STAT115: Introduction to Biostatistics

University of Otago
Ōtākou Whakaihu Waka

Lecture 26: ANOVA in Action

Outline

- Fitting ANOVA model
- Understanding ANOVA table
  - Comparing the variance within a group, to the variance between groups
- Look at multiple comparisons
  - Pairwise differences

Recall: hair colour and pain threshold data

- We are looking at pain thresholds may differ with hair colour in students
  - Four colours: light blond, dark blond, light brunette, dark brunette
- Import the data

```
blonds = read.csv('blonds.csv')
```

- Look at the data

Recall: ANOVA

- One-way ANOVA model with K groups
  - Outcome variable in group j is normally distributed with mean μ_j and variance σ^2
- We want to know how the mean outcome differs among groups
  - Potential problems with multiple comparisons
- Are there any differences in mean outcome among the groups?
- This takes the form of a hypothesis test
  - $H_0 : \mu_1 = \mu_2 = \ldots = \mu_K$
  - H_A : at least one mean is different

In R

- As with categorical variables with 2 levels
  - Special case of linear regression
  - Categorical variables can be included in R as factors

```
blonds$HairColour = as.factor(blonds$HairColour)
```

- We can then fit a linear regression model

```
m_blonds = lm(Pain ~ HairColour, data = blonds)
```

- This fits the ANOVA model
- Problem: output from `m_blonds` is not in a convenient form
  - Output is in terms of particular pairwise comparisons

In R

- We use the aov function to get the results in more convenient form

```
a_blonds_lm = aov(m_blonds)
```

- We can also use aov directly

```
a_blonds = aov(Pain ~ HairColour, data = blonds)
```

