PS303: Week 8

pp. 425-452

Recap

- When working with nominal/ordinal (categorical) data, we tested whether *observed* distributions of proportions were statistically different relative to *expected* proportions. This included:
  - Binomial tests for testing binary data distributions (categorized as **successes** and **failures**)
  - χ^2 tests for assessing proportion distributions of >2 categories.
- In the case of continuous data, we test whether *observed* point estimates (e.g., means, medians) vary from *expected* estimates (typically a null difference).Because point estimates refer to *sample properties* whereas we are generally interested in *population parameters*, we typically supplement point estimates with range estimates. The latter may include standard deviation, standard error, and confidence intervals.
- We have discussed test statistics for contrasting pairs of continuous data using Student's and Welch's *t*-tests. This included:
  - 1-sample tests for contrasting a single sample's parameter with a known population parameter.
  - 2-sample tests for contrasting between two sample parameters. When data was sampled from two independent groups, we ran an independent/Welch's *t*-test. When data was sampled from the same group at two time-points, we ran a paired *t*-test.
  - We focused on Welch's test since it is more robust to violations of homogeneity of variance and balanced samples.
- When we are interested in analyzing the variance of ≥ 3 groups, we can run an Analyses of Variance (*ANOVA*). Instead of t or χ^2 test statistics, ANOVA generates *F-ratios* that identifies the likelihood of current observations relative to a null hypothesis (H_O).
  - If H_O is rejected, we can follow up our *ANOVA* with *post-hoc* two-sample tests to estimate whether there may be any meaningful differences between group pairs.
  - We don't *begin* our analysis with two-sample tests as the number of contrasts that are run is positively associated with increased Type-1 error (false positives become more likely to be detected). Think of why this may be the case in light of the $p < .05$ threshold that is common in psychological research (see here for details).
  - We will begin with running a one-way ANOVA on a real dataset
  - Similar to *t*-tests, ANOVAs are least biased when samples are balanced, data is normal, and there is homogeneity of variance. When the latter assumption is not met, than Welch's *F*-test and Kruskal-Wallis tests are commonly used alternatives.

Table 1: Anxiety scores across 3 studying spaces

Green Space	Urban Space	Mixed Space
6	3	3
1	8	2
2	7	6
6	6	6
4	5	7
4	3	8
3	5	3
1	8	8
5	7	6
3	3	3
3	9	8
4	7	6
2	5	4
6	9	3
7	3	6
4	5	5
1	9	4
5	6	4
6	9	3
2	3	5

Note:

All scores are simulated.

One-way ANOVA

> Use: You have ≥ 3 levels of an independent variable and a continous dependent outcome. The sample is balanced, there is homogeneity of variance, and the data is normally distributed.

Example: You remember reading somewhere that *increased proportion of (perceivable) green space (is) associated with decreased anxiety/mood disorder… in an urban environment* (Nutsford, Pearson & Kingham, 2013). You speculate that a similar ‘green advantage’ may help in reducing students’ anxiety before their exam.

To test your hypothesis, you recruit 60 students ($N = 60$). Each student is required to study for an hour in one of 3 environments before sitting for an exam. The environments include a ‘green space’, an ‘urban space’ and a ‘mixed space’. After students have spent an hour, you ask them to complete an anxiety questionnaire. You tabulate their outcomes across the following table (higher scores indicate greater anxiety).

