Examining a single factor (univariable)

Content

- The difference between paired and unpaired data
- Analysis for continuous data: ‘taking measurements on people/things’
- Analysis for categorical data: ‘counting people/things’
- Analysis for time-to-event data
- This session introduces you to the common statistical tests used

Paired or Unpaired Data?

Unpaired data (independent data)

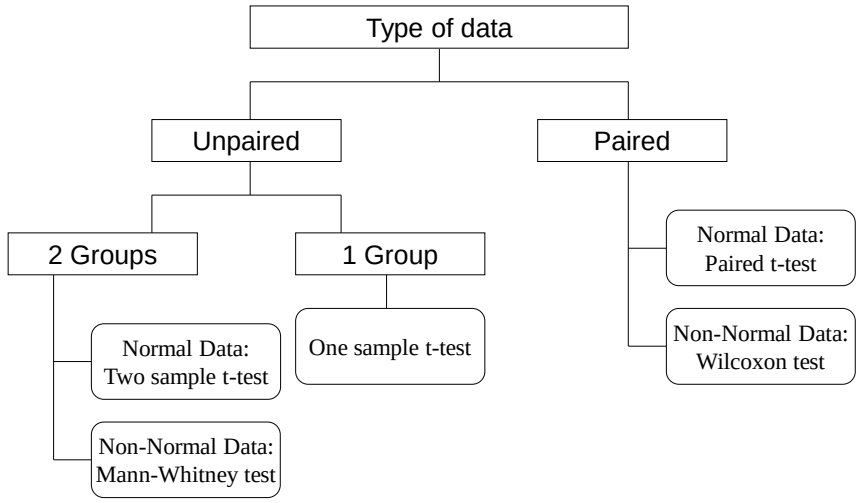
- Two (or more) groups of interest: the subjects in one group are **entirely separate** from subjects in all other groups
- The same outcome measure is taken on each group
- **Example 1**: A clinical trial of heart failure patients, randomly assigned to either beta-blocker or placebo, looking at the effect on mortality
- **Example 2**: A cross-sectional survey comparing respiratory function in males and females with COPD
- **Example 3**: Comparing tumour growth between two different groups of mice

Paired or Unpaired Data?

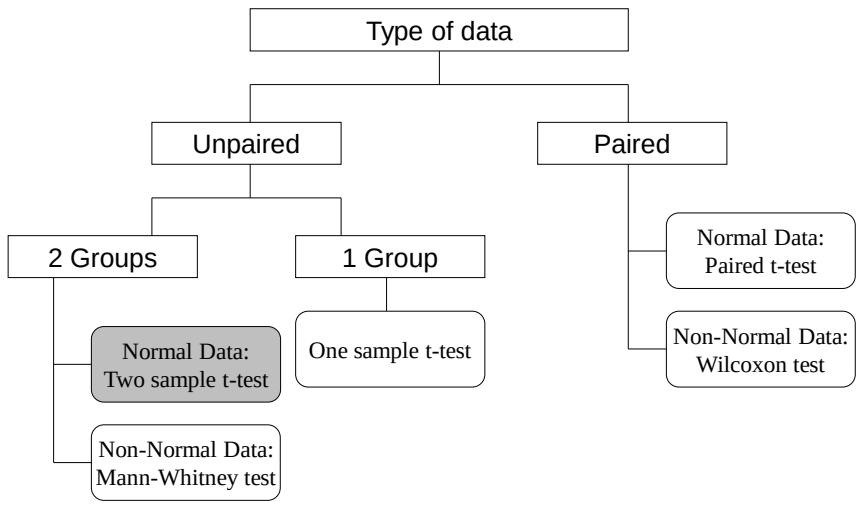
Paired data

- Two (or more) measurements of the same outcome measure, made on the **same subject**
- These are usually made at different time points
- **Example 1**: Heart rate measurements made on a group of healthy volunteers before and after exercise
- **Example 2**: Lung function measurements in asthma patients made before and after taking a new drug
- **Example 3**: Voting preference in a group of floating voters before and after seeing a series of party political broadcasts

Decision tree for continuous data: 'Taking measurements on people'



