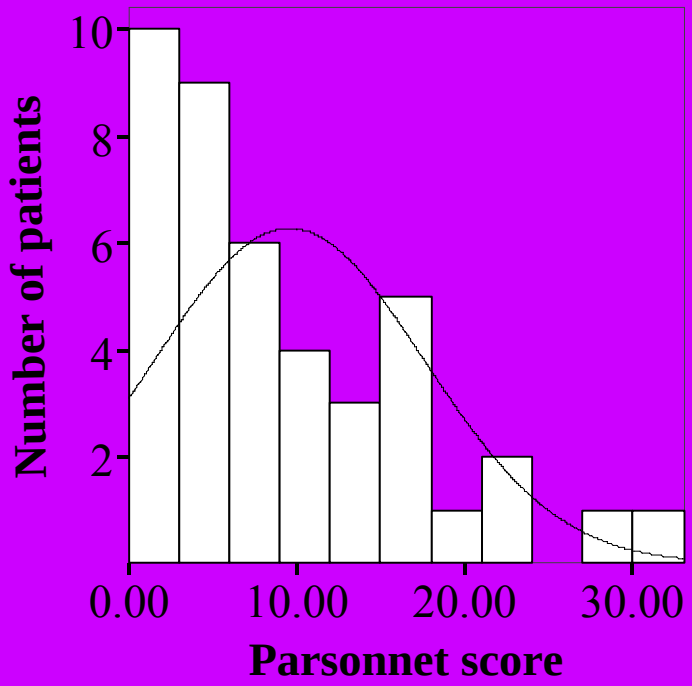
Candidates for independent samples t-test

Deviations from Normality in distributions representing Parsonnet scores for a) cell salvage group and b) non-cell salvage group



a) cell salvage

Shapiro-Wilks p-value: 0.002(1518…)



b) no cell salvage

Shapiro-Wilks p-value: 0.002(1818)

Candidates for Mann-Whitney U-test

Two-sample t-tests

Case 1: the two groups to be compared are related in some way – examples:

· each individual from a single group has pulse wave velocity measured using pressure sensor and light sensor: single group perceived as two related groups – one for each sensor

· each individual from a single group has pulse wave velocity measured before and during exercise: single group perceived as two related groups – one for each time point

· two groups – mothers and daughters; wish to compare weights in both groups according to parental relationship

Procedure: check flowchart (if, appropriate, use paired samples t-test)

e.g. Weights

mother	daughter	diff
70	40	30
100	105	-5
.	.	.
.	.	.

mean(diff),
standard deviation(diff)

obtain a column of differences and calculate the mean, $mean_{diff}$, and standard deviation, sd_{diff}, for these differences

H_0: the mean of the differences between the population groups is zero

H_A (for a two-tailed test): the mean of the differences between the population groups is not zero

Test statistic: $t = mean_{diff} / SE(mean_{diff})$,

where for an entire sample size of n, $SE(mean_{diff})$ is the standard error of $mean_{diff}$ and

$$SE(mean_{diff}) = sd_{diff} / \sqrt{n}$$

Result: t follows the t-distribution for df=n-1

Example (fasting plasma C-lysine flux before and after treatment with hydrazine sulphate

See handout: '**Worked example on the paired t-test (repeated measures)**'

Case 2: the two groups to be compared are unrelated

Procedure: check flowchart (if, appropriate, use independent samples t-test)

Example: comparison of lengths of hospital stays, where group 1 has seen consultant early on and group 2 has not

Null and alternative hypotheses: similar to those for paired t-test

Notation:

calculate these separately (c.f. Case 1)

$mean_1$, $mean_2$: sample means for groups 1 and 2, resp.

n_1, n_2: sample sizes for groups 1 and 2, resp.

Test statistic

$t=(mean_1-mean_2)/SE(mean_1-mean_2)$

where $SE(mean_1-mean_2)$ is the standard error of the difference between the mean for group 1 and the mean for group 2 – see Petrie and Sabin (topic 21) for details and related example (compares peak flow readings for treatment and placebo groups) .

t follows t-distribution for n_1+n_2-2 degrees of freedom.