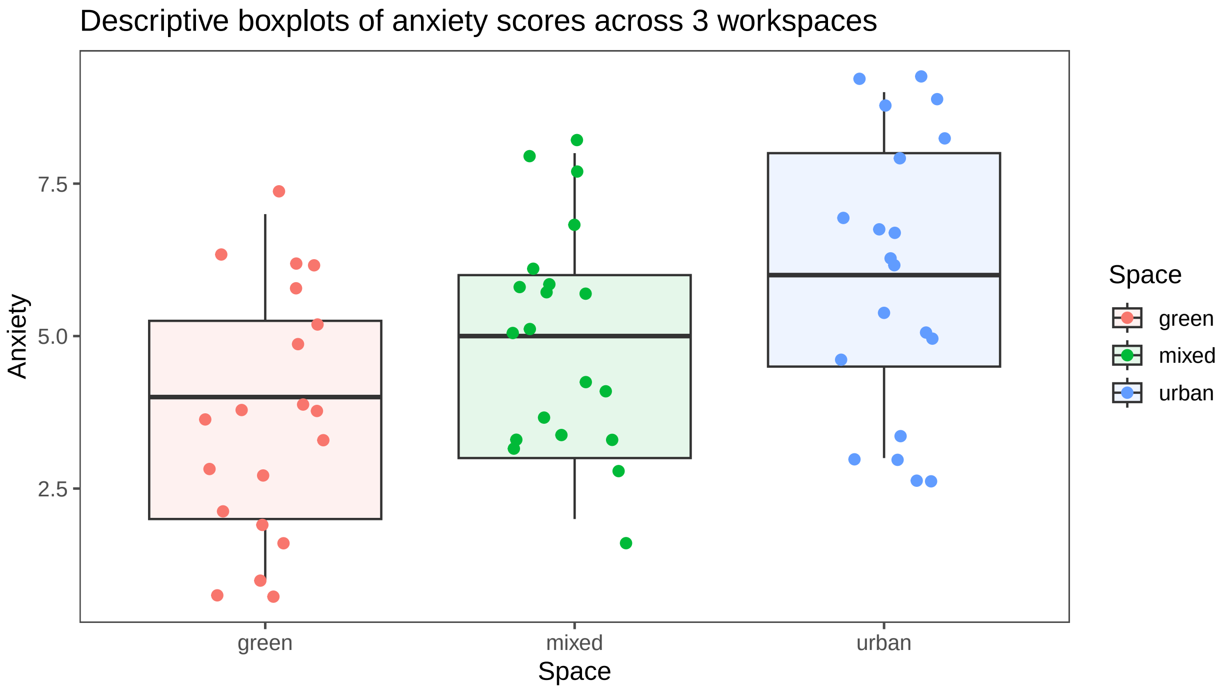
We can describe estimate the mean, standard deviation (SD) and standard error ($\frac{SD}{\sqrt{N-1}}$) to get an overview of the data.

```
# Function for estimating mean, sd and se for a numeric variable (all values rounded to 2 decimal place

mean_sd_se <- function(x){
  m  <- round(mean(x),2)
  sd <- round(sd(x),2)
  se <- round(sd(x)/sqrt((length(x)-1)),2)
  msd <- paste0("M = ", m, "; SD = ", sd, "; SE = ", se)
  msd
}
```

- Green space (n= 20): M = 3.75; SD = 1.89; SE = 0.43
- Urban space (n= 20): M = 6; SD = 2.25; SE = 0.52
- Mixed space (n= 20): M = 5; SD = 1.89; SE = 0.43

We could explore the data visually



Anxiety scores collected from participants studying in green space appears to be the lowest. *But is this statistically significant*? We can test the null hypothesis that the mean anxiety scores ($\mu's$) across the three study spaces are not statistically different.

$$H_O : \mu_{Green} = \mu_{Urban} = \mu_{Mixed}$$

.

Caution: Before running a test, it is necessary to check whether the **assumptions** for a test are satisfied - is the data normal? Are the samples balanced? Is there homogeneity of variance?

Preparing the data

Let's set up our variables

```
green<-c(6,1,2,6,4,4,3,1,5,3,3,4,2,6,7)
urban<-c(6,3,7,8,4,3,8,7,6,5,3,5,8,7,3)
mixed<-c(8,6,4,8,2,4,8,5,8,2,3,2,6,6,7)
```

You can use the function (from the previous slide) to extract means, standard deviations and standard error. *For example*

```
# Get summary stats for anxiety scores are green space users
mean_sd_se(green)
```

```
## [1] "M = 3.8; SD = 1.9; SE = 0.51"
```