Unpaired Normal data with 2 groups

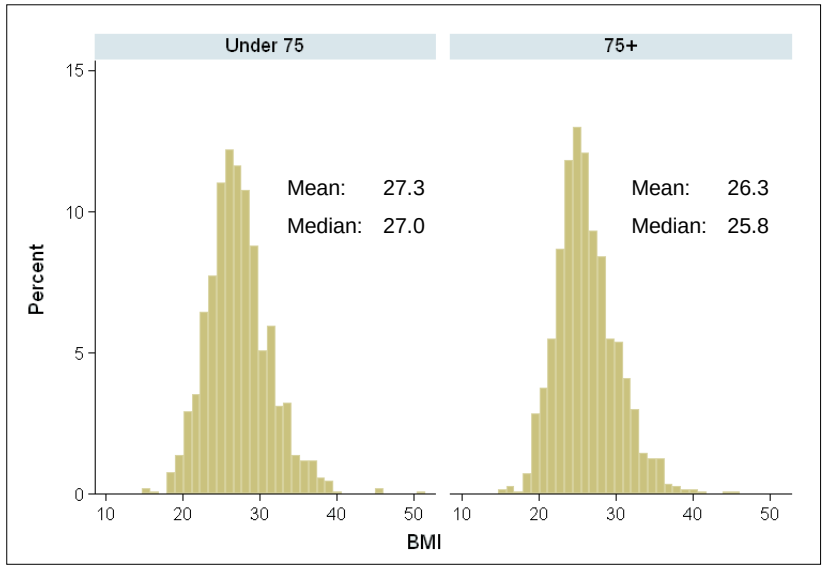


Unpaired Normal data with 2 groups

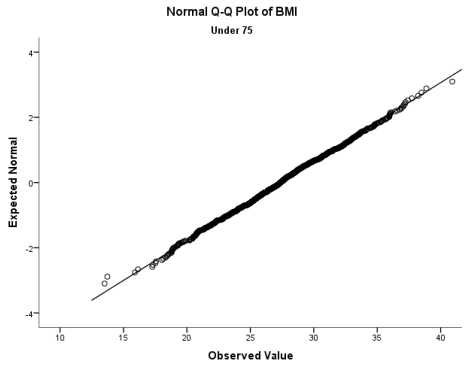
- The **two sample t-test** is used to compare the means of two independent groups
- **Assumptions**
  - The two groups are independent
  - The data are Normally distributed in both groups
- There are 2 different versions of the two sample t-test, depending on whether the two samples have equal variances (i.e. Standard Deviation^2) or not
- You need to examine the data and decide which method is more appropriate

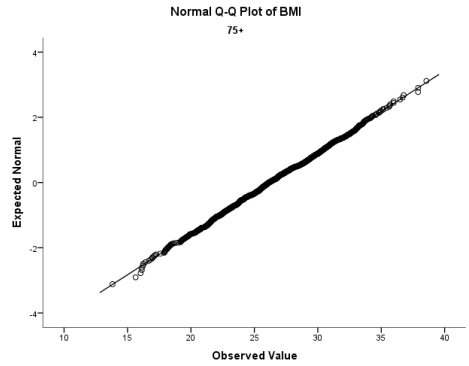
- **Example**: BMI of heart failure patients who are younger or older than 75 years

Unpaired Normal data with 2 groups



Unpaired Normal data with 2 groups





Unpaired Normal data with 2 groups

Two-sample t test with equal variances

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Under 75	1024	27.2847	.1272961	4.073476	27.03491	27.53449
75+	1093	26.27192	.1196679	3.956288	26.03711	26.50672
combined	2117	26.7618	.0878979	4.044259	26.58943	26.93418
diff	**1.012786**	**.1745471**			**.6704838**	**1.355088**

diff = mean(Under 75) - mean(75+) t = 5.8024

Ho: diff = 0 degrees of freedom = 2115

Ha: diff < 0 | **Ha: diff != 0** | Ha: diff > 0

Pr(T < t) = 1.0000 | **Pr(|T| > |t|) = 0.0000** | Pr(T > t) = 0.0000

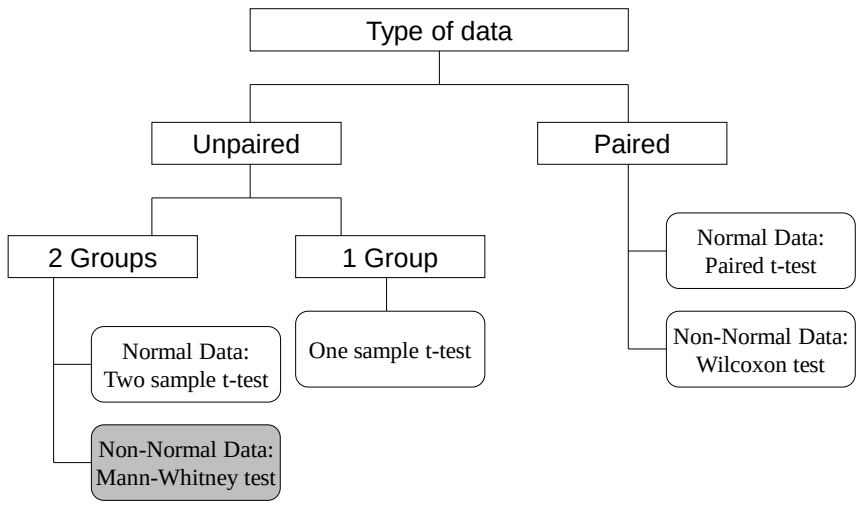
Effect size & 95% CI

P-value

Unpaired Normal data with 2 groups

- The p-value from the t-test is very small (<0.001)
- If we **assume the true difference is zero**, a difference in BMI of 1.01 or more, in either direction, would only be seen in less than 1 in 1000 similar studies due to chance
- Because the p-value is very small, we conclude that this result is unlikely to be due to chance
- Therefore there is likely to be a real difference in BMI between heart failure patients age <75 and ≥75
- However, we also need to consider the **clinical importance** of the result
- A difference in BMI of 1 unit is not important
- The p-value is a result of the large sample size (>2000)

