PS303: Week 11

pp.497-508;512-520

Recap

NHSTs for estimating differences between continuous parameters

- 1-sample t-test: Is the sample mean (M) different relative to a population mean (μ), or $H_0 : M = \mu$?
- 2-sample independent t-test: Is the difference between two sample means different from a null estimate, or $H_0 : M_1 - M_2 = 0$ (if a two-sided test is being run)?
- 2-sample pairwise t-test: Does the same sample measured at different times produce different results, or $H_0 : Time_1 - Time_2 = 0$?
- One-way/independent ANOVA: Are the means of $k \geq 3$ *independent* groups statistically equivalent, or $H_0 : M_1 = M_2 = M_3$?
- Linear regression: Do predictors (X_i) account for outcome (Y) variance, or $H_0 : \hat{Y}_i = b_0 + \epsilon_i$? Do individual coefficients predict for outcome variance, or $H_o : b_0 = 0$?

> As our analysis of regression models demonstrate, there may be multiple predictors that contribute towards observed variance. When we have *multiple* predictors, we can run factorial ANOVAs to test whether multiple independent variables singly or interactively explain outcome variances.

Balanced factorial ANOVA: Main effects

By *factorial*, we imply ANOVAs with more than 1 independent variable (factor)

By *balanced*, we imply all levels of our design contain equal numbers of participants.

A *balanced factorial* ANOVA involves looking at multiple independent variables that have equal numbers of observations for each level.

Suppose we want to know whether $n = 12$ participants with *low and high* levels of depression (Factor 1) from *Fiji and Singapore* (Factor 2) drink different quantities of alcohol and we have three participants' data from each country and location:

ID	Location	Depression	Weekly alcohol consumption (ml)
1	Fiji$_1$	Low$_1$	311
2	Fiji$_2$	Low$_2$	320
3	Fiji$_3$	Low$_3$	413
4	Singapore$_1$	Low$_4$	343

ID	Location	Depression	Weekly alcohol consumption (ml)
5	Singapore$_2$	Low$_5$	341
6	Singapore$_3$	Low$_6$	380
7	Fiji$_1$	High$_1$	375
8	Fiji$_2$	High$_2$	420
9	Fiji$_3$	High$_3$	412
10	Singapore$_1$	High$_4$	357
11	Singapore$_2$	High$_5$	519
12	Singapore$_3$	High$_6$	448

We can summarize the mean alcohol consumed for each factor combination (Location $\times$ Depression), which constitutes a 2×2 design.

	Fiji$_{Column1}$	**Singapore**$_{Column2}$	**Total row (R) means**
Low depression $_{Row1}$	348	354.67	$\frac{\sum Row_1}{N_{Rows}} = \mu_{Row1} = 351.34$
High depression $_{Row2}$	402.33	441.33	$\frac{\sum Row_2}{N_{Rows}} = \mu_{Row2} = 421.83$
Total column (C) means	$\frac{\sum Column_1}{N_{Cols}} = \mu_{Column1} = 375.16$	$\frac{\sum Column_2}{N_{Cols}} = \mu_{Column1} = 398$	$R \times C = 351.34$

> Rows ($R = 2$) and columns ($C = 2$) refer to the different factors we are interested in. Row and column averages (marginal means) refer to summary statistics of each factorial level. The grand average ($R \times C$) is the average of all marginal means across our data.

If we are interested in estimating whether factors can *individually* predict amount of alcohol drunk, we can declare the following null hypotheses:

- $H_0 1$: There is no difference in alcohol drunk between low and high depressed groups ($\mu_{Column1} = \mu_{Column2}$)
- $H_0 2$: There is no difference in alcohol drunk between Fijians and Singaporeans ($\mu_{Row1} = \mu_{Row2}$)

Remember that the row (R) and column (C) marginal estimates that refer to levels of factor. Across both H_0's, the claim being tested is whether the marginal means associated with each factor are statistically equivalent ($p \geq .05$).

