Oneway ANOVA

Recap

Model comparison approach

- Restricted/Full is the embodiment of the Null/Alternative
- Why are we doing this?

Why are we doing this?

- The right balance between simplicity and complexity — parsimony
- Full model will have a smaller (or equal) error than Restricted model, but it will also have a smaller degrees of freedom.
- We are essentially asking whether the increase in explained variance is "worth" the loss of degrees of freedom.

This time

Translating that into classic ANOVA terminology

Starting with Oneway ANOVA

Oneway ANOVA

Sometimes we have more than two groups. t-tests won't cut it.

CDR - Clinical Dementia Rating

Assesses performance in 6 areas: memory, orientation, judgement & problem solving, community affairs, home & hobbies, and personal care.

Composite Rating

- 0 = none
- .5 = very mild dementia
- 1 = mild dementia
- 2 = moderate dementia
- 3 = severe dementia

CDR - Clinical Dementia Rating

Assesses performance in 6 areas: memory, orientation, judgement & problem solving, community affairs, home & hobbies, and personal care.

Composite Rating

- **0 = none**
- **.5 = very mild dementia**
- **1 = mild dementia**
- 2 = moderate dementia
- 3 = severe dementia

We want to know if participants' anxiety scores are different for each of these groups.

- Null hypothesis?
- Alternative hypothesis?

The data

Participant Anxiety Scores (20-80) for each of the groups:

NoImpairment	VeryMildImpairment	MildImpairment
30	35	45
35	40	55
30	45	55
25	45	50
40	55	50

The data in long format

```
cdrLong = cdr %>%
  pivot_longer(cols = 1:3,
               names_to = "Imp
               values_to = "An
  mutate_at(vars(1), list(fact
  arrange(ImpairmentType)
```

ImpairmentType	AnxietyScores
MildImpairment	45
MildImpairment	55
MildImpairment	55
MildImpairment	50
MildImpairment	50
NoImpairment	30
NoImpairment	35
NoImpairment	30
NoImpairment	25
NoImpairment	40
VeryMildImpairment	35
VeryMildImpairment	40
VeryMildImpairment	45
VeryMildImpairment	45
VeryMildImpairment	55

Means

```r
grandMean = mean(cdrLong$AnxietyScores)
grandMean
```

```
## [1] 42.33333
```

```r
meansCdr =  cdrLong %>%
  group_by(ImpairmentType) %>%
  summarize(meanAnxiety = mean(AnxietyScores))
meansCdr
```

```
## # A tibble: 3 × 2
##   ImpairmentType     meanAnxiety
##   <fct>                    <dbl>
## 1 MildImpairment              51
## 2 NoImpairment                32
## 3 VeryMildImpairment          44
```

```r
## create a new column repeating the correct means

cdrLong$grandMean = rep(grandMean, times = nrow(cdrLong))
cdrLong$groupMeans = c(rep(meansCdr$meanAnxiety[1], times = 5),
                       rep(meansCdr$meanAnxiety[2], times = 5)
```

ImpairmentType	AnxietyScores	grandMean	groupMeans
MildImpairment	45	42.33333	51
MildImpairment	55	42.33333	51
MildImpairment	55	42.33333	51
MildImpairment	50	42.33333	51
MildImpairment	50	42.33333	51
NoImpairment	30	42.33333	32
NoImpairment	35	42.33333	32
NoImpairment	30	42.33333	32
NoImpairment	25	42.33333	32
NoImpairment	40	42.33333	32
VeryMildImpairment	35	42.33333	44
VeryMildImpairment	40	42.33333	44
VeryMildImpairment	45	42.33333	44
VeryMildImpairment	45	42.33333	44
VeryMildImpairment	55	42.33333	44