Unpaired non-Normal data with 2 groups

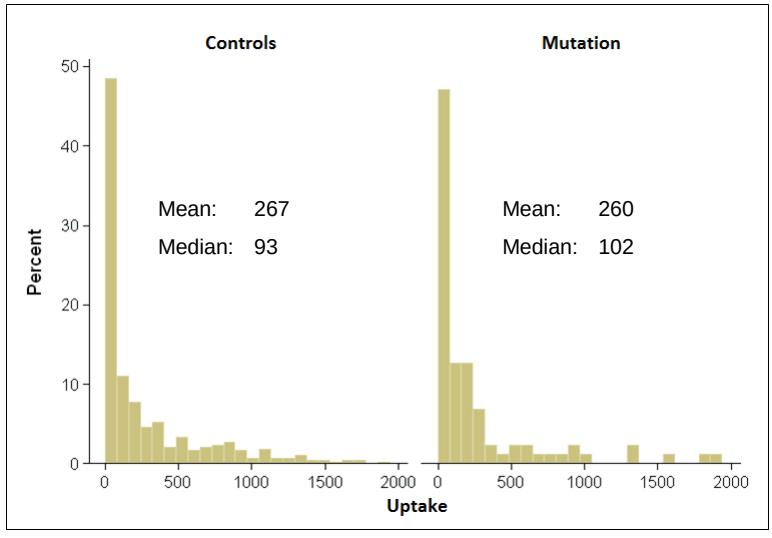


Unpaired non-Normal data with 2 groups

- If the data are skewed then a **Mann-Whitney test** is the most appropriate test to use
- **Assumptions**
  - The two samples are independent
  - The data are distributed similarly in the two groups
- This test uses the ranks of the data, not the actual values of the measurements (therefore it is not affected by very large or very small values)
- It determines whether the distribution of one group is shifted to the left or the right of the other group

- **Example**: Comparison of protein uptake in mice with a gene mutation and controls (i.e. no mutation)

Unpaired non-Normal data with 2 groups



Unpaired non-Normal data with 2 groups

- Each observation is ranked (from lowest to highest)

Uptake	20	23	30	37	44	61	92	128	452	1013
Rank	1	2	3	4	5	6	7	8	9	10
Group	A	B	B	A	B	B	A	A	B	A

- The ranks for observations in the two groups are added to calculate the **ranksum** for each group
- These are compared to what would be expected if the two groups had the same distribution, in order to calculate the p-value

Unpaired non-Normal data with 2 groups

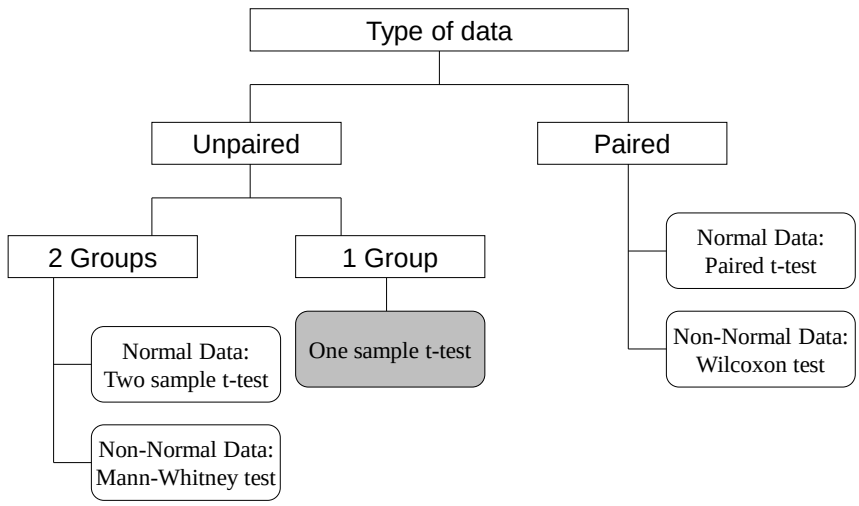
```
Mann-Whitney test

    Group |      obs    rank sum    expected
----------+---------------------------------
 Controls |      478      135311      135274
 Mutation |       87       24584       24621
----------+---------------------------------
 combined |      565      159895      159895

unadjusted variance  1961473.00
adjustment for ties    -1038.48
                     ----------
adjusted variance    1960434.52
Ho: Uptake(Group==Controls) = Uptake(Group==Mutation)
                           z =    0.026
                Prob > |z| =    0.9789
```

P-value

Unpaired non-Normal data with 2 groups

- The p-value is large (0.979), so there is little evidence of a difference in uptake between mice with or without mutation
- In this case we would conclude the two groups had the same distribution of protein uptake
- When reporting these results, you should provide the median in each group and the p-value
- We report the median because it is easier to interpret, and less influenced by outliers than the mean
- But remember that the p-value is not comparing the median in each group (it compares the ranks of the data)
- Some statistical packages will also calculate the difference between two medians, and the 95% CI for this difference