Chi-squared tests for relationship between two groups

You will need …

* 2 categorized groups (e.g. age and presence or absence of chest pain) each with at least two categories
* a Null hypothesis: e.g. there is no association between age and presence or absence of pain
* a significance level: we shall choose 5%
* a contingency table (which cross-tabulates frequencies for the two groups) - I will provide a general one here:

rows might represent cancer types

columns might represent age categories

Group2 (e.g. cancer)	Group1 (e.g. age) col1	col2	col3	…… colc	Marginal row totals
row1	f_{11}	f_{12}	f_{13}	…… f_{1c}	R_1
row2	f_{21}	f_{22}	f_{23}	…… f_{2c}	R_2
row3	f_{31}	f_{32}	f_{33}	…… f_{3c}	R_3
…	……				…
…	……				…
rowr	f_{r1}	f_{r2}	f_{r3}	…… f_{rc}	R_r
Marginal column totals	C_1	C_2	C_3	… C_c	Total (sample size: n)

Entries in above contingency table are said to be contained in **cells** of the table. These entries are the observed frequencies (O) of individuals falling under the cross-tabulated categories for the two variables (e.g. age and cancer) considered.

Expected frequencies (E): the frequencies that would be obtained if there was no association between the two groups (and hence H_0 was true). These are obtained as follows:

Expected frequency for observed frequency f_{ab} $=(R_a \times C_b)/n$

For the chi-squared test of association, the test statistic may be understood as a measure of the discrepancy between the observed and expected frequencies:

Test statistic (version 1):
$\chi^2 = \Sigma\ (O-E)^2/E$,

where, for any cell of the contingency table, O and E are the observed and expected frequencies, respectively.

Yates' correction: for the 2x2 case (i.e. when there are two categories in each group), a continuity correction which is a applied to the formula for the test statistic to allow for the fact that O and E are discrete variables but χ^2 is to be associated with a continuous distribution

Test statistic (version 2) with Yates' correction:
$\chi^2=\Sigma\ (|O-E|-0.5)^2/E,$

where |.| is the absolute value function

Degrees of freedom and when to use Yates' correction

Abbreviation
df: degrees of freedom

The number of row (r) and columns (c) in the contingency table determines the particular chi-squared distribution we should use when consulting look-up tables.

Where there are r categories in one group and c in the other, we assume that the test statistic follows the chi-squared distribution for the case where df= (r-1)x(c-1) degrees of freedom.

N.B. when df > 1, version 1 of the test statistic follows the chi-squared distribution sufficiently well for us not to need Yates' correction.

The need for Fisher's Exact test

Before applying the chi-squared test for association, carry out check of frequencies

Yes

No

chi-squared test for association

Fisher's exact test: provides more exact p-value

[Calculations tedious by hand, so best performed by a statistical Package]

[Note: if we are seeking a linear trend between the categories of a group and the occurrence of an event (as could be the case in seeking an association between stage of cancer (under a staging system) and mortality, we should consider the chi-squared test for a linear trend. See, for example:

Topic 25 of Petrie, A. and Sabin, C., 2000. Medical Statistics at a Glance. Blackwell Science Ltd, Oxford

or

http://www.uvm.edu/~dhowell/StatPages/More_Stuff/OrdinalChisq/OrdinalChiSq.ht

General point about chi-squared test for association for 2x2 case:

may be appropriate to use when comparing the proportions of those individuals within two groups which have a particular condition [think this through in exam – see handouts]: want to see if group proportions significantly different

Wound infection?	Use of cell salvage procedure: cell salvage	no cell salvage
yes	9	6
no	33	36

Worked example on the chi-squared test of association (2 x 2 case)

Q. The table below presents the observed frequencies for absence and presence of wound infections according as to whether or not the cell salvage procedure was used during operation. By choosing a suitable hypothesis test, decide whether the proportions of individuals for which wound infection occurred differed significantly according as to whether or not the cell salvage procedure was used and therefore whether it is likely that there is an association between the variables 'cell salvage' and 'wound infection'.

Cell salvage	Wound infection	
	Present	Absent
Yes	9	33
No	6	36

Solution. Use chi-squared test of association

Expected frequencies are in square brackets in table below.

Cell salvage	Wound infection		
	Present	Absent	Total
Yes	9 [7.5]	33 [34.5]	42
No	6 [7.5]	36 [34.5]	42
Total	15	69	84

- Null hypothesis: there is no association between 'cell salvage' and 'wound infection' [two-tailed test].

- Equivalently, there is no difference between the proportions of individuals with wound infections across the two cell salvage categories.