It's good practice to keep all relevant data within a single structure, such as a **data frame**. We will combine the three variables (green, urban, mixed) as columns, then re-organize the structure for easier analyses further on.

```
# Install package (first time only)
# install.packages('tidyverse')

# Load package
require('tidyverse')

# Let's set up the variables from before into a dataframe
df <- cbind.data.frame(green,urban,mixed)

# Re-organize the dataframe (so factors are variables)
df.1<-gather(data = df,                       # Data frame created above
             key="study.space",                 # Factor (IV) label
             value = "anxiety.score",             # Outcome (DV) label
             green,urban,mixed) %>%                 # Factor levels
             mutate(study.space=factor(study.space)) %>%  # Identify as factor
             mutate(ID = row_number())        # Create ID variable (optional for now)

# Print outcome
kbl(df.1,booktabs = T) # Using the 'kableExtra::kbl()' function for prettier output
```

study.space	anxiety.score	ID
green	6	1
green	1	2
green	2	3
green	6	4
green	4	5
green	4	6
green	3	7
green	1	8
green	5	9
green	3	10
green	3	11
green	4	12
green	2	13
green	6	14
green	7	15
urban	6	16
urban	3	17
urban	7	18
urban	8	19
urban	4	20
urban	3	21
urban	8	22
urban	7	23
urban	6	24
urban	5	25
urban	3	26
urban	5	27
urban	8	28
urban	7	29
urban	3	30
mixed	8	31
mixed	6	32
mixed	4	33
mixed	8	34
mixed	2	35
mixed	4	36
mixed	8	37
mixed	5	38
mixed	8	39
mixed	2	40
mixed	3	41
mixed	2	42
mixed	6	43
mixed	6	44
mixed	7	45

Scores represent individual participants. For most cases, individual participant data should be presented as rows.

Once you have a data structure organized, it becomes easier to run operations whole scale. For example, using the `rstatix` package, we can get a summary of the data

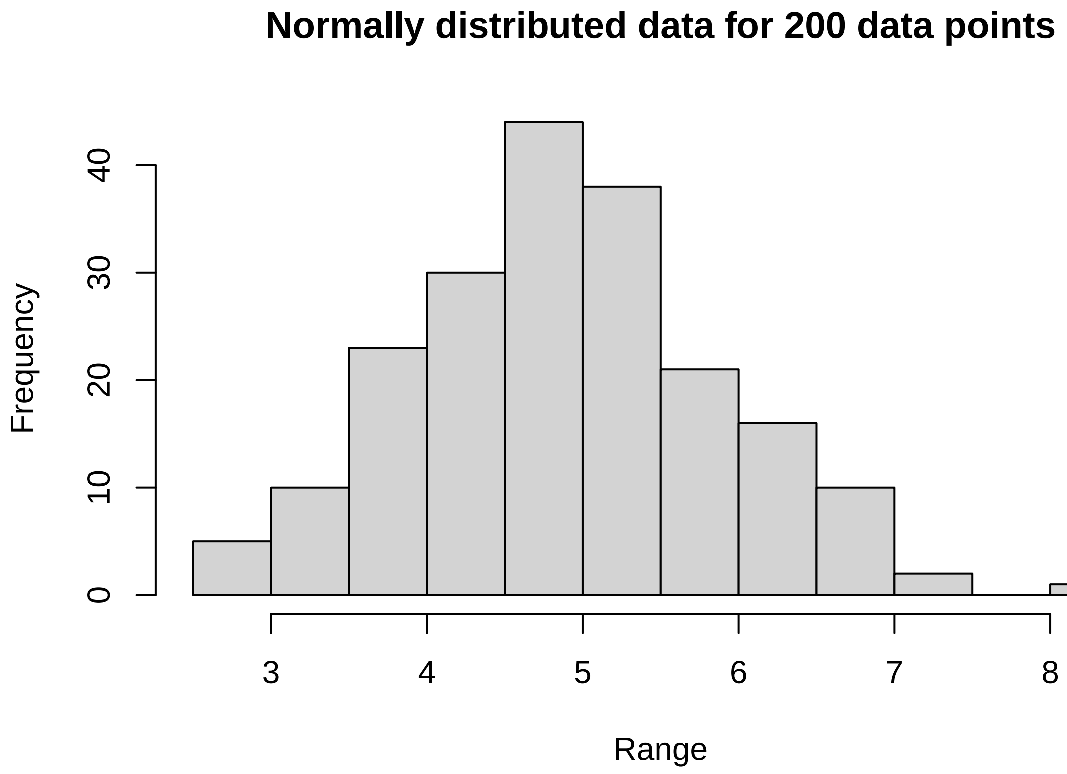
```r
# Load package
require(rstatix)

df.1 %>%                       # Select the data frame of interest
group_by(study.space) %>%       # Group our data by the factor of interest (IV)
get_summary_stats(anxiety.score) # Select the DV to summarize by factor
```

```
## # A tibble: 3 x 14
##   study.s~1 varia~2     n   min   max median    q1    q3   iqr   mad  mean    sd
##   <fct>     <fct>   <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 green     anxiet~    15     1     7      4   2.5   5.5   3    2.96  3.8   1.90
## 2 mixed     anxiet~    15     2     8      6   3.5   7.5   4    2.96  5.27  2.31
## 3 urban     anxiet~    15     3     8      6   3.5   7     3.5  2.96  5.53  1.96
## # ... with 2 more variables: se <dbl>, ci <dbl>, and abbreviated variable names
## #   1: study.space, 2: variable
```