Setting up the data

```
ID         <- seq(1:12)                                             # 12 participants
Location   <- rep(c(rep("Fiji",3),rep("Singapore",3)),2)            # Locations levels
Depression <- c(rep("Low",6),rep("High",6))                         # Depression levels
Alcohol    <- c(311,320,413,343,341,380,375,420,412,357,519,448)     # Alcohol drunk
df <- cbind.data.frame(ID,Location,Depression,Alcohol)               # Combine into data frame

# Convert non-Alcohol variables into factors
df$ID          <- as.factor(df$ID)
df$Location    <- as.factor(df$Location)
df$Depression  <- as.factor(df$Depression)

# Print the data frame (named 'df')
df
```

```
##    ID  Location Depression Alcohol
## 1   1      Fiji        Low     311
## 2   2      Fiji        Low     320
## 3   3      Fiji        Low     413
## 4   4 Singapore        Low     343
## 5   5 Singapore        Low     341
## 6   6 Singapore        Low     380
## 7   7      Fiji       High     375
## 8   8      Fiji       High     420
## 9   9      Fiji       High     412
## 10 10 Singapore       High     357
## 11 11 Singapore       High     519
## 12 12 Singapore       High     448
```

Let's run ANOVAs for each factor (similar to running 2 one-way ANOVAs) and assign the results into `mod1` and `mod2`

```
mod1 <- aov(formula=Alcohol~Depression,data=df)  # Does depression predict alcohol consumed?
mod2 <- aov(formula=Alcohol~Location,data=df)    # Does depression predict alcohol consumed?
```

We can extract the F-ratio,sum of squares and p-values for each model by using the `summary()` function

```
summary(mod1)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Depression   1  14911   14911   6.204 0.0319 *
## Residuals   10  24032    2403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We found a main effect for depression, $F_{1,10} = 6.204, p = .032$, towards predicting alcohol consumption.