Normal data with 1 sample



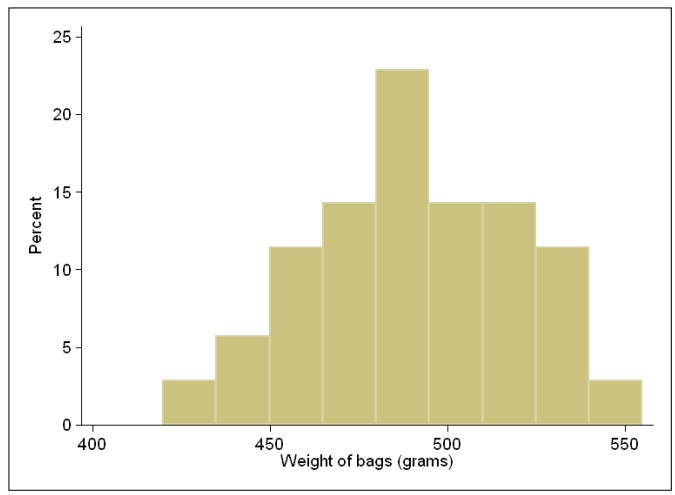
Normal data with 1 sample

- The **one sample t-test** is used when comparing an observed mean from a sample against a **known** or target value
- **Assumptions**
  - The data are Normally distributed
  - With the one sample t-test, we assume the true mean to be equal to the target value

- **Example**: Actual weights of bags of sweets coming out of a machine, which have a claimed weight of 500 grams

Normal data with 1 sample

517.7	487.7
494.5	495.7
539.7	456.1
485.7	474.7
438.6	472.1
490.4	516.8
527.3	499.5
511.6	509.0
523.1	471.8
489.4	477.2
460.8	529.7
445.5	461.3
482.6	554.0
499.5	497.1
429.0	476.7
452.6	538.8
513.2	481.0
489.8	
Mean	**491.2**



Normal data with 1 sample

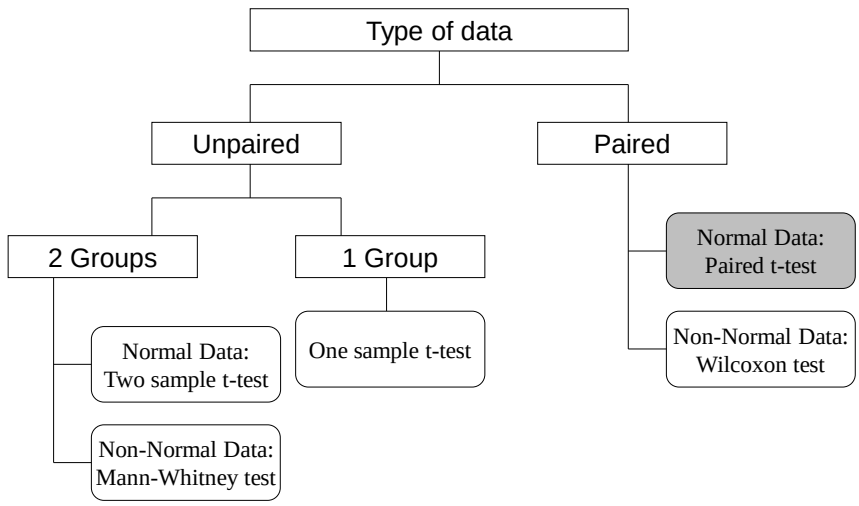
```
One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Weight   |      35    491.1523    5.026074    29.73465    480.9381    501.3665
------------------------------------------------------------------------------
    mean = mean(weight)                                     t =  -1.7604
Ho: mean = 500                                 degrees of freedom =       34

    Ha: mean < 500               Ha: mean != 500                 Ha: mean > 500
 Pr(T < t) = 0.0437         Pr(|T| > |t|) = 0.0873          Pr(T > t) = 0.9563
```

Estimate & 95% CI

P-value

Since the p-value is greater than 0.05 we would accept that this sample of sweets came from a population whose mean weight could be 500 grams.
However, in this instance the machine may need to be checked since the upper limit of the 95% CI is only just above the target of 500 grams.

Paired Normal data with 1 group

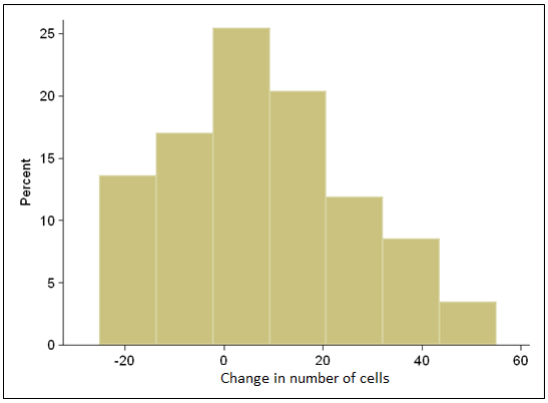


Paired Normal data with 1 group

- The **paired t-test** is used to test for differences in means between measurements made at two time points on the same subject
- We therefore work with the **difference** between the measurements for each subject, not the before and after measurements
- **Assumptions**
  - The two samples are not independent, but paired data taken from the same subject at different times
  - The differences between the measurements at the two points are Normally distributed (not the distribution of each measurement)

- **Example**: Number of bacteria cells in a Petri dish at baseline and after 30 minutes