Restricted Model

ImpairmentType	AnxietyScores	grandMean	groupMeans	restrictedDev	restrictedDev2
MildImpairment	45	42.33333	51	2.666667	7.111111
MildImpairment	55	42.33333	51	12.666667	160.444444
MildImpairment	55	42.33333	51	12.666667	160.444444
MildImpairment	50	42.33333	51	7.666667	58.777778
MildImpairment	50	42.33333	51	7.666667	58.777778
NoImpairment	30	42.33333	32	-12.333333	152.111111
NoImpairment	35	42.33333	32	-7.333333	53.777778
NoImpairment	30	42.33333	32	-12.333333	152.111111
NoImpairment	25	42.33333	32	-17.333333	300.444444
NoImpairment	40	42.33333	32	-2.333333	5.444444
VeryMildImpairment	35	42.33333	44	-7.333333	53.777778
VeryMildImpairment	40	42.33333	44	-2.333333	5.444444
VeryMildImpairment	45	42.33333	44	2.666667	7.111111
VeryMildImpairment	45	42.33333	44	2.666667	7.111111
VeryMildImpairment	55	42.33333	44	12.666667	160.444444

```
## [1] 1343.333
```

Restricted Model

Degrees of freedom = $n - 1$

We had to estimate the overall grand mean

```
dfr = nrow(cdrLong) - 1
dfr
```

```
## [1] 14
```

ImpairmentType	AnxietyScores	grandMean	groupMeans
MildImpairment	45	42.33333	51
MildImpairment	55	42.33333	51
MildImpairment	55	42.33333	51
MildImpairment	50	42.33333	51
MildImpairment	50	42.33333	51
NoImpairment	30	42.33333	32
NoImpairment	35	42.33333	32
NoImpairment	30	42.33333	32
NoImpairment	25	42.33333	32
NoImpairment	40	42.33333	32
VeryMildImpairment	35	42.33333	44
VeryMildImpairment	40	42.33333	44
VeryMildImpairment	45	42.33333	44
VeryMildImpairment	45	42.33333	44
VeryMildImpairment	55	42.33333	44

Full Model

ImpairmentType	AnxietyScores	grandMean	groupMeans	fullDev	fullDev2
MildImpairment	45	42.33333	51	-6	36
MildImpairment	55	42.33333	51	4	16
MildImpairment	55	42.33333	51	4	16
MildImpairment	50	42.33333	51	-1	1
MildImpairment	50	42.33333	51	-1	1
NoImpairment	30	42.33333	32	-2	4
NoImpairment	35	42.33333	32	3	9
NoImpairment	30	42.33333	32	-2	4
NoImpairment	25	42.33333	32	-7	49
NoImpairment	40	42.33333	32	8	64
VeryMildImpairment	35	42.33333	44	-9	81
VeryMildImpairment	40	42.33333	44	-4	16
VeryMildImpairment	45	42.33333	44	1	1
VeryMildImpairment	45	42.33333	44	1	1
VeryMildImpairment	55	42.33333	44	11	121

```
## [1] 420
```

Full Model

Degrees of freedom = $n - 3$

We had to estimate 3 things; each of the group means

```
dff = nrow(cdrLong) - 3
dff
```

```
## [1] 12
```

Wrapping it up!

$$F = \frac{(1343.333 - 420)/(14 - 12)}{420/12} \quad F = \frac{461.667}{35} \quad F = 13.190$$

Critical value is 3.885

Tricks are for kids

- E_R = squared deviations from the grand mean
- E_F = squared deviations from the group means
- If equal n,
  $E_R - E_F = n \times \Sigma(\text{Grand Mean - Group Mean})^2$

Tricks are for kids

$$E_R - E_F = n \times \Sigma(\text{Grand Mean - Group Mean})^2$$

$$= 5 \times ((32 - 42.333)^2 + (44 - 42.333)^2 + (51 - 42.333)^2)$$

$$E_R - E_F = 5 \times (106.778 + 2.778 + 75.111)$$

$$E_R - E_F = 923.333$$

```
Er - Ef
```

```
## [1] 923.3333
```

ANOVA Formula

$$F = \frac{(E_R - E_F)/(df_R - df_F)}{E_F/df_F} \quad F = \frac{SS_B/df_{\text{numerator}}}{SS_W/df_{\text{denominator}}}$$

$$F = \frac{923.333/2}{420/12}$$

Er - Ef

- Sum of square deviations from the grand mean *minus* sum of squared deviations from the group mean
- $E_R - E_F = n \times \Sigma(\text{Grand Mean - Group Mean})^2$
- Sum of squared deviations ***between*** groups
- ***How different each group is from other groups?***

