Module 3B: ANOVA & Correlation

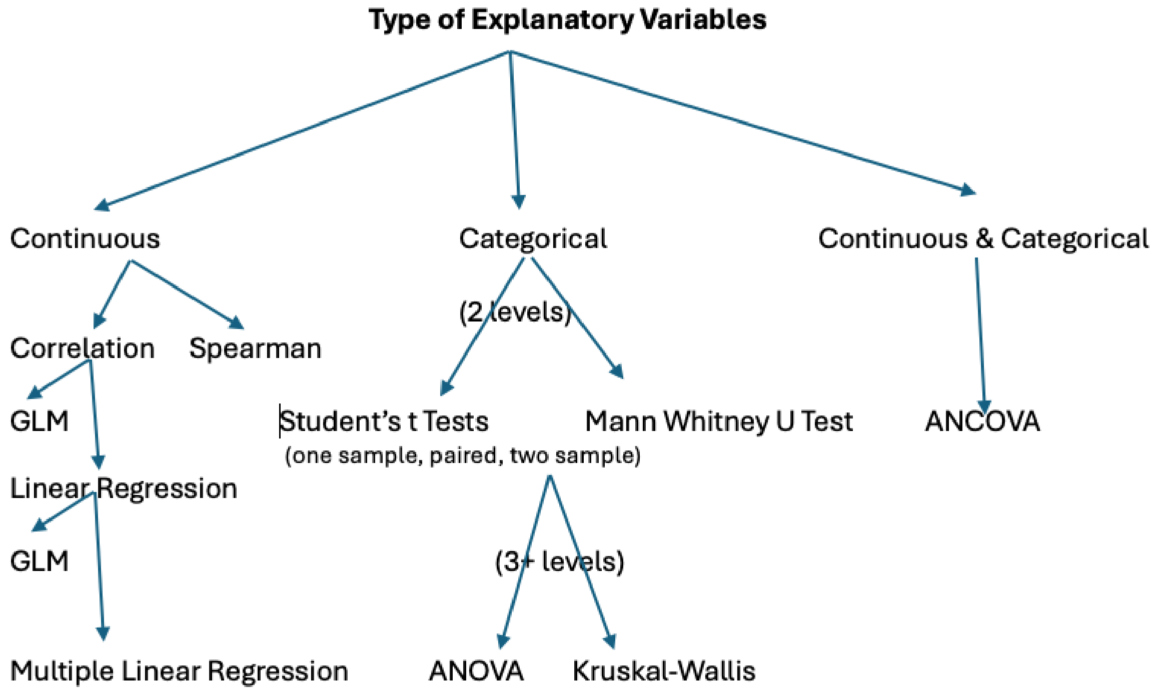
Assigning signal and noise to variation

Agenda:

1. ANOVA: Nuts & Bolts

2. Worked Example
   A. One way ANOVA
   B. Post-hoc tests: Tukey-Kramer
   C. Kruskal-Wallis (nonparametric)

3. Linear Correlation
   A. Spearman's rank



A worked example of ANOVA:

Researchers are investigating the effect of three different diets (A, B, and C) on body weight in genetically modified mice that are prone to obesity. After 8 weeks, the body weights of the mice are measured (in grams). The data is as follows:

Body weights after 8 weeks (grams)

Diet A: 32, 30, 29, 34, 35

Diet B: 40, 42, 43, 45, 41

Diet C: 38, 35, 39, 37, 36

Step 1:

H_0: There is no difference in the mean body weight after 8 weeks among the diets.

H_A: At least one diet group has a different mean body weight.

Step 1:

H_0: There is no difference in the mean body weight after 8 weeks among the diets.

H_A: At least one diet group has a different mean body weight.

Step 2:

GROUP	mean	s	n
A	32	2.55	5
B	42.2	1.92	5
C	37	1.58	5

$N = \sum n = 15$

Mean square error:

$$SS_{error} = \square\, df_i s_i^2$$

df$_{\mathbf{error}}$ = 4+4+4 = 12

$$= 4(2.55)^2 + 4(1.92)^2 + 4(1.58)^2 = 50.74$$

MS$_{\mathbf{error}}$ = 50.74/12 = 4.23

mean squares groups:

$$\bar{X}_G = \frac{5(32)+5(42.2)+5(37)}{15} = 37.07$$

df$_{\mathbf{groups}}$ = k - 1 = 3 -1 = 2

SS$_{\mathbf{groups}}$ = $5(32.0 - 37.07)^2 + 5(42.2 - 37.07)^2 + 5(37.0 - 37.07)^2$

=260.13

MS$_{\mathbf{groups}}$ **=** $SS_{groups} / df_{groups}$ =260.13/2 =130.07

The test statistic for ANOVA is F:

$$F = MS_{groups} / MS_{error}$$
$$= 130.07/4.23$$
$$= 30.25$$

$$F_{0.05(1),2,12} = 3.88$$

Since 30.25 >> 3.88, we know that $P<0.05$ and we can reject H_0.

