Multiple groups or comparisons

- Comparing several groups
  - When outcome measure is based on ‘counting people’
  - When the outcome measure is based on ‘taking measurements on people’
- Having several outcome measures

Multiple groups or comparisons

- When the outcome measure is based on ‘counting people’, this is categorical data.
- The groups can be compared with a simple chi-squared (or Fisher’s exact) test.

Comparing multiple groups
ANOVA – Analysis of variance

When the outcome measure is based on 'taking measurements on people data'

- For 2 groups, compare means using t-tests (if data are Normally distributed), or Mann-Whitney (if data are skewed)
- Here, we want to compare more than 2 groups of data, where the data is continuous ('taking measurements on people')
- For example, comparing blood pressure between 3 dose groups (5mg, 10mg, 20mg) and determine which dose reduces blood pressure the most
- For normally distributed data we can use ANOVA to compare the means of the groups.

ANOVA – Example

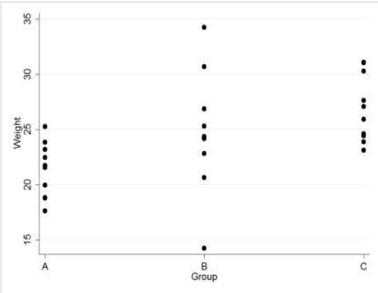
Example: Weight lost by rats given 3 diets; A, B & C .

- Question: Are there differences in mean weights between any of the 3 diets?
- If there are differences, where do they lie?
- Note this is a one-way ANOVA – only considering one source of variability (Diet).
- If gender or another appropriate covariate were also important, then a 2-way ANOVA might be considered instead.

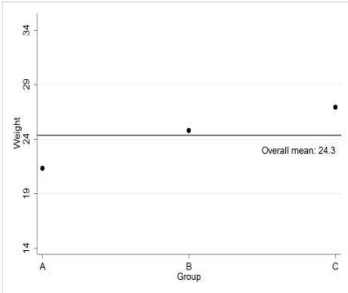
Within vs Between Variability

Concept of ANOVA is to separate the SOURCES of variability:

Total Variability = Variability within groups + Variability between groups



Raw data: can look at variability **within** in each group



Mean values: can look at variability **between** groups, i.e. how they differ from the overall mean of 24.3

Rat sample 1

	Diet A	Diet B	Diet C
	21.3	25.3	27.3
	22.0	24.6	26.7
	21.1	25.0	26.9
	21.2	24.6	27.1
	21.1	25.2	26.7
	21.5	24.4	26.9
	21.5	25.0	27.0
	20.4	25.0	26.6
	21.3	24.6	27.0
	21.5	24.2	27.0
Mean	**21.3**	**24.8**	**26.9**
SD	**0.4**	**0.4**	**0.2**

Rat sample 2

	Diet A	Diet B	Diet C
	23.84	20.66	23.90
	23.21	24.34	31.10
	21.73	14.27	24.42
	18.79	30.69	31.06
	22.46	22.84	27.63
	19.96	24.18	23.14
	17.64	26.88	25.91
	21.58	25.31	27.10
	18.83	24.29	30.29
	25.27	34.26	24.60
Mean	**21.3**	**24.8**	**26.9**
SD	**2.5**	**5.4**	**3.0**

- Both samples have the same differences between group (A, B or C) means – we can say variation between means in each data set is the same
- But the variability within a sample in each set is different. Set 1 is tighter around its mean (lower SDs) than set 2.
- Although they have the same difference between the means, which data set is more reliable to make judgements about real differences between A B and C?

- The ratio of the variability **Between** means to variability **Within** the samples is used to determine whether differences in means exist:
- There appears to be stronger evidence supporting true differences between means in data set 1 than in data set 2 because the within group variability (i.e. within A, B or C) is smaller when compared to the between group variability

Variability	Rat sample 1	Rat sample 2
Between	Same	Same
Within	Smaller	Larger
Ratio	Larger in sample 1 than sample 2	

- **If the ratio of Between to Within is > 1 then it indicates that there may be differences between the groups .**
- **Results displayed in an ANOVA table**

Data entry

Rat ID number	Diet	Weight(g)
1	A	23.84
2	A	23.21
3	B	20.66
4	B	24.34
5	C	23.90
6	C	31.10
etc.		

Most stats packages will require data to be in the form above (rather than in separate columns for each diet as in the previous slide).