- The output we will consider is an ANOVA table

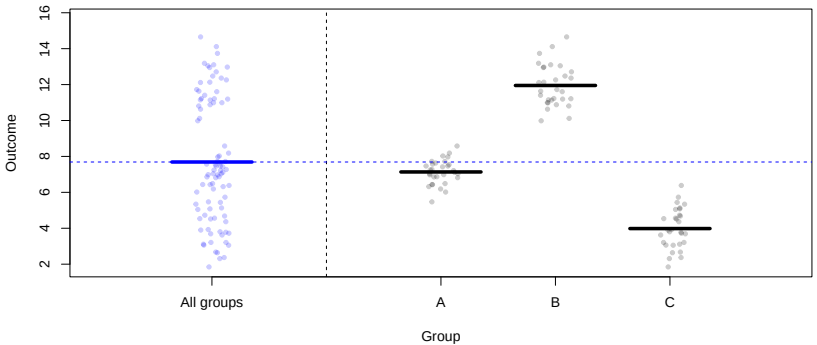
```
summary(a_blonds)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## HairColour   3   1361   453.6   6.791 0.00411 **
## Residuals   15   1002    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

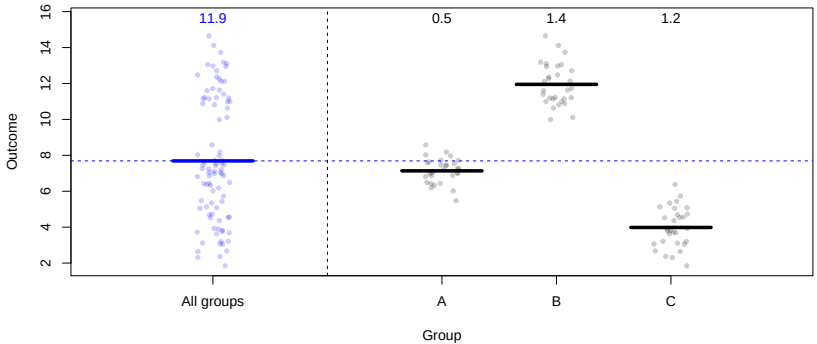
- Take a graphical look at the ANOVA model to help explain what this tells us

Understanding ANOVA (analysis of variance)

- Left plot (blue): plot of all outcome variables (irrespective of group)
- Right three plots (black): plot of outcome variables by group
- Solid horizontal lines: means
  - Dashed blue line is the overall mean

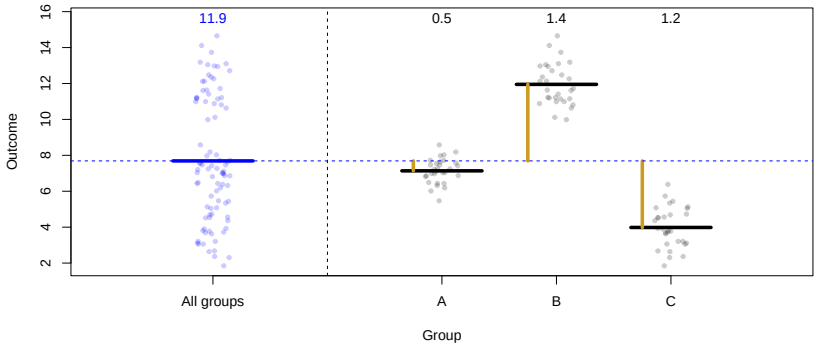


Comparing variance



- The sample variance for each group is given on the plot above
  - Combined data (blue): outcomes are highly variable
  - Data from each group (black; A, B, C): outcomes have much lower variability
- The group variable has explained a lot of the variability in the data

Comparing variance



- Overall variability partitioned into:
  - Variability in group means (indicated by gold lines)
  - Variability within the groups (points around their mean)
- This is the information summarized in the ANOVA table

ANOVA table

- The ANOVA table for the pain threshold data is

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## HairColour   3   1361   453.6   6.791 0.00411 **
## Residuals   15   1002    66.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- To explain what this represents we will use the table:

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	GMS $= \frac{\text{GSS}}{\text{DF}}$	F $= \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	RMS $= \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

ANOVA table: rows

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	GMS $= \frac{\text{GSS}}{\text{DF}}$	F $= \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	RMS $= \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

- Group row: describes the variation between group means
  - Variation represented by gold bar in plot above
- Residuals row: describes the variation within each group
- Total row: describes the variation when we combine across groups
  - Data represented in blue in plot above
  - This row is not in R output

ANOVA table: columns

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	GMS $= \frac{\text{GSS}}{\text{DF}}$	F $= \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	RMS $= \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

- Mean Sq[uare]
  - Group (GMS): related to the between-group variance
  - Residual (RMS): estimate of within-group variance