The variance between the sample group means is bigger than expected given the variance within sample groups so at least one of the groups has a population mean different from another group

Source of variation	Sum of Squares	df	Mean Squares	F-ratio	P
Groups (treatment)	260.13	2	130.07	30.25	**<0.001**
Error	50.80	12	4.23		
Total	310.93	14			

$$R^2 = SS_{groups}/SS_{total} = 260.13/310.93 = 0.84$$

Experimental Design:

*How do we identify **which** means are different and the **magnitude** of their difference?*

1. Planned comparisons:
   - **A priori** comparison between means of groups that were previously identified as particularly interesting
     - **Baked into the study design**
     - **Determined BEFORE data are examined**

   - Only small number allowed so that α isn't inflated

1. Unplanned comparisons:

Experimental Design:

*How do we identify **which** means are different and the magnitude of their difference?*

1. Planned comparisons:
   - **A priori** comparison between means of groups that were previously identified as particularly interesting
   - Only small number allowed so that α isn't inflated
   - If used two-sample t-test instead, your answer would be less precise and would have less power

Experimental Design:

*How do we identify **which** means are different and the magnitude of their difference?*

Example: You run an experiment with **3 diet groups** and measure **12-week weight gain**:

- Group 1: **Chow**
- Group 2: **Low-fat**
- Group 3: **High-fat (HFD)**

You run a **one-way ANOVA** and find a significant overall F-test → at least one group mean differs.

1. Planned comparisons:

- **A priori** comparison between means of groups that were previously identified as particularly interesting
- In this example, *before* you even collected data, your specific scientific hypothesis was:

  "***High-fat diet mice gain more weight than the average of the Chow and Low-fat groups.***"

- That's a **planned comparison** (a contrast you decided *in advance*), and it's *more specific* than "some group is different from some other group."
- Formally, that might be written as a contrast: $H_0: \mu_{HFD} = \frac{\mu_{chow} + \mu_{Low\text{-}Fat}}{2}$
- You then test **that one contrast** (or a small number of pre-specified contrasts). Because they're planned and limited, you:
  - **Don't** usually correct as harshly for multiple comparisons
  - Get **more power** to detect exactly the pattern you care about

Experimental Design:

*How do we identify **which** means are different and the **magnitude** of their difference?*

- Planned comparisons
- Unplanned comparisons:
  - Post hoc
  - Multiple comparisons
  - Determine which means and their magnitude
  - Type **of data dredging** so protect against increasing α
  - Tukey-Kramer procedure tests all pairs of means

2. Unplanned Comparisons(Tukey HSD):

Method:

- Like two-sample t-tests
- Use t distribution
- Different standard error: pooled sample variance (MS_{error}) based on all *k* groups (i.e. using all the information about variance rather than just a subset)

$$SE = \sqrt{MS_{error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

- df of Ms_{error}

Why use MS_{error} instead of a two-sample t-test?

- Increased precision
- Increased power

Assumptions:

- Same as ANOVA but not as robust to violations

What do I mean by inflation of α?

- For a two-sample t test, you are dividing up the variance of only **two** groups into the two samples.

$$\boxed{\boxed{\frac{(n_1-1)s_1^2}{N-k}+\frac{(n_2-1)s_2^2}{N-k}}+\ldots+\frac{(n_k-1)s_k^2}{N-k}}$$

$\mathbf{s_p^2}$ $\mathbf{MS_{error}}$

- For a planned comparison, you are dividing up **ALL** the variance (all the total deviations of the data points) into **only two** of the **k** groups (note: you can do this because H_o assumes variance is same in all groups)

Big idea: this means that you have access to all the degrees of freedom provided by the data points event he ones that are in the groups we are not comparing!