Decide on significance level (α) of 0.05

Check degrees of freedom and calculate test statistic:

$$df = (r-1)(c-1) = (2-1)(2-1) = 1$$

Hence use test statistic with Yates' correction:

$$\chi^2 = \Sigma\ (|O-E|-0.5)^2/E$$

$$= (|9-7.5|-0.5)^2/7.5 + (|33-34.5|-0.5)^2/34.5$$
$$+ (|6-7.5|-0.5)^2/7.5 + (|36-34.5|-0.5)^2/34.5$$

$$= (2/7.5 + 2/34.5)$$

$$= 0.325$$

Using look-up table for 1 df (see Appendix A), critical value is 3.84. Hence the test-statistic does not lie in the rejection region for the null hypothesis. (It is not at least as extreme as the critical value.) Therefore, $p > 0.05$.

Conclusion.
For a sample size of 84, there is insufficient evidence at the 5% significance level to reject the null hypothesis, as there is not a significant difference in the proportions of individuals who had wound infections across the two cell salvage groups (21.4% for the cell salvage group and 14.3% for the non-cell salvage group).

Hence it cannot be concluded (from a population perspective) that there is probably an association between 'cell salvage' and 'wound infection'.

Example for you to try: see the handout 'Exercise on the chi-squared test of association (2 x 2 case)'

Simple linear Regression

Conditions

. We wish to explore the linear relationship between two continuous variables.

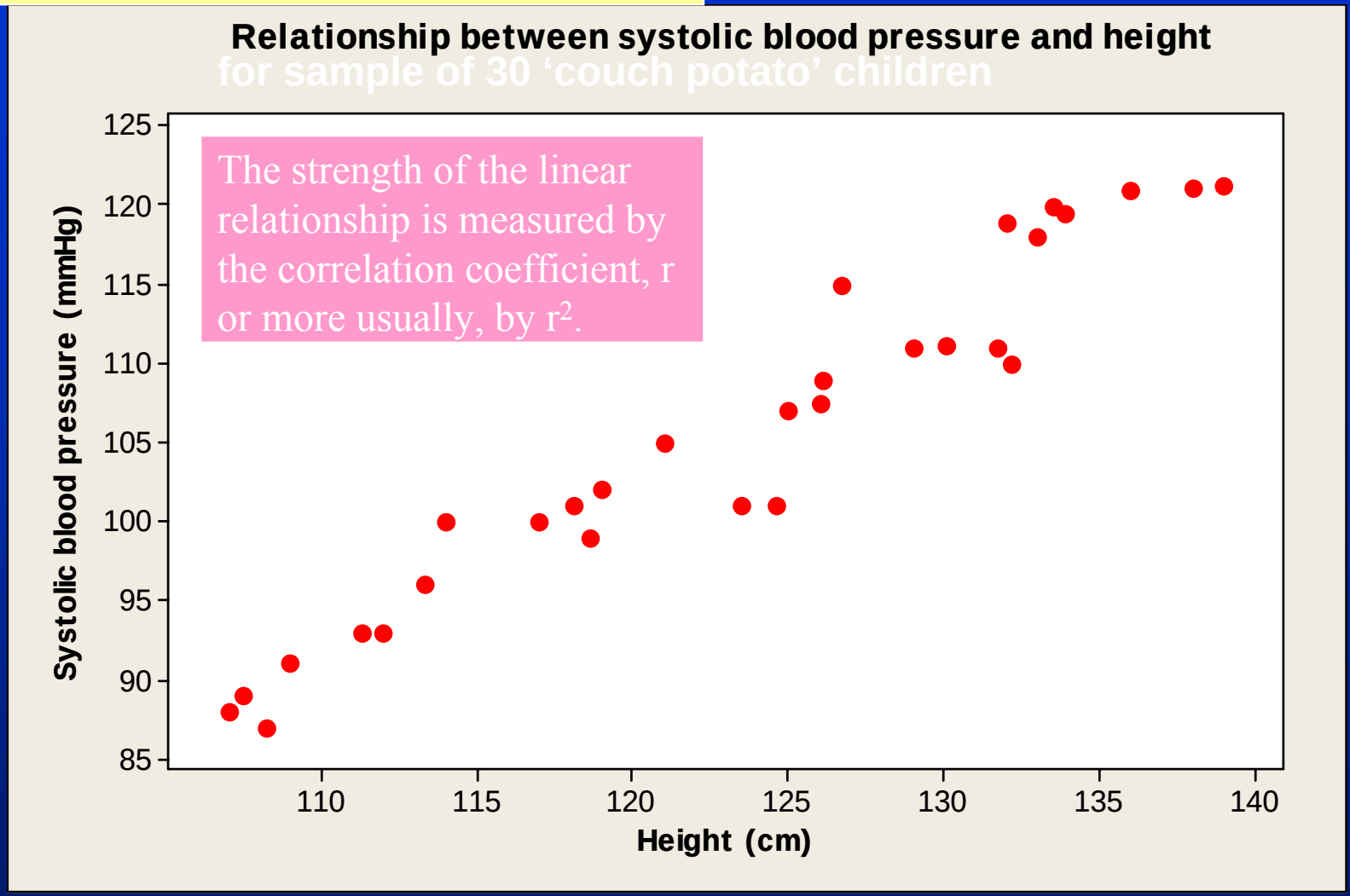
. A scatter plot of one variable against the other reveals a linear trend.