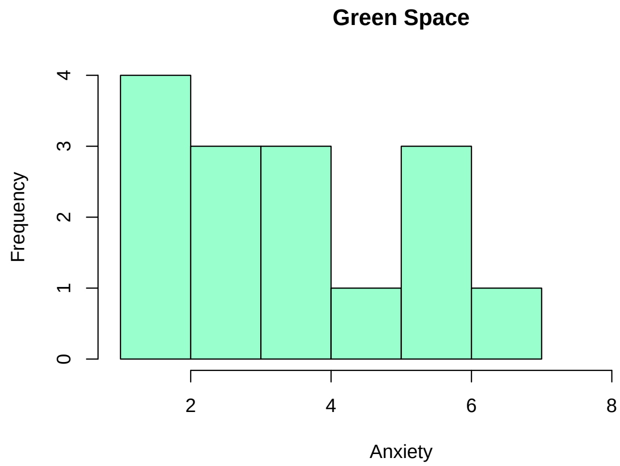
Assumption checks: Is the data normally distributed?

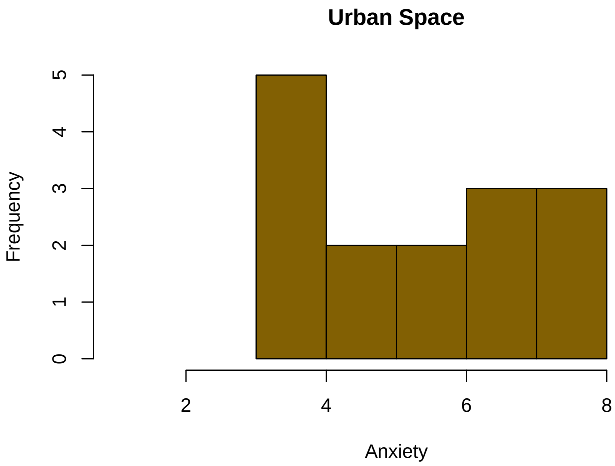
There are a variety of methods for assessing a normal distribution. First, we can construct histograms to visually inspect the 'shape' of the distributions. In a normal distribution, the majority of data are clustered around the mean. Here is an example of a normally distributed dataset