One-way ANOVA in SPSS

Below is the output from SPSS, comparing the mean weights of the rats in Sample 2

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	158.744	2	79.372	5.370	.011
Within Groups	399.062	27	14.780		
Total	557.805	29			

P-value for the differences between diet groups

Between Diet (group) variability

Within Diet (group) variability

Ratio of between To within variability

One-way ANOVA in SPSS

What would happen if we ran the same test on sample 1? (The sample of rats with less variability)

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	158.331	2	79.165	515.294	.000
Within Groups	4.148	27	.154		
Total	162.479	29			

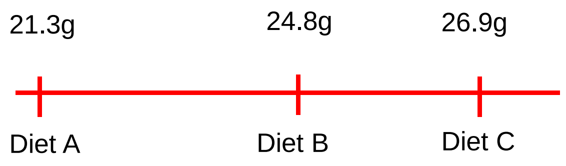
The smaller within diet variability leads to a much larger ratio

Between Diet (group) variability is the same (~79)

Within Diet (group) Variability is MUCH smaller

The larger ratio, gives us an even smaller p-value; we can be more sure that there is a real difference between the diets.

- Now that we know that the <u>mean</u> weights are different (F-test) across the diets groups, which particular diets are different from each other?



- Mean weight lost after diet C is greater than the other 2 diets
- There are larger differences in weight lost between diets A vs. C than diet B vs. C (5.6g difference and 2.1g difference)
- Diets B and C might be more similar because the mean rat weights are closer together.
- Need to do pairwise tests (A vs. B, A vs. C) to confirm whether diet A (standard) is significantly different to the other 2 diets

---

- Many researchers are interested in pairwise comparisons.
- They often do several independent t-tests (for continuous data)
- E.g.: if there are 3 groups of people, A, B & C, there is a separate t-test for
  - A vs B
  - A vs C
  - B vs C
- Suppose we wanted to examine differences between 5 groups; there are 10 possible pairs, and therefore 10 effect sizes and 10 p-values.

- The usual error rate for a single comparison is 5%, i.e. we allow ourselves to falsely conclude that there is an effect, when there really isn't one, 5% of the time (or 1 in every 20 studies of the same size)
- If each t-test is done at the 5% level, then the ***overall error*** *rate* for 10 comparisons is $1 - [(1\text{-}a)^{10}]$, i.e. $1 - [(1\text{-}0.05)^{10}]$ which = 0.4, or 40%!
- We could make a mistake (false conclusion) 40% of the time
- Do not perform lots of <u>independent</u> pairwise comparisons

There are different approaches to control the false positive error rate, but a simple way for continuous data ('taking measurements on people') is as follows:

a) First we see if at least one of the means differ from the others. We use a test called the F-test from the ANOVA

b) If the p-value is small (i.e. <0.05), we know that at least one of the groups is different to the others, so then we can proceed to look at differences between specific groups. Otherwise, conclude there is no evidence of a difference

- We are not usually interested in all comparisons
- They are usually pre-planned
- E.g., New Diets B and C, each compared with a standard Diet A (i.e. 2 main comparisons)
- Comparing between B and C might be of less importance

Comparison	Estimate	Raw p-value (95% CI)	Adjusted p-value
A vs. B	-3.4	0.055 (-7.0 to 0.09)	0.17 (-7.8 to 0.9)
A vs. C	-5.6	0.003 (-9.1 to -2.1)	0.009 (-10 to -1.2)

- The ANOVA provides a p-value and 95% CI that allows for having several other comparisons. These are the ones to interpret

- Note the impact on p-values and wider confidence intervals to adjust for having ≥2 comparisons (i.e. a higher false positive rate)

Non-Parametric testing

- When the data are not Normally distributed, we use a non-parametric analogue of one-way ANOVA (called Kruskal-Wallis ANOVA)

- It is an extension of the Wilcoxon Rank Sum test

- The analysis is based on the ranks of the data (not the actual values)

- Example uses the same data – Compare Weights of Rats from 3 diets A, B and C

Kruskal-Wallis Test in SPSS

Kruskal-Wallis Test

Ranks

	Diet	N	Mean Rank
Weight	A	10	8.30
	B	10	16.60
	C	10	21.60
	Total	30	

Test Statistics[a,b]

	Weight
Chi-Square	11.646
df	2
Asymp. Sig.	.003

a. Kruskal Wallis Test

b. Grouping Variable: Diet

- The p-value tells you if there is a difference somewhere between the groups. As with ANOVA we would need to inspect the data/perform pairwise tests to find out where.
- When presenting/interpreting the results we would present the medians along with the p-value.
- The Wilcoxon Mann-Whitney test can be used to perform pairwise comparisons, but as before, you may need consider adjusting the p-values for multiple tests.

Multiple endpoints

- The more analyses done on the same dataset, the more likely that you are to find a statistically significant result, when there really isn't an effect

Example: examining a new treatment for COPD (New vs Standard)

- Possible primary endpoints are:
  - FEV1 (Forced Expiratory Volume in 1 Second)
  - FVC (Forced Vital Capacity)
  - FEV1/FVC
  - Number of exacerbations
  - Time until first exacerbation
  - Time until treatment stopped early
- Each of these can be examined in relation to the treatment allocation to produce an effect size (a mean difference, or regression coefficient), 95% CI and p-value
- But this increases the error rate (as in having multiple comparisons)

Simple solutions

- Have 1 or 2 pre-specified endpoints
- These are the ones on which you will make major decisions.