. There are no outliers in the data to be used in the regression analysis.

. One of the two variables (the dependent variable) is understood to be statistically dependent on the remaining variable (the independent variable), which does not guarantee a causal relationship, clinically.

. The dependent variable is Normally distributed.

How do we carry out simple linear regression?



We choose to approximate the data by means of a linear model.
In linear regression, if we have a single independent variable, we choose the line of best fit, with equation of the form

$$Y=a+bX$$

to represent the linear model.

Some Properties of the correlation coefficient, r [otherwise know as the Pearson (product moment) correlation coefficient]

- r is unitless
- r is a measure of how close the sample points are to the line of best fit for these points
- r is an estimate of the strength of linearity in the population data
- r lies between -1 and 1, inclusive
- if r is strictly negative (strictly positive) then the data exhibits a decreasing (increasing) trend

. if r = -1 or r = 1 (both cases unusual) there is a perfect correlation and perfect linearity

. if r = 0 then there is no linear relationship between the y- and corresponding x-coordinates for the sample data

. the value of r is only valid within the coordinate value ranges represented by the sample

More generally, correlation coefficients measure the strength of a particular relationship between two variables

See handout '**A summary of some useful correlation coefficients and when to use them**' for more details

The regression equation

a is the regression constant; b is the regression coefficient

$$Y=a+bx$$

gradient (slope) of regression line

independent variable (predictor variable)

regression line intercepts vertical axis at point (0,a)

response variable (dependent variable)

(x,y) on scatter plot → (x,Y) on regression line

Residual for point (x,y) : Y-y

Fitted value – actual value

The regression coefficient b represents the average change in Y if x increases by one unit.

Why perform simple linear regression?

Linear regression allows us to predict the value of the dependent variable given any value of the predictor variable which lies within the range represented by the sample.

(Note: we cannot use the regression model to extrapolate beyond the sample range for the predictor variable.)

Two further checks (both of which should be applied) to ensure that linear regression analysis is the best way forward:

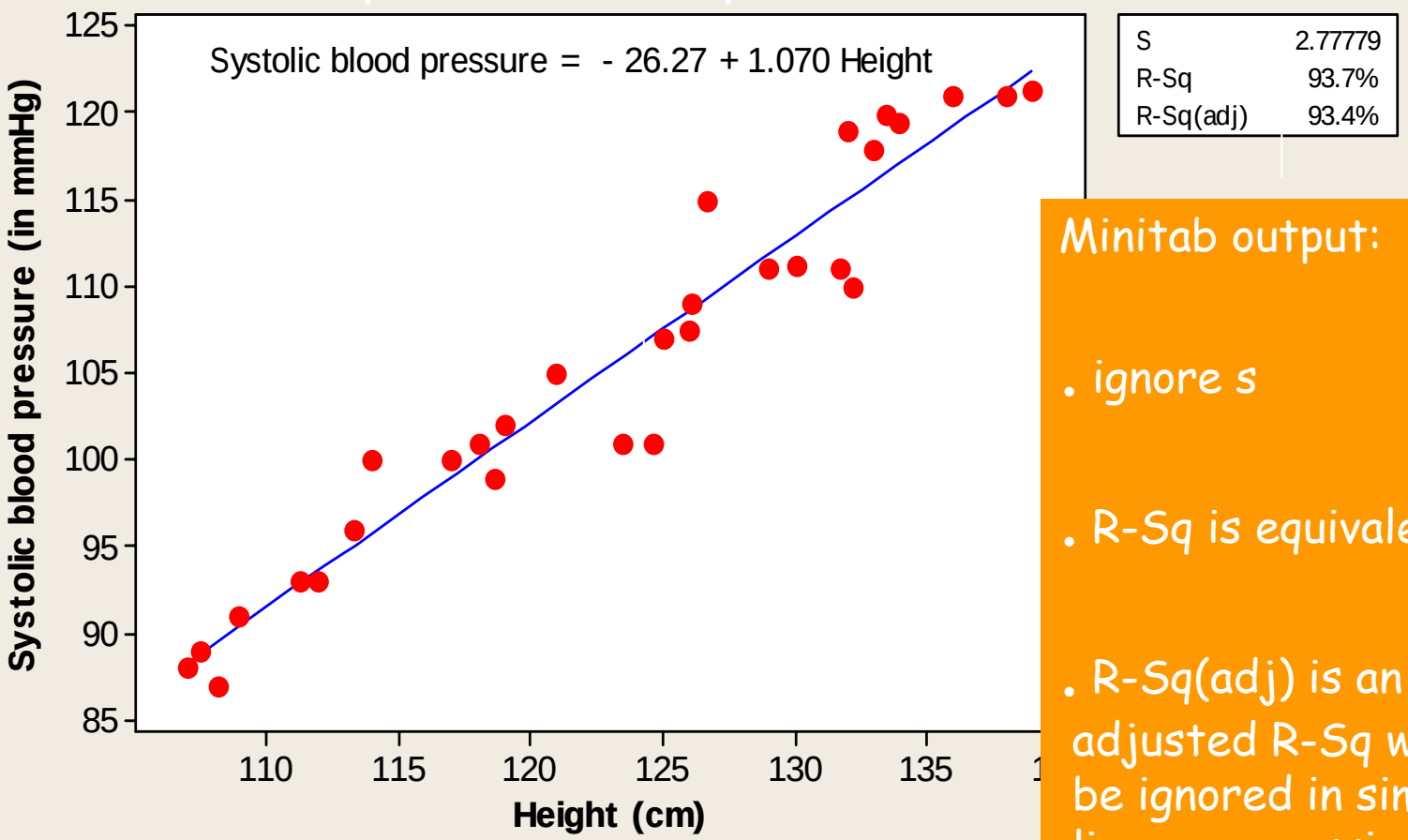
1) check that the distribution of the residuals is Normal

2) check that in a plot of the standardized residuals against the standardized fitted values, the residuals do not systematically increase or decrease with increasing fitted value [the dependent variable has constant variation across the range of the predictor variable]

CAN USE SPSS* FOR 1) AND 2)!

*SPSS: statistical package for the social sciences

Regression plot: Minitab uses the least squares method to obtain the line of best fit. This method proceeds by minimizing the sum of the squares of the residuals



Minitab output:

. ignore s

. R-Sq is equivalent to r^2

. R-Sq(adj) is an adjusted R-Sq which can be ignored in simple linear regression analysis

The interpretation of the results of diagnostic tests:

Sensitivity, Specificity, Positive Predictive Value (PPV) and Negative Predictive Value (NPV)

Motivating table (e.g. for pain prediction test for cancer patients)

Motivating table (e.g. for Pain prediction test for cancer patients)

TP, FP,FN,TN: true positive, false positive, false negative and true negative		TARGET DISORDER (e.g. pain)		
		Present	Absent	
DIAGNOSTIC TEST RESULT (e.g. pain prediction)	+	a (TPs) (997)	b (FPs) (950)	a + b
	-	c (FNs) (6)	d (TNs) (998,010)	c + d
Totals		a + c	b + d	a+b+c+d

Sensitivity (of test) =a/(a+c)
=997/(997+6)=0.99

Specificity (of test) =d/(b+d)
=998010/(950+998010)=1.00 (to 2 dp)

PPV=a/(a+b)
=997/(997+950)=0.51

NPV=d/(c+d)
=998010/(6+998010)=1.00 (to 2 dp)

Conceptual interpretations

Sensitivity: proportion of individuals with disorder who have a positive diagnostic test result

Specificity: proportion of individuals without disorder who have a negative diagnostic test result

---

PPV: proportion of people with positive diagnostic test result who have the disorder

NPV: proportion of people with negative diagnostic test result who do not have the disorder

---

Remember!

Above example (with calculations) shows sensitivity and specificity can fail to pick up the extent of possible distress caused through false positives!

[From PPV, 49% of positive diagnoses are wrong in this case!]