- We saw a different test that also 'absorbed' inflated error by tweaking df (Welch's approximate t test, this reduced df instead of expanding it)

Tukey-Kramer test*:

- Already carried out a single-factor ANOVA and rejected H_0
- Compares all group means to all other group means

$$H_0: \mu_1 = \mu_2$$
$$H_0: \mu_1 = \mu_3$$
$$H_0: \mu_2 = \mu_3$$

* Tukey's Honestly Significant difference (HSD) test

So why not just use a series of two-sample t-tests?

Data Dredging:

When you use multiple tests on a data set, the **actual** probability of making **at least one** type I error, α, is larger than the significance level states

- each hypothesis test has a probability of error and these errors compound as more tests are conducted

- Example: two independent studies are performed to test the same null hypothesis. What is the probability that at least one study obtains a significant result and rejects the null hypothesis **even if the null hypothesis is true**? Assume that in each study there is a **0.05** probability of rejecting the null hypothesis (Answer is **0.0975**)

P(No type I errors)= $(1-\alpha)^N$, *where N = independent tests*

P(≥1 type I error)= $1-(1-\alpha)^N$

Why not use a series of two sample t-tests?

- Multiple comparisons would cause the t-test to reject too many true null hypotheses
- Tukey-Kramer adjusts for the number of tests

Uses larger critical value to limit Type I error

P(≥1 Type I error)= α

- Tukey-Kramer also uses information about the variance within groups from all the data, so it has more power than a t-test with a Bonferroni correction: $\alpha^*=\alpha/\#$ of tests

Tukey-Kramer test:

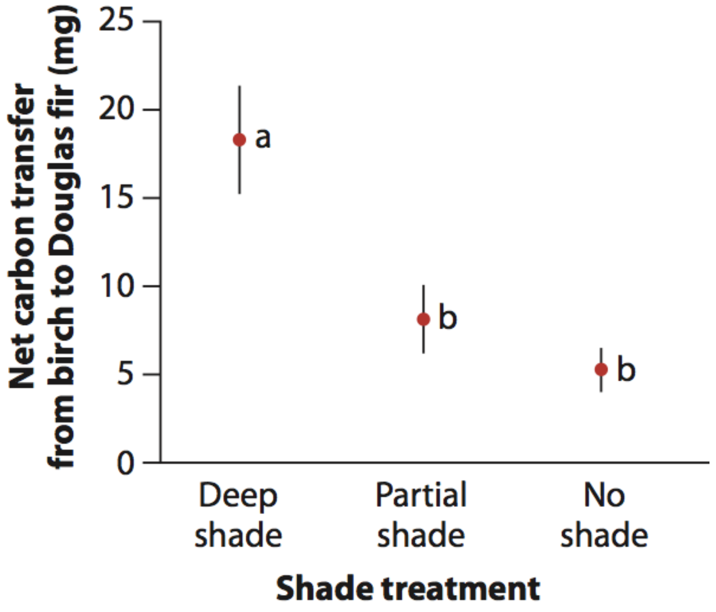
- Uses **q test statistic**
- Method:

1. Order group means from smallest to largest
2. Compare each pair of group means

Ex: First comparison:

$H_0: \mu_1 - \mu_2 = 0$

$H_A: \mu_1 - \mu_2 \neq 0$



3. Calculate **q** test statistic:

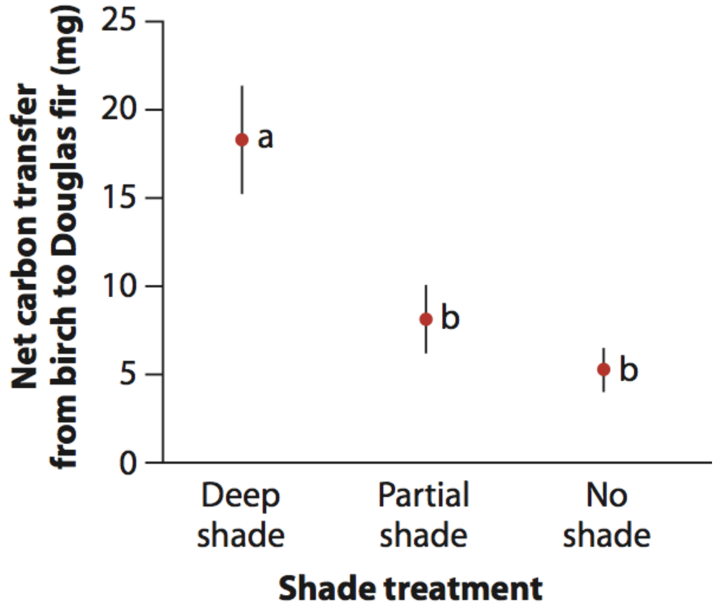
Standard error: MS_{error} df: **k** and **N - k**