Paired Normal data with 1 group



Dish	Cells_00	Cells_30	Difference
1	53	43	-10
2	52	54	2
3	46	61	15
4	63	45	-18
5	37	49	12
6	44	70	26
7	33	71	38
8	31	59	28
9	34	64	30
10	68	37	-31
11	17	52	35
12	25	32	7
.	.	.	.
.	.	.	.
.	.	.	.
Mean	**44.4**	**52.6**	**8.2**

Paired Normal data with 1 group

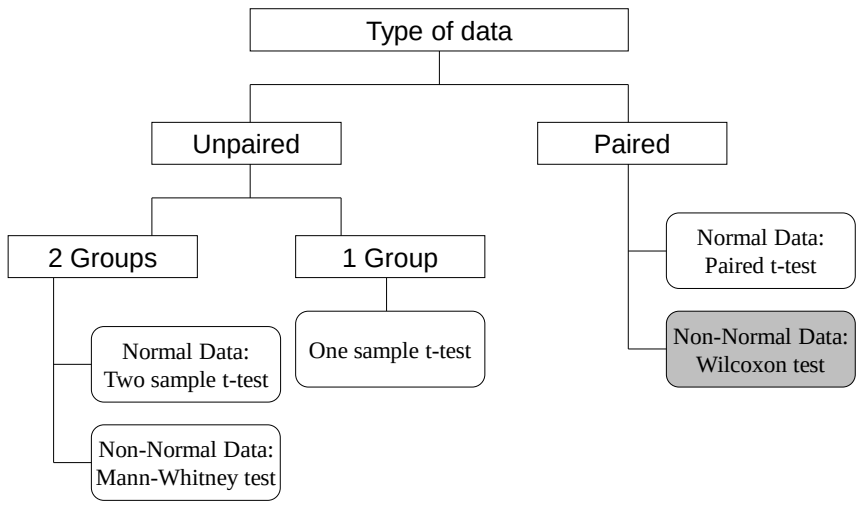
```
Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
cells_30 |      59    52.59887    2.534066    19.46453    47.52639    57.67135
cells_00 |      59    44.42561    2.473531    18.99955     39.4743    49.37692
---------+--------------------------------------------------------------------
    diff |      59    8.173258    2.398973    18.42686    3.371191    12.97532
------------------------------------------------------------------------------
     mean(diff) = mean(qol_30 - qol_00)                           t =   3.4070
 Ho: mean(diff) = 0                           degrees of freedom =       58

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9994         Pr(|T| > |t|) = 0.0012          Pr(T > t) = 0.0006
```

Effect size & 95% CI

P-value

Since the p-value is small (0.001) we would conclude that there is a real effect, i.e. the number of bacteria cells has increased in 30 minutes

Paired non-Normal data with 1 group

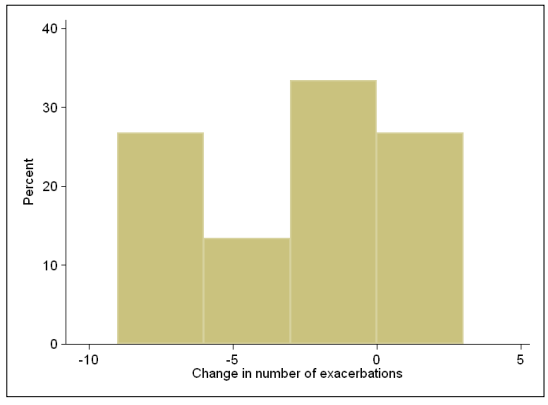


Paired non-Normal data with 1 group

- If the assumption of Normally distributed differences does not hold then it is not appropriate to use the paired t-test
- In this case we would use the **Wilcoxon sign rank test**
- This is like the Mann-Whitney, i.e. it uses the ranks of the data values, instead of the actual data when calculating the p-value
- **Assumptions**
  - The two samples are not independent, but paired data taken from the same subject at different times

- **Example**: Number of exacerbations per month in COPD patients before and after a light exercise program

Paired non-Normal data with 1 group



Patient	Before	After	Difference
1	26	20	-6
2	26	18	-8
3	27	18	-9
4	22	22	0
5	23	25	2
6	22	25	3
7	26	19	-7
8	25	24	-1
9	25	18	-7
10	25	19	-6
11	22	23	1
12	26	24	-2
13	22	21	-1
14	23	21	-2
15	27	26	-1
Median	**25**	**21**	**-2**

