Introducing ANOVA

PS203

Review of *t*-tests

- Last week we discussed t-tests for identifying statistically significant differences between group means. This included:
    - Single-sample *t*-tests for exploring whether a sample mean is statistically different from that population's mean.
    - Two-sample *t*-tests for exploring whether two sample means are statistically different. These include: ++ Independent *t*-tests, which contrast samples representing independent participant groups ++ Paired *t*-tests contrast across the same group at different time points
- The produces a test statistic, which can inform us how probable our observations are assuming the null hypothesis is true. If the observed data is 'extremely unlikely' ($p < .05$), we reject the null hypothesis. Values of *p* higher then .05 means we may *retain* the null hypothesis
- What if we are interested in estimating differences between more than 2 groups?

Analysis of Variance (ANOVA)

- Whenever we have an factorial variable with at least 3 levels ($k \geq 3$) and are interested in estimating group differences, we can run an Analysis of Variance, or ANOVA.
- The ANOVA is a statistical method used to test the difference between the means of more than two groups. The purpose of the test is to determine if there is a significant difference between the means of the groups, and if so, which specific groups are different from one another.
- Similar to t-tests, ANOVAs can be set to run *between* independent levels of a factor, or *across* repeated levels within a factor, or some combination of both.
- If an ANOVA is significant, we can follow this up with two-sample tests between levels of a statistically relevant facttpr.

Conditions for running an ANOVA

- Before running ANOVAs, our data must meet certain assumptions:
    - *Independence of observations*: Each observation in each group should be independent of the others.
    - *Normality*: The observations within each group should be normally distributed.
    - *Equal variances*: The variances of the groups should be equal and homogeneous.
- Violating these assumptions can affect the validity of the results and alternative statistical tests, such as the Kruskal-Wallis test or Welch's ANOVA, may be more appropriate.
- Today we will go over how to run a ANOVA. We will expand on the topic in PS303.

A simulated dataset

> Suppose we have collected physical activity data for six students. The data includes amount of time spent on Walking, Vigorous and Moderate physical activity in minutes for a given week. We may be interested to know whether the mean times spent engaging in these various activities statistically vary between groups.

	Walking (W)	**Vigorous (V)**	**Moderate (M)**
$Student_1$	240	120	80
$Student_2$	300	130	240
$Student_3$	250	110	190
$Student_4$	420	212	200
$Student_5$	300	98	199
$Student_6$	290	150	321

- All columns represent the same scale (minutes)
- Research hypothesis: People spend more time walking than engaging in vigorous or moderate activity

Null hypothesis: There is no statistical difference in the average time spent on the three activities.

$$H_0 : W_\mu = V_\mu = M_\mu$$

Let's prepare the data in `R`

```r
# Install and load the tidyverse package
require('tidyverse')

# Assign the values to variables
walk     <- c(240,300,250,420,300,290)
vigor    <- c(120,130,110,212,98,150)
moderate <- c(80,240,190,200,199,321)

# Combine the values into a single data frame called 'df'
df <- cbind.data.frame(walk,vigor,moderate)

# Tidy the data and store into a new data object called 'df1'
df1 <- gather(df,key="Activity",value="Minutes",convert = T)

# Create an ID variable
df1$ID <- rep(c(1:6),3)

# View output
df1
```

```
##    Activity Minutes ID
## 1      walk     240  1
## 2      walk     300  2
## 3      walk     250  3
## 4      walk     420  4
## 5      walk     300  5
## 6      walk     290  6
## 7     vigor     120  1
## 8     vigor     130  2
## 9     vigor     110  3
## 10    vigor     212  4
## 11    vigor      98  5
## 12    vigor     150  6
## 13 moderate      80  1
## 14 moderate     240  2
## 15 moderate     190  3
## 16 moderate     200  4
## 17 moderate     199  5
## 18 moderate     321  6
```

The data above contains data from multiple (*k*=3) factorial levels for each participant. Because we are exploring group means across multiple time points across the *same* participants, we can run a within-subjects ANOVA.