```
summary(mod2)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Location     1   1564    1564   0.418  0.532
## Residuals   10  37379    3738
```

We did not find a main effect found for location, $F_{1,10} = .418, p = .532$, towards predicting alcohol consumption.

The F-ratio is the mean sum of squares of the factor variance

$$MS_{Factor} = \frac{SS_{Factor}}{df_{Factor}}$$

divided by the mean sum of squared residuals

$$MS_{Residuals} = \frac{SS_{Resiuduals}}{df_{Resiuduals}}$$

which gives us

$$F = \frac{MS_{Factor}}{MS_{Residuals}}$$

The p-value tells us how likely is the F-ratio to be observed assuming the null hypothesis (there is no relationship) is true.

A key difference between running multiple one-way ANOVAs and a factorial ANOVA has to do with how the residuals (difference between predicted and observed estimates) are calculated. Note that the present data included $k > 1$ predictors, where *each* predictor would be

> associated with a degree of outcome variability. During a factorial ANOVA, the residuals associated with each predictor are taken into account when estimating main effects. In contrast,a one-way ANOVA takes a single predictor into consideration exclusively, meaning variances across *all* predictors are attributed to a *single* predictor, which would render the data more ‘noisy’, In practical terms, a one-way ANOVA is less likely to detect a significant effect (have greater Type-2 error) relative to a factorial ANOVA, even when both models involve a single predictor.

Compared to a one-way ANOVA, a two-way factorial ANOVA can provide us with four possible outcomes:

1. Only Factor A matters
2. Only Factor B matters
3. Neither Factors matter
4. Both Factors matter (there is an *interaction* between A & B)

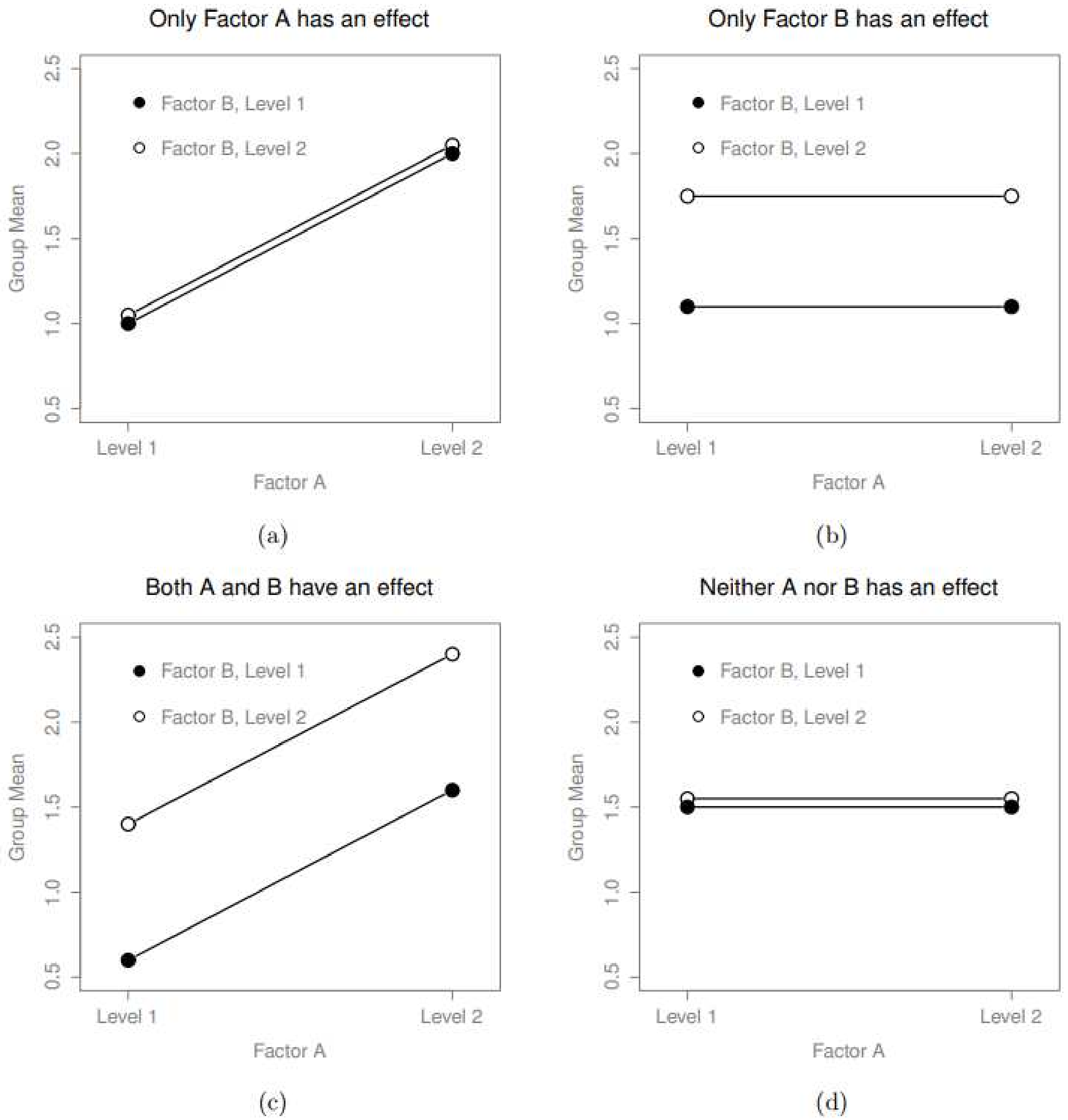
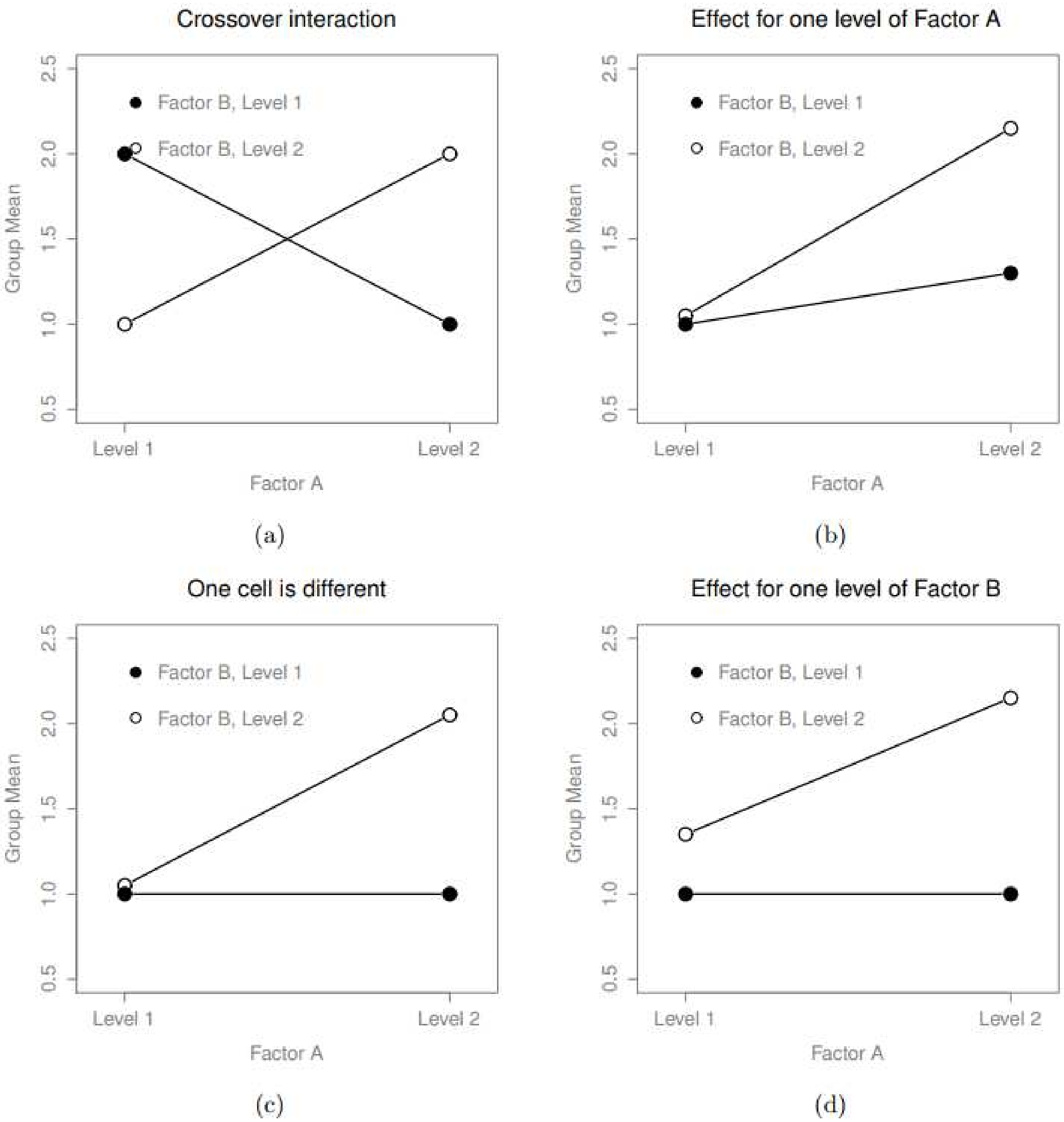


Figure 16.1: The four different outcomes for a 2 × 2 ANOVA when no interactions are present. In panel (a) we see a main effect of Factor A, and no effect of Factor B. Panel (b) shows a main effect of Factor B but no effect of Factor A. Panel (c) shows main effects of both Factor A and Factor B. Finally, panel (d) shows no effect of either factor.

Balanced factorial ANOVA: Main effects + Interactions

An *interaction* between factors implies levels of *one factor* vary with levels of the *other factor*



(a) (b) (c) (d)

Some varieties of interactions

To explore for interactions across multiple predictors, we can either specify each of three terms (2 for main effects, 1 for the interaction) within the ANOVA model