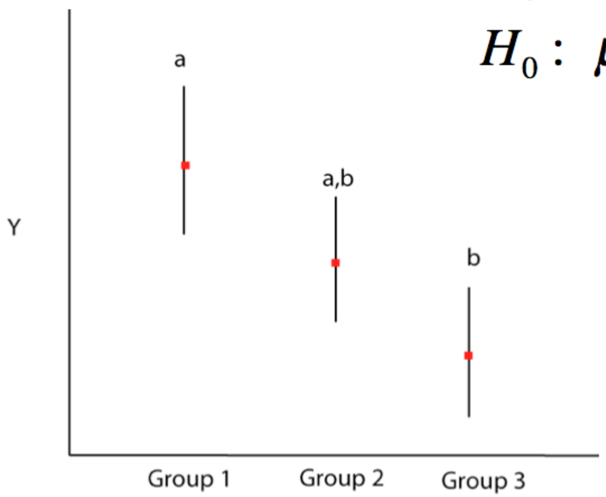
Q-distribution (statistical tables for this online)

- Same assumptions as ANOVA but not as robust
- P value is correct when design is balanced (approximately same number of data points in each category) but it is **conservative** when unbalanced (makes it more difficult to reject the null hypothesis)

https://www.socscistatistics.com/pvalues/qcalculator.aspx#google_vignette

How Tukey-Kramer results are displayed:

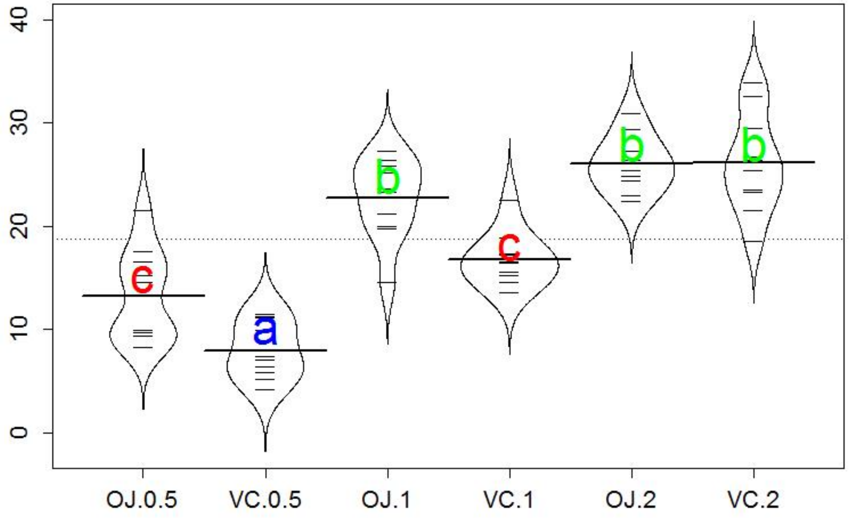




$H_0: \mu_1 = \mu_2$ Cannot reject

$H_0: \mu_1 = \mu_3$ Reject

$H_0: \mu_2 = \mu_3$ Cannot reject



Compact Letter Display

The Tukey test compares the means between each pair of diets (A, B, C) to see which groups differ significantly:

Group 1	Group 2	Mean Difference	P-Value	95% Confidence Interval	Reject Null Hypothesis
Diet A	Diet B	-10.2	<0.001	[-13.51, -6.88]	yes
Diet A	Diet C	5.0	0.007	[1.69, 8.30]	yes
Diet B	Diet C	5.2	0.006	[1.88, 8.51]	Yes

https://www.socscistatistics.com/pvalues/qcalculator.aspx#google_vignette

The Tukey test compares the means between each pair of diets (A, B, C) to see which groups differ significantly:

Group 1	Group 2	Mean Difference	P-Value	95% Confidence Interval	Reject Null Hypothesis
Diet A	Diet B	-10.2	<0.001	[-13.51, -6.88]	yes
Diet A	Diet C	5.0	0.007	[1.69, 8.30]	yes
Diet B	Diet C	5.2	0.006	[1.88, 8.51]	Yes