- Adjust p-values using the **Bonferroni adjustment** (there are other, more complex methods to adjust p-values)

- Alternatively, specifically state that the study is a pilot or feasibility (hypothesis generating), and don't adjust the p-values. But make clear that further confirmatory studies are needed.

Bonferroni adjustment

There are two ways we can perform a Bonferroni adjustment:

1. Reduce the p-value cut off. If you are performing 4 comparisons, your cut off is divided by 4 so, for a result to be counted as statistically significant it needs to be <0.0125 (i.e. 0.05/4).

2. If the raw p-value is <0.05 adjust it by multiplying by the number of comparisons performed.
    - The advantage of this method is that, we are still looking for value below 0.05. As we are so wedded to this cut off, there can be a temptation (when using method 1) for the investigator (or reader) to consider p-values <0.05 but greater than 0.0125 (or whatever reduced limit is being used) to be significant.

Bonferroni adjustment

Endpoint	Raw/unadjusted p-value	Adjusted p-value (simply multiply the raw p-value by 4 if p<0.05, i.e. the number of comparisons)
FEV1	0.001	0.004
FVC	0.03	0.12
No. of exacerbations	0.04	0.16
Time until first exacerbation	0.67	0.67

Bonferroni adjustment

- When adjusting p-values this way, we do need to be wary of borderline values.
- Any value originally over 0.05 should be considered non-significant (i.e. a p-value of 0.06 (>0.05 so not adjusted) should not be considered better evidence of an effect that a p-value of 0.049 (raw) which becomes 0.196 (adjusted).
- Another problem with the Bonferroni adjustment, is that it assumes no relationship between the endpoints, and this is unlikely to be true in most situations.
- You could therefore be inflating each p-value too much, and so could miss a real effect.

Two possible solutions:
1) Provide the unadjusted p-values.

Small ones ($p<0.001$) should be OK, since they should remain small even after allowing for several comparisons

2) Be cautious about $0.05<p<0.01$

- Provide 97.5% confidence intervals for the effects sizes if there are, say 2 or 3 endpoints, and 99% CIs for >3 endpoints
- These give a conservative range of true effect sizes.
- If a 99% CI still does not include the no effect value, then there is likely to be a real effect
- If using 97.5% CI, then the p-value cut-off to determine statistical significance should then be 0.025 (not 0.05 as is usual)
- If using 99% CI, then the p-value cut-off to determine statistical significance should then be 0.01 (not 0.05 as is usual)

Changing the confidence interval: an example

Example: quality of life measured in a trial comparing 2 treatments.
The survey has 25 items so each is an endpoint (5 are shown below), the survey is administered several times per patient over time (i.e. repeated measures)
We could therefore have 25 separate (mixed modelling) results

Endpoint (all are measured on a 0 to 100 scale)	**Effect size (mean difference), 99% CI**	**Unadjusted p-value**	**Possible conclusions**
Global health status	-0.6 (-4.0, +2.7)	0.62	
Nausea & vomiting	-1.9 (-4.4, +0.6)	0.048	
Insomnia	-10.0 (-14.5, -5.5)	<0.0001	
Constipation	+10.6 (+6.2, +15.0)	<0.0001	
Financial difficulties	+3.6 (-1.5, +8.8)	0.07	

Note: For 99% CI, the p-value should be < 0.01 and not < 0.05

Changing the confidence interval: an example

Endpoint (all are measured on a 0 to 100 scale)	Effect size (mean difference), 99% CI	Unadjusted p-value	Possible conclusions
Global health status	-0.6 (-4.0, +2.7)	0.62	
Nausea & vomiting	-1.9 (-4.4, +0.6)	0.048	
Insomnia	-10.0 (-14.5, -5.5)	<0.0001	
Constipation	+10.6 (+6.2, +15.0)	<0.0001	
Financial difficulties	+3.6 (-1.5, +8.8)	0.07	

Note: For 99% CI, the p-value should be < 0.01 and not < 0.05

Changing the confidence interval: an example

Endpoint (all are measured on a 0 to 100 scale)	Effect size (mean difference), 99% CI	Unadjusted p-value	Possible conclusions
Global health status	-0.6 (-4.0, +2.7)	0.62	No evidence of an effect
Nausea & vomiting	-1.9 (-4.4, +0.6)	0.048	Insufficient evidence of an effect, but there might be
Insomnia	-10.0 (-14.5, -5.5)	<0.0001	Evidence of an effect
Constipation	+10.6 (+6.2, +15.0)	<0.0001	Evidence of an effect
Financial difficulties	+3.6 (-1.5, +8.8)	0.07	No evidence of an effect

Note: For 99% CI, the p-value should be < 0.01 and not < 0.05