> The three columns on the printed table refer to: - *Activity*: Three levels of the independent factorial variable (physical activity). - *Minutes*: Dependent numeric variable (minutes spent) - *ID*: Six levels corresponding to individual students

To run an ANOVA on `R`, ensure all your independent variable levels are represented in a single column. Each row should represent a single observation/participant. The ID column should inform us which observation corresponds with which participant. This is necessary for running *within-subjects* ANOVAs

> We can represent the ANOVA as a 'model' to R, which can then run the necessary procedures to derive whether the observed test statistic is significant.

Recall that we had created the dataframe `df1` earlier, which contained the variables `Activity` and `Minutes`, with `ID` as the identifier variable.

This can be built into a model using the following formula:

```
aov(formula = Dependent_Variable~Indepentent_Variable+Error(ID/Independent_Variable),data = data_frame)
```

```
# Assign the ANOVA results to a variable called 'mod1'
mod1 <- aov(formula = Minutes~Activity + Error(ID/Activity),
            data=df1)

# Summarize the results
summary(mod1)
```

```
##
## Error: ID
##           Df Sum Sq Mean Sq F value Pr(>F)
## Residuals  1  13249   13249
##
## Error: ID:Activity
##          Df Sum Sq Mean Sq
## Activity  2  69822   34911
##
## Error: Within
##           Df Sum Sq Mean Sq F value Pr(>F)
## Activity   2  17576    8788   2.658  0.111
## Residuals 12  39670    3306
```

```
## Warning: A call to `aov` does not support sphericity correction, continuing
## without correction of possible violated sphericity
```

Our ANOVA model is *not* statistically significant, $F(2, 12) = 2.66$, $p = .111$, $\eta_p^2 = .31$. We can retain the null hypothesis ($W_\mu = V_\mu = M_\mu$) which stated there would be no difference between times spent on the different activities. A common effect size parameter (η_p^2) reported following ANOVAs is known as partial-eta squared.

Post-hoc tests

We can run t-tests to identify differences between pairs of groups. Although we had retained $H_0 : W_\mu = V_\mu = M_\mu$, we will still test the following null hypotheses to illustrate how post-hoc tests would be reported:

1. $H1_0 : W_\mu = V_\mu$ (Walking = Vigorous)
2. $H2_0 : W_\mu = M_\mu$ (Walking = Moderate)
3. $H3_0 : M_\mu = V_\mu$ (Moderate = Vigorous)

You have to load the `tidyverse` and `rstatix` packages prior to running the following.

```
# Load packages
require('tidyverse')
require('rstatix')

# Run pairwise tests between levels of each factor
paired<- df1 %>% rstatix::pairwise_t_test(Minutes~Activity,p.adjust.method = "holm",paired=TRUE)

# Output
paired
```

```
## # A tibble: 3 x 10
##   .y.     group1   group2    n1    n2 statistic    df        p    p.adj p.adj.~1
## * <chr>   <chr>    <chr>  <int> <int>     <dbl> <dbl>    <dbl>    <dbl> <chr>
## 1 Minutes moderate vigor      6     6      2.10     5 0.089    0.09     ns
## 2 Minutes moderate walk       6     6     -2.66     5 0.045    0.09     ns
## 3 Minutes vigor    walk       6     6    -11.1      5 0.000104 0.000312 ***
## # ... with abbreviated variable name 1: p.adj.signif
```

Before correcting for multiple comparisons (see the $p.adj$ column), we can see that the average amount of time spent walking is significantly different from time spent on vigorous activity,$t_5 = 11.095, p = .0001$, and from time spent on moderate activity, $t_5 = 2.662, p = .045$. After applying a Holm correction (to reduce inflated Type-1 error rates due to multiple comparisons), only the difference between walking and vigorous activity remain significant ($p = .0003$).