```
mod2 <- aov(Alcohol~Depression+Location+Depression:Location)
mod2
```

```
## Call:
##    aov(formula = Alcohol ~ Depression + Location + Depression:Location)
##
## Terms:
##                 Depression  Location Depression:Location Residuals
## Sum of Squares   14910.750  1564.083             784.083 21684.000
## Deg. of Freedom          1         1                   1         8
##
## Residual standard error: 52.06246
## Estimated effects may be unbalanced
```

Which is the same as

```
mod3 <- aov(Alcohol~Depression*Location)
mod3
```

```
## Call:
##    aov(formula = Alcohol ~ Depression * Location)
##
## Terms:
##                 Depression  Location Depression:Location Residuals
## Sum of Squares   14910.750  1564.083             784.083 21684.000
## Deg. of Freedom          1         1                   1         8
##
## Residual standard error: 52.06246
## Estimated effects may be unbalanced
```

Applying the `summary()` function reveals whether either factor predicted variances in alcohol consumption (main effects), and whether the two factors *interacted* to influence alcohol consumption

```
summary(mod3)
```

```
##                     Df Sum Sq Mean Sq F value Pr(>F)
## Depression           1  14911   14911   5.501  0.047 *
## Location             1   1564    1564   0.577  0.469
## Depression:Location  1    784     784   0.289  0.605
## Residuals            8  21684    2710
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can manually compute the effect size for individual terms

$$\eta_p^2 = \frac{SS_{Factor}}{SS_{Factor} + SS_{Residuals}} = \frac{14911}{14911 + 21684} = .407$$

.