- F value: ratio of group mean square and residual mean square
- Df: degrees of freedom
- Sum Sq: sum of squares
  - Convenient when calculating by hand

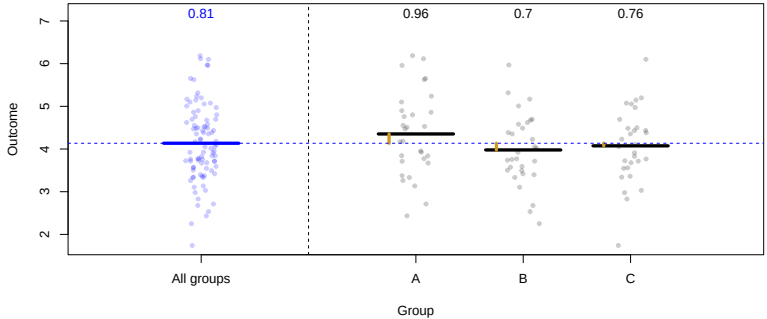
ANOVA table

Source	Df	Sum Sq	Mean Sq	F value
Group	$K-1$	GSS	GMS $= \frac{\text{GSS}}{\text{DF}}$	F $= \frac{\text{GMS}}{\text{RMS}}$
Residuals	$n-K$	RSS	RMS $= \frac{\text{RSS}}{\text{DF}}$	
Total	$n-1$	TSS		

- If the groups explain a lot of variability (like our plots above)
  - The group mean square will be large relative to residual mean square
  - F-value will be relatively large
    - ANOVA table below is for data from plots above

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2  966.2   483.1   472.5 <2e-16 ***
## Residuals   87   88.9     1.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example II: group does not explain much variation



- The group mean square will not be large relative to residual mean square
- The F-value is not large

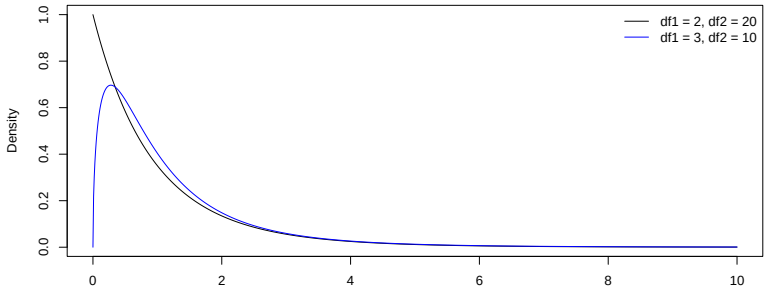
```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2   2.27  1.1354    1.41   0.25
## Residuals   87  70.08  0.8055
```

ANOVA table: F column

- The F-value is comparing the variance among groups (the variability in the group means) to the variance within the groups
  - It is a measure of how much variation in the data is explained by the groups compared to unexplained variation
- If the null hypothesis is true
  - Data come from the ANOVA model with all means equal ($\mu_1 = \mu_2 = \ldots = \mu_k$)
    - The data are normally distributed with the same mean and variance
  - F-statistic will have an F-distribution with Df (group), Df (residual) degrees of freedom
- We can use this to find a p-value
  - Quantify the incompatibility between the data and null hypothesis
  - Are the data unusual given that the null hypothesis is true (group means are the same)
- If null hypothesis is true, we expect an F-value of around 1

Detour: F-distribution

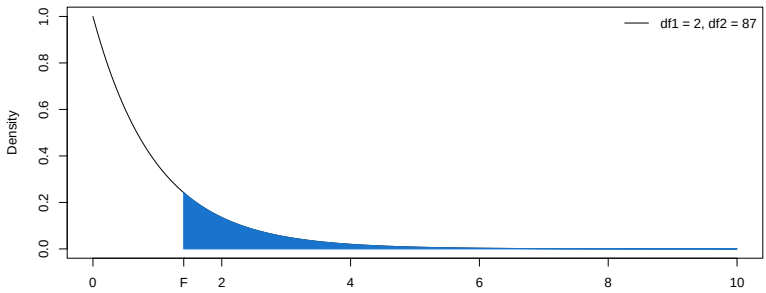
- The F-distribution is a distribution for positive random variables



  - It is asymmetric (positively skewed)
  - It has two parameters:
    - Degrees of freedom for the numerator (df1)
    - Degrees of freedom for the denominator (df2)

Finding a p-value

- An extreme F-value is as large, or larger, than that observed
  - Indicative of groups explaining as much, or more, variation in the data



- The blue area is given by `1-pf(F, df1, df2)`
  - `pf(F, df1, df2)` gives probability of a value less than F

Example II

- The ANOVA table for example II is

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2   2.27  1.1354    1.41   0.25
## Residuals   87  70.08  0.8055
```