- We can estimate effect sizes to estimate whether the observed differences might be 'practically important'.
- We will estimate the Hedge's *g* statistics, which is more robust to small sample sizes and unequal group variances relative to the more commonly reported Cohen's *D* statistic.

```
# Estimate Hedge's g effect sizes
effects <- df1 %>%
  rstatix::cohens_d(Minutes~Activity,
                    hedges.correction = TRUE,
                    paired = TRUE)

# View output
effects
```

```
## # A tibble: 3 x 7
##   .y.     group1   group2 effsize    n1    n2 magnitude
## * <chr>   <chr>    <chr>    <dbl> <int> <int> <ord>
## 1 Minutes moderate vigor    0.723     6     6 moderate
## 2 Minutes moderate walk    -0.915     6     6 large
## 3 Minutes vigor    walk    -3.81      6     6 large
```

All effect sizes shown under the `effsize` column, with interpretations shown under the `magnitude` column

> A very large effect ($g_{Hedge} = 3.81$) was observed for the difference between walking and vigorous physical activity durations, which corresponds with the statistically significant difference noted earlier.

We can combine all our results into a single table. Recall that `paired` was the variable we assigned all *t*-test outcomes to, and `effects` was the variable we assigned all Hedge's *g* outputs to.

```
# Combine results tables using `cbind()` argument
results <- cbind(paired,effects)

# Output
results
```

```
##       .y.   group1 group2 n1 n2  statistic df        p    p.adj p.adj.signif
## 1 Minutes moderate  vigor  6  6   2.103167  5 0.089000 0.090000           ns
## 2 Minutes moderate   walk  6  6  -2.661855  5 0.045000 0.090000           ns
## 3 Minutes    vigor   walk  6  6 -11.095177  5 0.000104 0.000312          ***
##       .y.   group1 group2    effsize n1 n2 magnitude
## 1 Minutes moderate  vigor  0.7230437  6  6  moderate
## 2 Minutes moderate   walk -0.9151138  6  6     large
## 3 Minutes    vigor   walk -3.8143891  6  6     large
```

> Too many columns, some of which are redundant? We can extract only the columns we are interested in…

```
table_df <- results[,c(2,3,6,7,9,14,17)]  # Select the columns of interest by order

#Output table
table_df
```

```
##     group1 group2  statistic df    p.adj    effsize magnitude
## 1 moderate  vigor   2.103167  5 0.090000  0.7230437  moderate
## 2 moderate   walk  -2.661855  5 0.090000 -0.9151138     large
## 3    vigor   walk -11.095177  5 0.000312 -3.8143891     large
```

We have enough information to reject/retain the null hypothesis (`*t* - ; *p* =`) with the corresponding effect size (`Hedge's g`) for pairwise contrasts. Specifically, we can reject the null $H1_0$ ($W_\mu = V_\mu$), which assumed the amount of time spent on walking and vigorous physical activity would be statistically equivalent.

> Note that it is **not** recommended to run post-hoc tests if the ANOVA was not significant, unless one has good theoretical justification for doing so. The post-hoc tests described above are for illustrating how to run and report said tests (*if* the ANOVA had been significant).