We can report our findings as as follows

> A 2×2 ANOVA with depression levels and location as independent factors did not interact to predict variance in alcohol consumption ($p = .605$). A reliable main effect was found for depression only, $F_{1,8} = 5.501, p = .047, \eta_p^2 = .41$.

We would be now justified in running post-hoc tests across our significant factor using two-sample t-tests. However, because there are only two levels of the depression factor, this is not necessary presently as a main effect *here* would be equivalent to a two-sample test.

Assumptions for running an ANOVA

Like all NHSTs run earlier, a factorial ANOVA requires the data to meet certain assumptions in order to generate a minimally biased estimate.

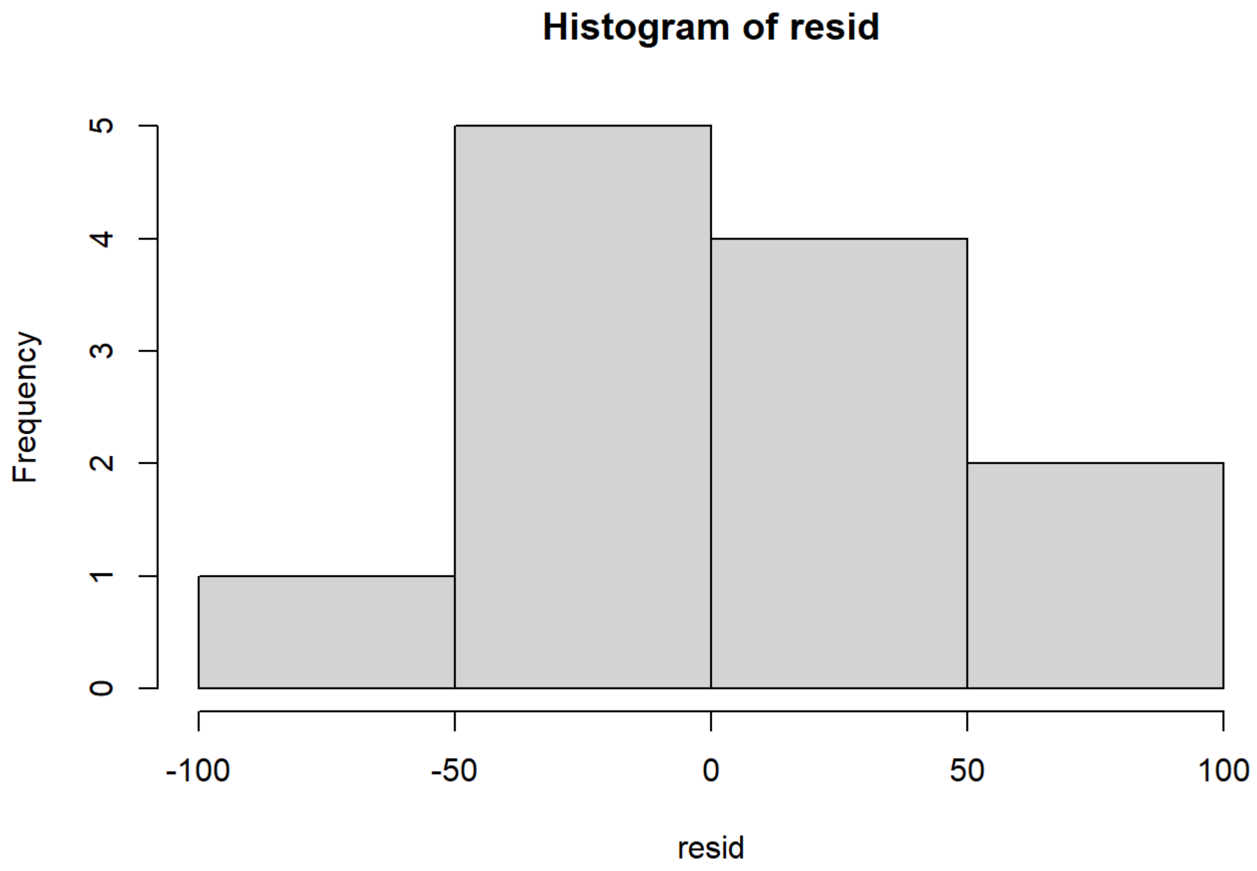
- *Homogeneity of variance*: Do the groups have statistically equivalent variance? We can run the Levene test on our full (**saturated**) model *viz.* with both the main effects and interactions specified. If the null hypothesis is *supported* ($p \geq .05$), we can assume data across all levels of a factor have statistically equivalent variances

```
require(rstatix)
levene_test(mod3)
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     3     8     0.769 0.543
```