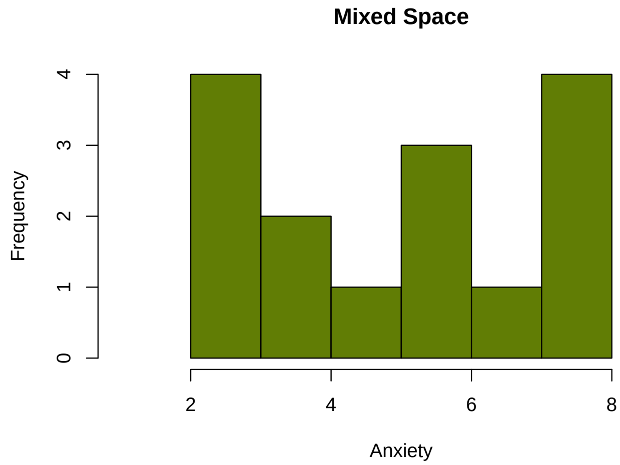


Now let's look at the distributions of data we have

```
# The scale ('xlim') ranges from '1' (Not anxious) to '9' (Very anxious)
hist(green,main = "Green Space",col = "#99ffcc",xlab = "Anxiety",xlim = (c(1,9)))
hist(urban,main = "Urban Space",col = "#826003",xlab = "Anxiety",xlim = (c(1,9)))
hist(mixed,main = "Mixed Space",col = "#617d05",xlab = "Anxiety",xlim = (c(1,9)))
```







The data does not look normal, although this may simply be due to an insufficient sampling distribution. We may also run `Shapiro-Wilk` tests to formally estimate if the observed distributions significantly vary from a normal distribution (here, a *lack* of a significant effect is desirable as it suggests the assumption for normality had not been violated).

```
shapiro.test(green)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  green
## W = 0.9431, p-value = 0.4229
```

```
shapiro.test(urban)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  urban
## W = 0.87376, p-value = 0.03834
```

```
shapiro.test(mixed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  mixed
## W = 0.8831, p-value = 0.05279
```

2 out of 3 distributions **do not** violate the normality assumption (but the data for Urban Spaces do) - this will be sufficient for us to continue with our assumption tests (we will go through an alternative procedure later).

> If you want to know more about the Shapiro-Wilkinson test, check out this link.

Assumption checks: Are sample variances homogenous?

```r
# Load the package
require(rstatix)

# We can run a Levene's test on the dataframe created previously
# A non-significant outcome imples the homogeneity of variance assumptions has NOT been violated
lev.test <- levene_test(data=df.1,formula = anxiety.score~study.space)
lev.test
```

```
## # A tibble: 1 x 4
##     df1   df2 statistic     p
##   <int> <int>     <dbl> <dbl>
## 1     2    42     0.451 0.640
```

Levene's test assumes homogeneity of variance as the null hypothesis

$$H_O : SD^2_{green} = SD^2_{urban} = SD^2_{mixed}$$

. The *p*-value suggests there is a 0.64 likelihood of observing the current distributions if the null is 'true'. Because this is larger than the significance threshold (p=.05), we can *retain* the null. Critically, the homogeneity of variance assumption is **conserved**.

Running a one-way ANOVA

Using R's built-in functions…

```
# Base functions ('aov' for anovas)
res.aov <- aov(anxiety.score~study.space, # Describe the formula; DV~IV
               df.1) # Name of the relevant data structure

# Use the summary argument to extract the relevant parameters
summary(res.aov)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## study.space  2  26.13  13.067   3.065 0.0572 .
## Residuals   42 179.07   4.263
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Another approach is to use the `rstatix` package (this will be useful for more complex models later).

```
# Ensure the 'rstatix' package is loaded beforehand
classic.aov <-anova_test(df.1,          # Data structure
              anxiety.score~study.space,# Formula for ANOVA
              wid=ID,                   # ID variable
              effect.size = "pes")      # Effect size

# Print
classic.aov
```

```
## ANOVA Table (type II tests)
##
##         Effect DFn DFd     F     p p<.05   pes
## 1 study.space   2  42 3.065 0.057       0.127
```

We can report our findings as follow:

A one-way ANOVA indicated that anxiety scores did not significantly vary between green, urban and mixed study spaces, $F(2, 42) = 3.06$, $p = .057$, $\eta_p^2 = .13$. The null hypothesis may be retained.

ANOVA alternatives: Welch's ANOVA

Welch's ANOVA is generally the preferred alternative when the homogeneity of variance has been **violated** (Levene's test generates a p-value less than .05) while the data remains balanced and normally distributed. Alternatively, if homoscedasticity *and* normality is conserved, you could run the classic ANOVA (even if your sample is unbalanced).

Although homoscedasticity was conserved across our simulated dataset, we will run Welch's ANOVA for the sake of illustration:

```
# Using the `rstatix` package
require(rstatix)
waov <- welch_anova_test(df.1,anxiety.score~study.space)
waov
```