Visualizing data using boxplots

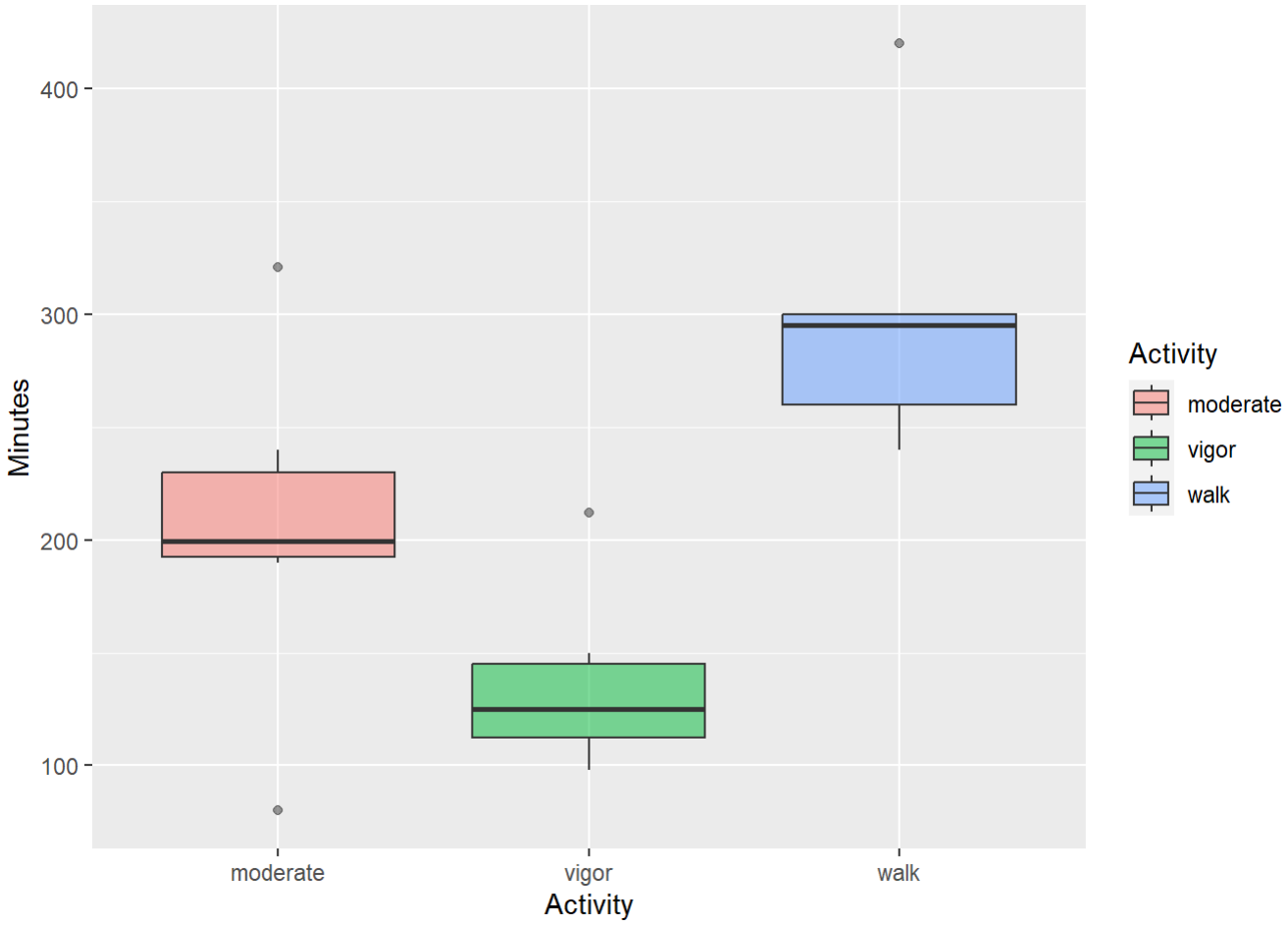
We can depict the distributions visually using a *boxplot*. Boxplots are a type of graphical representation used in statistics to display the distribution of a dataset. The benefits of using boxplots include:

- Boxplots show the central tendency (median) and dispersion (quartiles) of the data.
- Easily identify 'outliers', which are observations that fall outside of the expected range of values.
- Compare multiple datasets by having boxplots next to each other, which can help to quickly identify differences and aid in interpretation.

Boxplots can display a lot of information in a compact space, making them a useful tool for exploring quantitative data across factorial levels.

We will construct a boxplot after loading the `ggplot2` package.