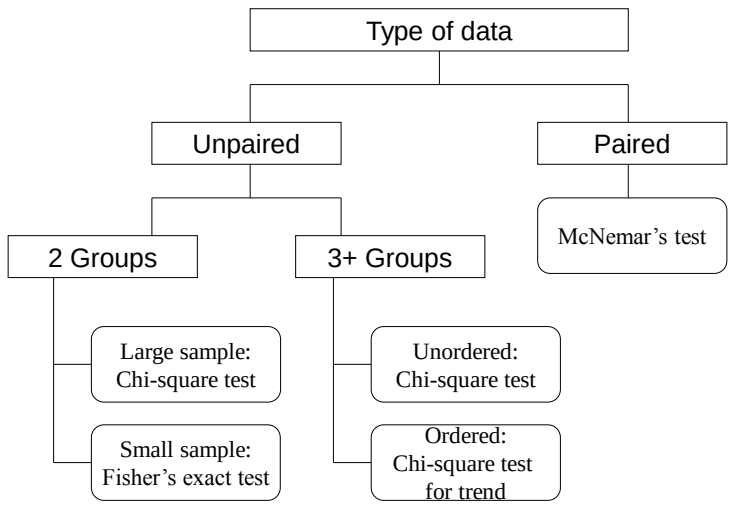
Paired non-Normal data with 1 group

```
Wilcoxon signed-rank test

       sign       obs   sum ranks    expected
   positive         3        19.5        59.5
   negative        11        99.5        59.5
       zero         1           1           1
        all        15         120         120

unadjusted variance      310.00
adjustment for ties       -2.00
adjustment for zeros      -0.25
adjusted variance        307.75

Ho: after = before
             z =  -2.280
Prob > |z| =   0.0226
```

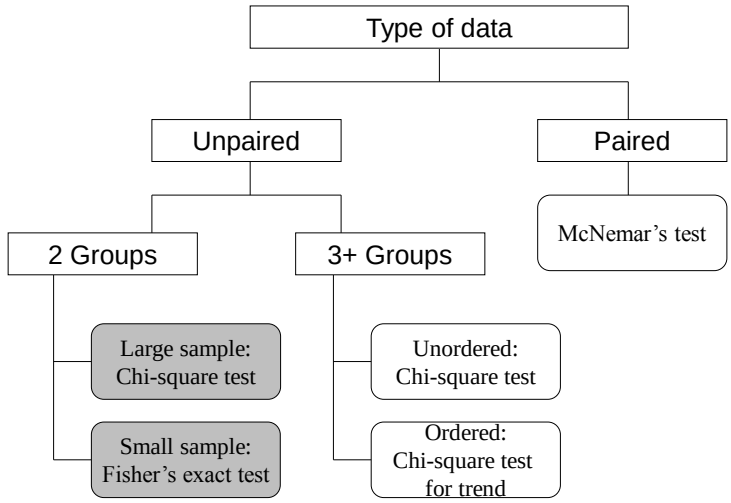
P-value →

Since the p-value is small (0.023) we would conclude that light exercise helped to reduce the number of exacerbations per month

Decision tree for categorical data: 'Counting people'



Comparing 2 independent proportions



Comparing 2 independent proportions

- One of the most common types of test performed is where both the explanatory factor and the outcome are **dichotomous** variables – only takes 2 values
- In clinical trials an example would be looking at the 6 month mortality status (outcome) of patients on one of two different treatments (factor)
- In observational studies we might look at the smoking history (factor) of subjects who do or do not have lung cancer (outcome)
- Results are often presented in a 2x2 table
- **Assumption**
  - The two groups are from independent samples

Comparing 2 independent proportions

- The **Chi-square test** is used to look for an association between the outcome and the factor. It works by comparing the **expected** frequency in each cell of the 2x2 with the **observed** result
- If the expected frequency in any cell of the 2x2 table is less than 5, then the chi-square test is not appropriate, and **Fisher's exact test** should be used. This usually occurs when the number of subjects is small

- The interpretation of the results is the same for both tests
- Software such as SPSS will warn you when the chi-square test is not appropriate

Comparing 2 independent proportions

- **Example**: Comparing the proportion of workers reporting symptoms of repetitive strain injury (RSI) in 2 different types of employment

	Data entry	Secretarial	Total
No symptoms	34 (63.0%)	15 (51.7%)	49 (59.0%)
RSI symptoms	20 (37.0%)	14 (48.3%)	34 (41.0%)
Total	54	29	83

- Essentially we are testing the difference between the proportion of employees with symptoms in each arm – 37.0% vs 48.3%
- We are testing this difference against the **assumption that the two proportions are the same** – 41.0%

Comparing 2 independent proportions

			Job		
			Data entry	Secretarial	Total
RSI	No	Count	34	15	49
		Expected Count	31.9	17.1	49.0
		% within Job	63.0%	51.7%	59.0%
	Yes	Count	20	14	34
		Expected Count	22.1	11.9	34.0
		% within Job	37.0%	48.3%	41.0%
Total		Count	54	29	83
		Expected Count	54.0	29.0	83.0
		% within Job	100.0%	100.0%	100.0%

Observed data

'Expected' assuming there is no association between Job & RSI

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	.985	1	.321
N of Valid Cases	83		

P-value

Here the p-value from the chi-square test is large (0.321), so there is little evidence that RSI symptoms differ between the two types of employment

Comparing 2 independent proportions

- Now consider a similar study, but with much smaller numbers

			Job		
			Data entry	Secretarial	Total
RSI	No	Count	4	8	12
		Expected Count	6.3	5.7	12.0
		% within Job	36.4%	80.0%	57.1%
	Yes	Count	7	2	9
		Expected Count	4.7	4.3	9.0
		% within Job	63.6%	20.0%	42.9%
Total		Count	11	10	21
		Expected Count	11.0	10.0	21.0
		% within Job	100.0%	100.0%	100.0%