- *Residual normality*: Are the residuals of the model normally distributed?

```
resid <- residuals(mod3)   # Extract the residuals
hist(resid)                # Draw a histogram
```



This looks approximately normal. We can also run a `shapiro test for normality` to quantify our results

```
shapiro.test(resid)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid
## W = 0.96931, p-value = 0.9035
```

The test supports the null hypothesis, meaning the assumption for normality has not been violated

Comparing models

Similar to our strategy with regression models, we may want to contrast *across* models to select the 'better' model. We can use the F-ratio to contrast between models.

$Model - 1$: *Alcohol~Depression*

$Model - 2$: *Alcohol~Depression + Location + Depression:Location*

We can estimate the sum of squares (SS) for each model by subtracting the residual variability from the total outcome variability

$$SS_{Model-1} = SS_{Total} - SS_{Residuals-1}$$

$$SS_{Model-2} = SS_{Total} - SS_{Residuals-2}$$

We can estimate the *difference* (Δ) between the sum of squares for the two models and the degrees of freedom

$$SS_{\Delta} = SS_{Model-2} - SS_{Model-1}$$

$$df_{\Delta} = df_{Model-2} - df_{Model-1}$$

Now we can estimate the mean square for the difference between models

$$MS_{\Delta} = \frac{SS_{\Delta}}{df_{\Delta}}$$

We extract the mean squares for the full model (with the main effects & interaction terms specified)