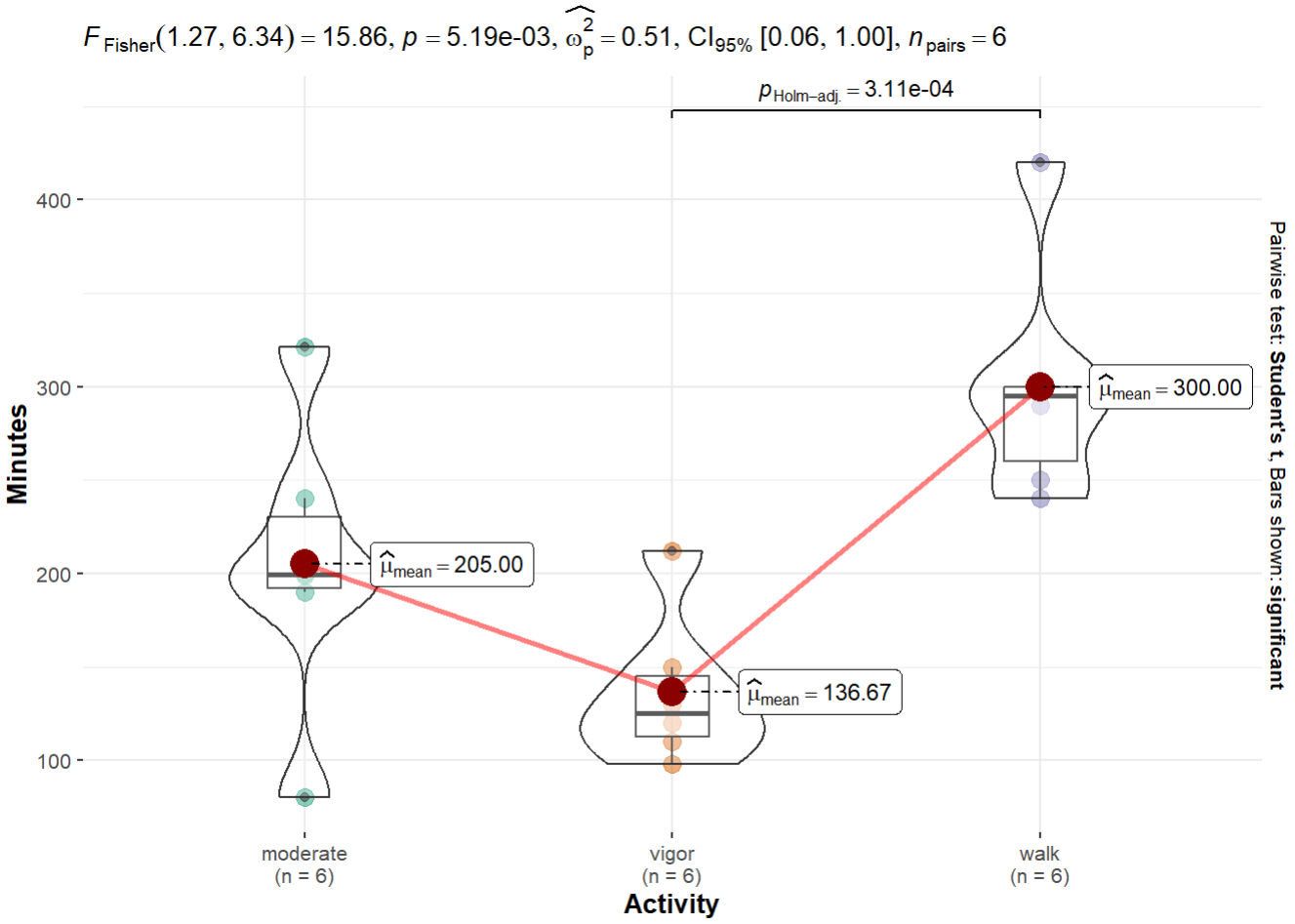
```
# Load package
require('ggplot2')

# Draw plot
ggplot(df1,
       aes(x=Activity,y=Minutes,fill=Activity))+
       geom_boxplot(alpha=.5)
```



It becomes easy to see how walking is the most frequently engaged physical activity (blue box), followed by moderate (red) and vigorous (green) activity durations.