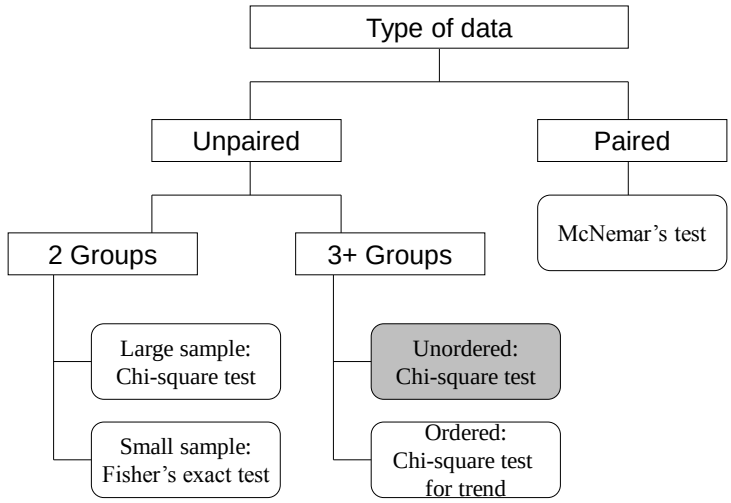
- Two of the cells have expected frequencies <5, so we should use **Fisher's exact** test for this example

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	4.073[a]	1	.044	
Fisher's Exact Test				.080
N of Valid Cases	21			

a. 2 cells (50.0%) have expected count less than 5. The minimum expected count is 4.29

- The two p-values here are different sides 0.05, so use of the chi-square test would have lead to erroneous conclusions

3+ unordered independent proportions



3+ unordered independent proportions

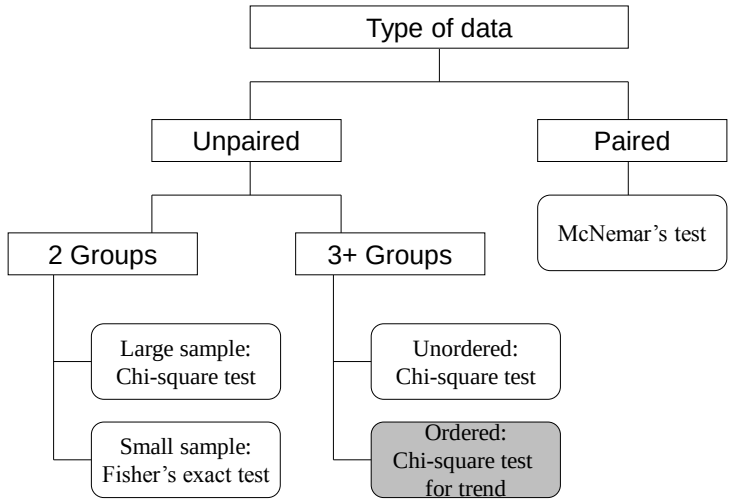
- We may want to compare proportions between several **unordered, independent categories**
- The **chi-square test** is also used in these situations
- As before, we are testing these observed proportions against the **assumption that the proportion in each group is the same**

RSI symptoms	Data entry	Secretarial	Clerical	Maintenance	Total
No	34 (63.0)	15 (51.7)	48 (70.6)	15 (75.0)	112
Yes	20 (37.0)	14 (48.3)	20 (29.4)	5 (25.0)	59
Total	54	29	68	20	171

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	4.167	3	.244
N of Valid Cases	171		

The p-value from the chi-square test is large (0.244), so there is little evidence that RSI is more/less associated with one of the types of employment

3+ ordered independent proportions



3+ ordered independent proportions

- One other possible scenario would be to look at proportions for **ordered, independent** groups
- The ordering of the groups could include escalating doses of a particular treatment or a child's order of birth within families with three children
- If the categories are ordered, it is important to consider this in the analysis; often we would be looking for a **linear trend** across the categories, in relation to the outcome
- The appropriate test to use for this analysis is the **chi-square test for trend**
- As with all previous uses of the chi-square test, the assumption of independence between the groups must still hold

3+ ordered independent proportions

- **Example**: Two year mortality rate of cystic fibrosis patients, based on their lung function assessment
- This time we are testing the proportions against an **assumption that no linear trend is present**

Dead at 2 years	Mild	Moderate	Severe	Total
No	52 (75.4)	43 (52.4)	26 (44.8)	121
Yes	17 (24.6)	39 (47.6)	32 (55.2)	88
Total	69	82	58	209