$$MS_{Model-1} = \frac{SS_{Model-1}}{df_{Model-1}}$$

Now we can estimate the F-ratio

$$F = \frac{MS_{\Delta}}{MS_{Model-2}}$$

We can use the `anova()` function directly on the model objects

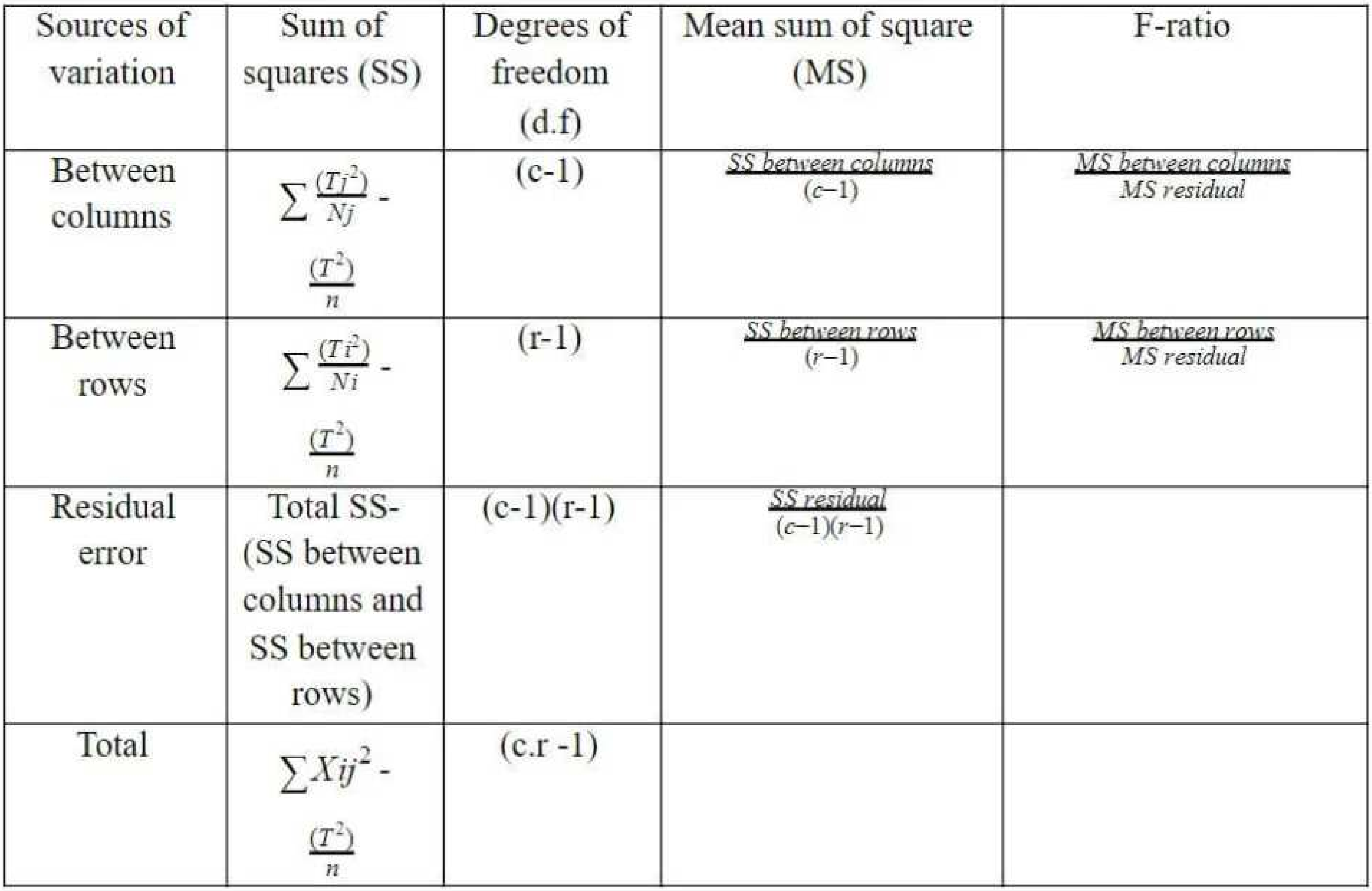
```
anova(mod1,mod2)
```

```
## Analysis of Variance Table
##
## Model 1: Alcohol ~ Depression
## Model 2: Alcohol ~ Depression + Location + Depression:Location
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     10 24032
## 2      8 21684  2    2348.2 0.4332 0.6628
```

Because there is no significant impact of including the additional terms (*Location* and *Depression:Location*), we can retain our initial model that explored for main effects across *Depression* exclusively.

Balanced groups across ANOVAs are ideal, but not always possible. After you finish data collection, you may have to drop participants from conditions due to missing data, programming errors, or a host of non-predicted reasons. Additionally, because all our predictors had two levels, there was no need for running post-hoc tests following significant main effects.

We continue our discussion of factorial ANOVAs the following week, where we will discuss factors with $k > 2$ levels and run post-hoc tests when factors are significant. The lab activity will be provided at the end of next week's class.



Sources of variation	Sum of squares (SS)	Degrees of freedom (d.f)	Mean sum of square (MS)	F-ratio
Between columns	$\sum \frac{(Tj^2)}{Nj}$ - $\frac{(T^2)}{n}$	(c-1)	$\frac{SS\ between\ columns}{(c-1)}$	$\frac{MS\ between\ columns}{MS\ residual}$
Between rows	$\sum \frac{(Ti^2)}{Ni}$ - $\frac{(T^2)}{n}$	(r-1)	$\frac{SS\ between\ rows}{(r-1)}$	$\frac{MS\ between\ rows}{MS\ residual}$
Residual error	Total SS-(SS between columns and SS between rows)	(c-1)(r-1)	$\frac{SS\ residual}{(c-1)(r-1)}$	
Total	$\sum Xij^2$ - $\frac{(T^2)}{n}$	(c.r -1)		

Source: https://microbenotes.com/anova/ (https://microbenotes.com/anova/)