```
## # A tibble: 1 x 7
##   .y.               n statistic   DFn   DFd     p method
## * <chr>         <int>     <dbl> <dbl> <dbl> <dbl> <chr>
## 1 anxiety.score    45      3.36     2  27.8 0.049 Welch ANOVA
```

In this case the null hypothesis is rejected. However, *because* the homogeneity of variance assumptions were **not** violated, it would be typically inappropriate to report Welch's F-test.

One recommended approach is to avoid testing for homoscedasticity in the first place and simply use Welch's ANOVA by default when running one-way tests (Delacre, Leys, Mora, & Lakens, 2019). According to those authors, two reasons for this are:

- Homogeneity of sample variance (homoscedasticity) is unlikely in real-life circumstances.
- There is a considerable gain in Type-1 error control rates (lower chance of detecting false positives).

For further details, check out the article by Delacre et al (2019) linked above.

We can report Welch's ANOVA in the same way as a classical ANOVA:

A Welch's F-test indicated that anxiety score variance across the three study spaces were significantly different from the null, $F_{2,27.8} = 3.36; p = 0.049$. We are now justified in running *post-hoc* tests to examine which group means vary significantly relative to one another.

Post-hoc comparisons

There are a variety of methods available for contrasting between group means following a significant F-test.
- If samples are **balanced and homoscedastic**, a common approach is Tukey's Honestly Significant Difference test (abbreviated as **Tukey HSD**).
- If samples are **unbalanced and/or heteroscedastic**, we can run the Games-Howell test.

Post-hoc tests vary from conventional t-tests by controlling for familywise error rates (the inflation of Type-1 error associated with multiple comparisons). Specifically, instead of a 'one-size-fits-all' p-value, the latter becomes adjusted in accordance with the number of comparisons being made. For additional details on *how* this is achieved, see the discussion here by Jim Frost..

We will use both methods for the sake of illustration:

1. Tukey's HSD

```
# When samples are balanced and homoscedastic
tukey_hsd(df.1,anxiety.score~study.space)
```

```
## # A tibble: 3 x 9
##   term        group1 group2 null.value estimate conf.low conf.h~1  p.adj p.adj~2
## * <chr>       <chr>  <chr>       <dbl>    <dbl>    <dbl>    <dbl>  <dbl> <chr>
## 1 study.space green  mixed           0    1.47   -0.365      3.30 0.139  ns
## 2 study.space green  urban           0    1.73   -0.0984     3.57 0.0669 ns
## 3 study.space mixed  urban           0    0.267  -1.57       2.10 0.933  ns
## # ... with abbreviated variable names 1: conf.high, 2: p.adj.signif
```

2. Games-Howell

```
# When samples are unbalanced and/or heteroscedastic
games_howell_test(df.1,anxiety.score~study.space)
```

```
## # A tibble: 3 x 8
##   .y.           group1 group2 estimate conf.low conf.high p.adj p.adj.signif
## * <chr>         <chr>  <chr>     <dbl>    <dbl>     <dbl> <dbl> <chr>
## 1 anxiety.score green  mixed     1.47  -0.449        3.38 0.159 ns
## 2 anxiety.score green  urban     1.73  -0.00916      3.48 0.051 ns
## 3 anxiety.score mixed  urban     0.267 -1.67         2.21 0.938 ns
```

We can expand our earlier outcome statements with the above results - if we ran a classical ANOVA with Tukey's HSD we can report:

A one-way ANOVA indicated that anxiety scores did not significantly vary between green, urban and mixed study spaces, $F(2,\ 42) = 3.06$, $p = .057$, $\eta_p^2 = .13$. Tukey's HSD confirmed that none of the group means significantly varied from each other (all p's $> .06$), as expected (**Note: You should not run post-hoc tests if your ANOVA is non-significant**).

Alternatively, if we ran Welch's F-test and Games-Howell tests, we may report:

A Welch's F-test indicated that anxiety score variance across the three study spaces were significantly different from the null, $F_{2,27.8} = 3.36; p = 0.049$. Post-hoc Games-Howell tests indicated none of the group means varied significantly from each other, although the difference between anxiety scores collected in green and urban studying spaces approached significance ($p = .051$).

Table 2: Personality scores across 5 dimensions

Extroversion	Openness	Conscientuousness	Neuroticism	Agreeableness
6	6	6	6	7
9	2	5	2	2
8	1	9	2	1
1	5	7	7	3
5	4	5	8	8
3	1	9	3	9
4	8	4	5	3
8	3	1	4	1
4	8	2	3	3
1	NA	1	4	NA

Note:
Missing data (NA) for Openness and Agreeableness

ANOVA alternatives: Kruskal-Wallis tests

You may occasionally come across data that is *unbalanced* designs (levels have varying numbers of observations), not normal and/or homoscedastic. When assumptions are not met, we run the Kruskal-Wallis test across parameter medians.

Suppose we have the following data from a personality test.

When setting up our data frame, you can include *NA* in the vector to account for missing values. By equating vector lengths, it is possible to combine them into a data frame.