3+ ordered independent proportions

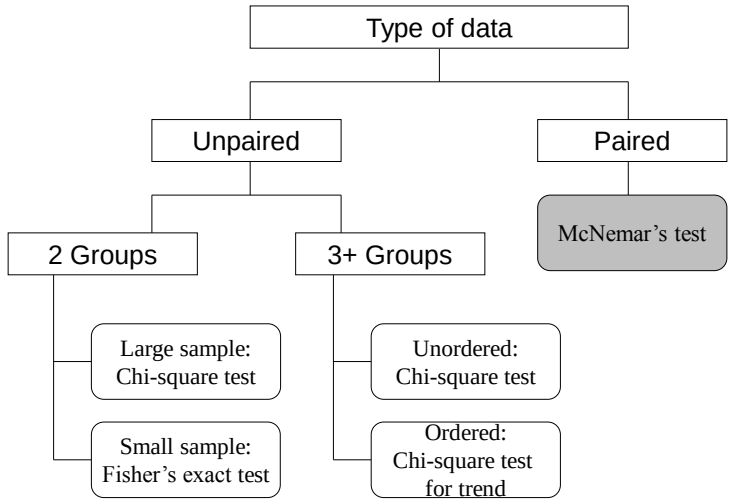
			Severity			
			Mild	Moderate	Severe	Total
2-year mortality	Alive	Count	52	43	26	121
		Expected Count	39.9	47.5	33.6	121.0
		% within Severity	75.4%	52.4%	44.8%	57.9%
	Dead	Count	17	39	32	88
		Expected Count	29.1	34.5	24.4	88.0
		% within Severity	24.6%	47.6%	55.2%	42.1%
Total		Count	69	82	58	209
		Expected Count	69.0	82.0	58.0	209.0
		% within Severity	100.0%	100.0%	100.0%	100.0%

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	13.700[a]	2	.001
Linear-by-Linear Association	12.446	1	.000
N of Valid Cases	209		

P-value

Here the p-value for the trend analysis is highly significant, so we would conclude that there was a linear association between disease severity and two year mortality

Paired categorical data



Paired categorical data

- As with the paired analysis of continuous data, we need to use the right test if we have paired categorical data
- Here we assume that the outcome is measured on the **same subject** at two different time points
- **McNemar's test** is the appropriate test in this instance. However, it only works when the outcome is binary

- In this analysis we are testing whether people are more likely to change in one direction than the other (i.e. we look at the discordant pairs)
- Subjects who give the same response at both time points do not contribute anything to the analysis, so are ignored

Paired categorical data

- **Example**: Voting preference in a ballot (for/against) before and after seeing a party political broadcast
- We are testing to see whether the opinions changed, against an **assumption that the broadcasts did not change opinions more in one way than the other**

	After		
Before	Against	For	Total
Against	25	62	87
For	40	60	100
Total	65	122	187

Paired categorical data

```
                |          After          |
     Before     |    Against       For    |      Total
----------------+-------------------------+-----------
     Against    |         25          62  |         87
     For        |         40          60  |        100
----------------+-------------------------+-----------
      Total     |         65         122  |        187
McNemar's chi2(1) =      4.75    Prob > chi2 = 0.0294
```

P-value

The p-value is <0.05 so we could conclude that the broadcast did effect people's opinion.

In this instance people who were initially 'against' the ballot question were more likely to change their view to being 'for' it

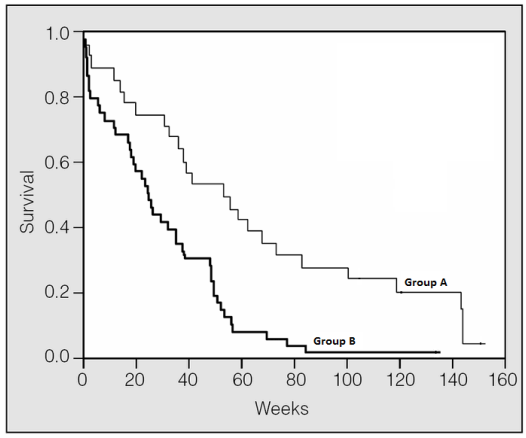
Paired categorical data

- With McNemar's test we are actually comparing the absolute values – in this example we have 62 vs. 40
- This is different from the chi-square test which compares the proportions in the separate groups
- From these results we can calculate an odds ratio, which is simply 62/40 = 1.55. The 95% confidence interval for this example is 1.03 to 2.37
- The 95% CI is completely above the critical value of 1, so the effect of the broadcast is significant
- The OR is interpreted as meaning that people were 55% more likely to change from being 'against' the ballot question to 'for' it than they were to change their mind in the other direction

Time-to-event data

- When data are looking at time until a specific event happens, **survival analysis** techniques need to be used
- Outcome is often related to survival, but can also include any definition of an event, including 'positive' outcomes, such as age at first child born or days until discharge from hospital ward
- Time-to-event data are usually presented using **Kaplan-Meier curves** and analysed using the **log-rank test**
- This test can easily compare two or more survival curves
- **Example**: Time to a major cardiac event after diagnosis in two groups of patients

Time-to-event data (log-rank test)



Overall Comparisons

	Chi-Square	df	Sig.
Log Rank (Mantel-Cox)	8.751	1	.000

P-value

The p-value here is highly significant (<0.001) indicating that the two survival curves are different.

Test Statistics

- All of the methods in this talk will calculate a **test statistic**
- This test statistic will then be converted into a p-value
- The p-value is interpreted to make conclusions regarding statistical significance
- Always important to use the appropriate test for the type of data that is being analysed
- **Type of data**
  - Continuous (‘taking measurements’)
  - Categorical (‘counting people’)
  - Paired or unpaired?
  - Number of groups?
  - Time-to-event data