Ef

- Sum of squared deviations from the group mean
- Errors reflect how much an individual deviates from the group
- Sum of squared deviations ***within*** groups
- ***How different each individual is from their own group?***

ANOVA Formula

$$F = \frac{SS_B/df_{\text{numerator}}}{SS_W/df_{\text{denominator}}}$$

$$F = \frac{MS_B}{MS_W}$$

ANOVA Table or Source Table

```
summary(aov(AnxietyScores ~ ImpairmentType, data = cdrLong))
```

```
##                Df Sum Sq Mean Sq F value   Pr(>F)
## ImpairmentType  2  923.3   461.7   13.19 0.000934 ***
## Residuals      12  420.0    35.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
2 + 12
```

```
## [1] 14
```

```
923.333 + 420
```

```
## [1] 1343.333
```

Why is it called the Analysis of Variance if we are interested in the means?

$$\sum_i \sum_j (Y_{ij} - \bar{\bar{Y}}_{..})^2 = (\bar{\bar{Y}}_{.j} - \bar{\bar{Y}}_{..})^2 + (Y_{ij} - \bar{\bar{Y}}_{.j})^2$$

$$SS_{\text{Total}} = SS_{\text{Between (Method)}} + SS_{\text{Within (Residual)}}$$

$$s^2_{\text{Total}} = s^2_{\text{Explained}} + s^2_{\text{Unexplained}}$$

The ratio of SS between (signal) with the SS within (noise)

Analysis of "Variance"

Total variability associated with the outcome variable (SStot) can be divided into “the variation due to the differences in the sample means for the different groups” (SSb) plus “all the rest of the variation” (SSw).

If the null hypothesis is true, then what would you expect? A larger SSb or a larger SSw?

The sample means should be pretty similar to each other, right? And that would mean SSb to be really small, in comparison to the “the variation associated with everything else”, SSw.

What does it mean if I have a large SSb but also a large SSw?

Note that the cut-off point will depend on the dfs.

Eta Squared

$$\eta^2 = \frac{SS_B}{SS_{\text{Total}}}$$

Interpretation: Proportion of the variability in the outcome variable that can be explained in terms of the predictor.

- $\eta^2 = 0$; there is no relationship at all between the outcome and the predictor
- $\eta^2 = 1$; the relationship between the outcome and predictor is perfect

If you take the square root of η^2, it can be interpreted as if it referred to the magnitude of a Pearson correlation.

see textbook page 440

Proportionate Reduction in Error (PRE)

To what extent does the full model reduce the errors made?

$$PRE = \frac{E_R - E_F}{E_R}$$

- Scale of 0 to 1
- If it's 0, then knowing X does not help predict Y. The full model does not reduce the errors made.
- If it's 1, then knowing X 100% predicts Y. The full model completely reduces the errors made.

Assumptions, assumptions

The following should hold in order to assume that your observed F maps on to the population distribution for F:

1. Normally distributed scores
2. Equal variances (homogeneity)
3. Scores should be independent

What happens when your n is unequal? Or your variances are not equal?

Assumptions, assumptions

If you have unequal n, we need to use a slightly different formula for E_F

- E_F reflects deviations from the group mean
- $\Sigma(\text{SS X sample size for the group})$
- Weight the sum of squares by the group size

When is the *F* test robust?

	Equal Sample Sizes	**Unequal Sample Sizes**
Equal Variances	Appropriate	Appropriate
Unequal Variances	Good, unless there's a very large difference	???

When is the *F* test robust?

- If you have large samples with very large variances, considered a *conservative* test

  A conservative test tends to reduce the chance of a Type I error (rejection of a true null hypothesis) at the cost of increasing the risk of a Type II error.

- If you have large samples with very small variances, considered a *liberal* test

  A liberal test tends to increase the chance of a Type I error but reduces the chance of a Type II error.

- Large variances make it harder to detect true differences (more conservative). Small variances make it easier to detect differences, including those that might not be meaningful (more liberal).

You try!

- Complete the first question in HW 4

Next time

- Multiple comparisons & contrasts