```
# 5 personality dimensions
e <- c(6, 9, 8, 1, 5, 3, 4, 8, 4, 1)
o <- c(6, 2, 1, 5, 4, 1, 8, 3, 8,"NA")
c <- c(6, 5, 9, 7, 5, 9, 4, 1, 2, 1)
n <- c(6, 2, 2, 7, 8, 3, 5, 4, 3, 4)
a <- c(7, 2, 1, 3, 8, 9, 3, 1, 3,"NA")

# Combine into data frame and print
df2 <- cbind.data.frame(e,o,c,n,a)
df2
```

```
##    e  o c n  a
## 1  6  6 6 6  7
## 2  9  2 5 2  2
## 3  8  1 9 2  1
## 4  1  5 7 7  3
## 5  5  4 5 8  8
## 6  3  1 9 3  9
## 7  4  8 4 5  3
## 8  8  3 1 4  1
## 9  4  8 2 3  3
## 10 1 NA 1 4 NA
```

We can use the `gather` function to prepare the data…

```
# Gather columns into a single ID variable
df3 <-gather(df2,                  # Name of data frame
             key = "Personality", # IV label
             value = "Ratings",   # DV label
             e,o,c,n,a)           # IV levels
df3
```

```
##    Personality Ratings
## 1            e       6
## 2            e       9
## 3            e       8
## 4            e       1
## 5            e       5
## 6            e       3
## 7            e       4
## 8            e       8
## 9            e       4
## 10           e       1
## 11           o       6
## 12           o       2
## 13           o       1
## 14           o       5
## 15           o       4
## 16           o       1
## 17           o       8
## 18           o       3
## 19           o       8
## 20           o      NA
## 21           c       6
## 22           c       5
## 23           c       9
## 24           c       7
## 25           c       5
## 26           c       9
## 27           c       4
## 28           c       1
## 29           c       2
## 30           c       1
## 31           n       6
## 32           n       2
## 33           n       2
## 34           n       7
## 35           n       8
## 36           n       3
## 37           n       5
## 38           n       4
## 39           n       3
## 40           n       4
## 41           a       7
## 42           a       2
## 43           a       1
## 44           a       3
## 45           a       8
## 46           a       9
## 47           a       3
## 48           a       1
## 49           a       3
## 50           a      NA
```

If homogeneity of variance is conserved (Levene's test is *not* significant), we could run a conventional ANOVA. Otherwise, we can run a non-parametric **Kruskal-Wallis** test which compares between group *median* (not mean) estimates. KW-tests are robust to violations of normality and large outliers, and preferred when assumptions are not met.

```
## Two ways of running the same KW test...