A more detailed plot with ANOVA and post-hoc test outcomes using the `ggstatsplot` package.

```
ggstatsplot::ggwithinstats(
  data=df1,
  x=Activity,y=Minutes,bf.message = F)
```



We have the output of the significant ANOVA near the top of panel (note that *p* = 5.19e-03 is the same as *p* = .00519, which is less then the conventional threshold of *p* = .05). The effect size reported for the ANOVA is 'omega-squared' (ω^2) with the 95% confidence interval of the observed effect.

We can see that the average amount of walking (300 minutes) was significantly higher (*p* = 3.11e-04, or *p* = 0.000311) then the average amount of vigorous activity (136.7 minutes).

> Note that the Fisher-corrected ANOVA *did* reach significance, whereas our earlier repeated-measures ANOVA did not. This is because Fisher's ANOVA assumes that the groups being compared are *independent* (which our samples were not). The error term correction used in our ANOVA calculation is not applied in Fisher's ANOVA so any dependencies between the groups remain unaccounted for.

Conclusion

- A one-way/independent ANOVA tests whether $k \geq 3$ independent group means are statistically equivalent to one another.
- ANOVAs should be run across normally distributed data that share homogeneous variances and balanced across conditions. *Not* meeting any of this criteria will lead to inflated Type-1 error (false positive rate).
- Just as there are independent and repeated t-tests to respectively estimate differences between independent and paired samples, ANOVAs can test between independent and/or across repeated groups.

- If an ANOVA rejects the null hypothesis, we run post-hoc tests to localize the source of the difference

ANOVAs, alongside regressions and t-tests, are the most common statistical approaches within Psychological and social science. When underlying assumptions are met, ANOVAs are powerful and flexible tools for detecting meaningful statistical differences between multiple group means.

Teaching Evaluation

- Familiarize yourself with installing and loading packages on RStudio Cloud.
- Familiarize yourself with the `script` and `console` windows in RStudio. You can write out your code/comments in your script, then run them on the console individually or collectively.
    - If you are unaware of the difference between `script` and `console`, or would like a refresher on the different R Studio windows, you can watch this informative video by R-tutorials (https://www.youtube.com/watch?v=Q3NxsSRxKek).

To install and load the `tidyverse` and `rstatix` packages, run the following code on the console window:

 # You have to install the packages only once
install.packages('tidyverse')
install.packages('rstatix')

 # Afterwards you only have to call the packages by using the `require()` function
require('tidyverse')
require('rstatix')

 # The packages should be loaded and ready to go

- Once you have loaded the packages, please examine the data on the next slide. For this week's TE, you have to replicate the series of steps discussed this week. This includes:

1. Creating a tidy dataframe with the values provided below (hint: all levels of a factor should be under a similar column - see the beginning of the document)
2. Run a repeated measures ANOVA and report whether the null hypothesis is retained ($p > .05$) or rejected ($p < .05$)
3. If you observe a significant ANOVA effect, run post-hoc tests and report which pairwise contrasts were significant

Solutions to the above are available in the current slides. Copy & paste your code, and the output, onto a word/pdf document and submit in the dropbox.

Data for teaching evaluation

The average wages for 5 industries across the years 2016, 2017 and 2018 were taken from the Fijian government website (https://www.statsfiji.gov.fj/statistics/social-statistics/employment-statistics44.html). Assuming the data for each year represents an independent sample, test whether the average wage across the following industries were statistically equivalent across the years 2016, 2017, and 2018. Interpret any significant effects using post-hoc tests.

	yr2016	yr2017	yr2018
$Agriculture$	9656.07	8187.40	9781.11
$Mining$	11798.02	9226.97	11107.74
$Manufacturing$	8078.75	6788.68	8624.29
$Construction$	9027.22	7867.25	10151.98
$Education$	10050.43	7099.13	8970.06

This can be re-written as

	yr2016	yr2017	yr2018
$Industry_1$	9656.07	8187.40	9781.11
$Industry_2$	11798.02	9226.97	11107.74
$Industry_3$	8078.75	6788.68	8624.29
$Industry_4$	9027.22	7867.25	10151.98
$Industry_5$	10050.43	7099.13	8970.06