- The observed F-statistic is 1.41
  - df1 is degrees of freedom for group: 2
  - df2 is degrees of freedom for residuals: 87
- The p-value is

```
1-pf(1.41, 2, 87)
```

```
## [1] 0.25
```

- In practice: refer to the `Pr(>F)` column in the output

In R: Pain Sensitivity Data

- The ANOVA table for the pain sensitivty and hair colour data is

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## HairColour   3   1361     454    6.79 0.0041 **
## Residuals   15   1002      67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The F-value is large, p-value is small
  - p-value $< \alpha$: evidence of incompatibility between data and null hypothesis
  - Data are (highly) unusual if all the means were truly the same
  - Providing evidence that at least one of the means differ
- Which groups have means that appear to differ?

Pairwise comparisons of group means

- To compare each group, there are (potentially) many comparisons
  - If we have $K = 3$ groups: 3 comparisons
  - If we have $K = 5$ groups: 10 comparisons
  - If we have $K = 10$ groups: 45 comparisons
- E.g. for $K = 3$: conduct hypothesis tests or find confidence intervals:
  - CI for $\mu_1 - \mu_2$; hypothesis test with $H_0 : \mu_1 - \mu_2 = 0$
  - CI for $\mu_1 - \mu_3$; hypothesis test with $H_0 : \mu_1 - \mu_3 = 0$
  - CI for $\mu_2 - \mu_3$; hypothesis test with $H_0 : \mu_2 - \mu_3 = 0$

Multiple comparisons

- The problem with multiple tests (or multiple confidence intervals) is that properties no longer hold. For hypothesis testing:
  - α gives the type I error rate for a single test
    - Probability of α of a 'false positive' given that the null hypothesis is true
  - In each test, there is a chance of a false positive (type I error)
  - With multiple tests, the overall chance of a type I error increases
  - Overall type I error rate: referred to as the family-wise error rate
    - Probability of making at least one type I error when performing multiple tests
  - Multiple comparisons increase the family wise error rate
    - e.g. if we perform 10 independent tests with $\alpha = 0.05$, then the probability of at least one type I error is $1 - 0.95^{10} = 0.4$, if the null hypothesis is true in each instance
    - Probability found using complements

Tukey HSD

- Tukey's honest significant difference (HSD) is a multiple comparison approach designed for ANOVA models
- If the sample sizes are the same in each group
  - Family-wise error rate is exactly α
- If the sample sizes are different among groups
  - It is conservative (family-wise error rate is less than α)
- The Tukey approach finds corrected confidence intervals and p-values
- It is easily implemented in R: `TukeyHSD`

In R: Pain Sensitity Data

```
TukeyHSD(a_blonds)

##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = Pain ~ HairColour, data = blonds)
##
## $HairColour
##                              diff   lwr  upr p adj
## DarkBrunette-DarkBlond     -13.8 -28.7  1.1 0.074
## LightBlond-DarkBlond         8.0  -6.9 22.9 0.436
## LightBrunette-DarkBlond     -8.7 -24.5  7.1 0.415
## LightBlond-DarkBrunette     21.8   6.9 36.7 0.004
## LightBrunette-DarkBrunette   5.1 -10.7 20.9 0.789
## LightBrunette-LightBlond   -16.7 -32.5 -0.9 0.037
```

Interpretation: Pain Sensitivity Data

- Intepret the adjusted confidence intervals, e.g.
  - We are 95% confident that the difference in mean pain threshold between light blonds and dark brunettes is between 6.9 and 36.7
- Interpret the adjusted p-values, e.g.
  - The p-value for the difference between dark brunette and dark blond mean pain thresholds is 0.074.
  - As p-value $> \alpha$ there is no evidence that the observed difference is unusual given the null hypothesis that the two means are the same
  - Note: the uncorrected p-value is 0.017

ANOVA: big picture

- We have looked at fitting one-way a ANOVA model
  - One-way refers to one categorical predictors: HairColour (for pain sensitivity example)
  - Two-way ANOVA: have two categorical predictors
- There might be many other potential predictors (categorical or continuous)
  - e.g. age, ethnicity, sex, etc.
- Recall: ANOVA is a special case of linear regression
  - We can use multiple linear regression to include these other variables
- There are lots of possible extensions
- There are also lots of ways to get ourselves into trouble
- These more complex models are explored in STAT 210

Summary

- Looked at the ANOVA summary table
  - Group: the variation between group means
  - Residuals: the variation within a group
  - F-value: comparing the variance within a group, to the variance between groups
- F-distribution to find p-value
- Look at multiple comparisons for pairwise differences
  - Tukey's honest significant difference
  - See multiple comparisons in general context later in the course