# Using base R
kruskal.test(Ratings~Personality,df3)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Ratings by Personality
## Kruskal-Wallis chi-squared = 0.19624, df = 4, p-value = 0.9955
```

```
# Using 'rstatix'
kruskal_test(df3,Ratings~Personality)
```

```
## # A tibble: 1 x 6
##   .y.         n statistic    df     p method
## * <chr>   <int>     <dbl> <int> <dbl> <chr>
## 1 Ratings    50     0.196     4 0.995 Kruskal-Wallis
```

The test was not significant. We can report this as follows:

> A Kruskal-Wallis test indicated the five group medians did not significantly vary from the null, $\chi^2(4) = .19; p = .996$.

Summary of steps in a classical ANOVA

Is the difference in mean estimates between (at least) three groups (call these **A, B, C**) significant?

- $H_0 : \mu_A = \mu_B = \mu_C$ - Null hypothesis
- $H_A : \mu_A \neq \mu_B \neq \mu_C$ - Alternative hypothesis

1. Estimate *between-group* variability (MS_B) to quantify how much individual group means ($\mu_A, \mu_B, \mu_C ... \mu_k$) vary from the overall/**grand** mean (μ). We first estimate the **Sum of Squared variance Between groups**, or SS_B. We weigh each group's variance in respect to it's sample size (n_A, n_B, n_C).

   $$SS_B = n_A(\mu_A - \mu)^2 + n_B(\mu_B - \mu)^2 + n_C(\mu_C - \mu)^2 = \sum n_k(\mu_k - \mu)^2$$

   We next estimate mean between-group variance (MS_B) by dividing SS_B with between group degrees of freedom ($df_k = k - 1$ where k is the number of groups).

   $$MS_B = \frac{SS_B}{df_k} = \frac{\sum n_k(\mu_k - \mu)^2}{k - 1}$$

   .

2. Estimate *within-group* variability (MS_W) to quantify 'spread' of individual means within respected groups. We first estimate the **Sum of Squared variance Within groups**, or SS_W. Assuming each individual observation within a single group as i, then

   $$SS_W = \sum(A_i - \mu_A)^2 + \sum(B_i - \mu_B)^2 + \sum(C_i - \mu_C)^2 = \sum(k_i - \mu_k)^2$$

   . We can next estimate the average within group variance, where

   $$MS_W = \frac{SS_W}{df_W} = \frac{\sum(k_i - \mu_k)^2}{N - k}$$

3. MS_B tells us how much group means vary from the grand mean, and MS_W tells us how much individual observations vary from their respective group means. The F-statistic is estimated as the ratio of the two variability estimates.

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

   A larger F-ratio implies the variance between groups is greater than the variance within groups. The larger the difference, the greater the likelihood of the tested samples representing different populations. We can check whether this difference is statistically significant ($p < .05$) by finding out whether the observed F is larger than the **critical** F value. This can be looked up manually or estimated in R using `qf(p,df1,df2,lower.tail=F)`, where `p` indicates the significance threshold, `df1` indicates between groups df_k, and `df2` indicates within groups df_W.

4. *Assuming a significant effect was found*, we can estimate an effect size (η^2) by dividing the between-groups variability (SS_B) by the total variability ($SS_{Total} = SS_B + SS_W$). So,

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

A significant ANOVA does not tell us which specific groups vary relative to each other, which we investigate using post-hoc tests.

Activity

You selected USP as a preferred college based on the notion that USP graduates have higher IQ scores than FNU or UoF graduates. To test this hypothesis, you collected the IQ scores for 20 students from USP, 20 students from FNU and 18 students from UoF. The data is listed below:
- 20 USP students' IQ scores:89,99,94,86,90,101,110,109,96,95,88,106,85,102,104,97,91,107,98,93
- 20 FNU students' IQ scores:110,105,109,87,86,104,106,103,92,95,98,89,101,99,100,90,108,93,97,85
- 18 UoF students' IQ scores:85,91,101,90,86,87,108,95,104,110,88,109,92,106,102,96,105,93,NA,NA

Please complete all activity questions

1. Assign the values provided above into separate variables, than combine them to a tidy dataframe (remember to use the `gather` function from the `tidyverse` package to ensure the factor and outcome variables are two seperate columns). Show your code and output.
2. Assess each variable's normality using histograms and Shapiro-Wilk tests.
3. Assess whether sample variances are homogeneous.
4. Run a one-way ANOVA *or* a Welch's ANOVA *or* a Kruskal-Wallis test. Provide reasons for test selection and report your findings in APA format.
5. If your model is significant, run post-hoc tests using Tukey's HSD **or** Games-Howell. Provide reasons for your selection. Report your output in APA format.

Submit your work in the dropbox for this week.