Introduction to Statistics with R

Session R05: ANOVA

Marvin Schmitt

The example data set

This example data set contains scores on a sustainability scale for three groups, each with a distinct diet: meat-eating (`meat`), vegetarian (`vegetarian`) and plant-based (`vegan`). Furthermore, each participant's self-reported gender is denoted.

Tip: Use the parameter `stringsAsFactors` to interpret character strings as factors automatically:

```
df = read.csv("R05_notes_dataset.csv", stringsAsFactors = TRUE)
str(df)
```

```
## 'data.frame':    150 obs. of  4 variables:
##  $ ID            : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ diet          : Factor w/ 3 levels "meat","vegan",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ gender        : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ sustainability: int  18 NA 12 14 NA 9 19 9 22 10 ...
```

We can change the **ordering** of the **factor levels**:

```r
df$diet = factor(df$diet, levels=c("vegan", "vegetarian", "meat"))
```

Missing data

In the `R` language, missing data are denoted as `NA` (not available). That's a dedicated symbol. It's not to be confused with `NaN` (not a number) that describes impossible data (e.g. from divison by 0).

```
head(df)
```

```
##   ID diet gender sustainability
## 1  1 meat   male             18
## 2  2 meat   male             NA
## 3  3 meat   male             12
## 4  4 meat   male             14
## 5  5 meat   male             NA
## 6  6 meat   male              9
```

Some functions do not work properly when `NA`s are present:

```
mean(df$sustainability)
```

```
## [1] NA
```

In many cases, these functions implement the argument `na.rm=TRUE` (= remove `NA` values):

```
mean(df$sustainability, na.rm=TRUE)
```

```
## [1] 15.6338
```

There are different strategies for missing values:

- Delete the data
  - Delete the entire observation (row)
  - Exclude the observation only from those analyses where the missing value would be required
- Substitute the missing value with a *typical* value, like ...
  - the mean $\bar{x}$
  - the median
  - the mode
  - ...
- Copy the last observation
- Estimate the missing value
  - Regression
  - Multiple imputation
  - ...

Introductory reading: https://doi.org/10.4097/kjae.2013.64.5.402

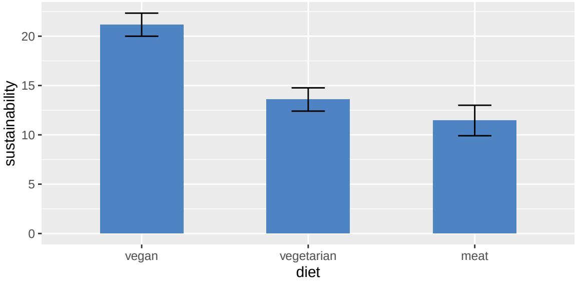
We can delete all rows that contain **any** missing data cell with `na.omit()`. That's simple but wasteful in applications where data are valuable.

```
df = na.omit(df)
```

Visualization: one-way ANOVA

For a one-way ANOVA (one grouping variable), we can simply use a bar plot with error bars:

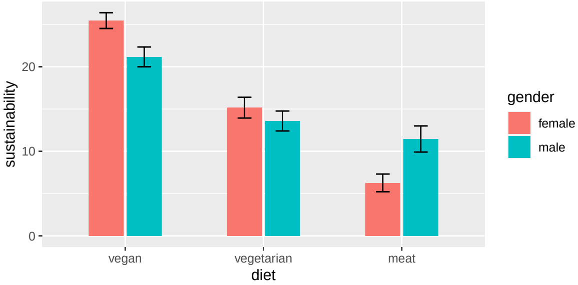
```
df %>% filter(gender=="male") %>%
  ggplot(., aes(y=sustainability, x = diet)) +
  stat_summary(fun = mean, geom = "bar", width=0.5, fill="#4E84C4") +
  stat_summary(fun.data = mean_se, geom = "errorbar", width=0.2)
```



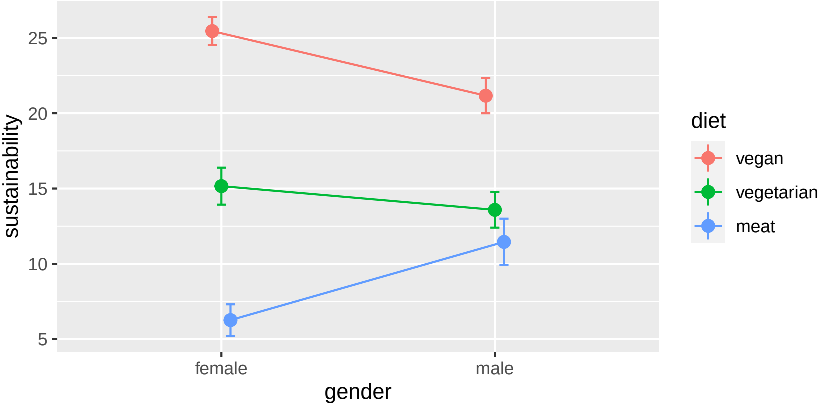
Visualization: Two-way ANOVA

For two factors, we can use a bar plot with different `fill`s. The bars and error bar need to be *dodged* a little:

```
ggplot(df, aes(y=sustainability, x = diet, fill=gender)) +
stat_summary(fun=mean, geom="bar", width=0.5, position=position_dodge(0.55)) +
stat_summary(fun.data=mean_se, geom="errorbar", width=.2,
             position=position_dodge(.55))
```



```
ggplot(df, aes(y=sustainability, x = gender, group=diet, color=diet)) +
  stat_summary(fun = mean, geom = "pointrange", position=position_dodge(0.1))+
  stat_summary(fun=mean, geom="line", position=position_dodge(0.1)) +
  stat_summary(fun.data = mean_se, geom = "errorbar", width=0.1,
               position=position_dodge(0.1))
```



Numerical ANOVA computation

The afex package implements convenient functions to compute ANOVAs. We will use the function aov_ez():

```
library(afex)
aov_ez(id = "ID",                    # identifier for subjects
       dv = "sustainability",  # dependent variable (y)
       between = "gender",      # between factor (group)
       data = df)                    # data frame
```

```
## Anova Table (Type 3 tests)
##
## Response: sustainability
##   Effect     df   MSE    F   ges p.value
## 1 gender 1, 140 72.35 0.03 <.001    .869
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '+' 0.1 ' ' 1
```

One-way ANOVA: Effect of gender

```r
m = aov_ez(id = "ID", dv = "sustainability",
           between = "gender", data = df)
```

```
## Contrasts set to contr.sum for the following variables: gender
```

```r
summary(m)
```

```
## Anova Table (Type 3 tests)
##
## Response: sustainability
##        num Df den Df   MSE      F        ges Pr(>F)
## gender      1    140 72.35 0.0273 0.00019465 0.8691
```

One-way ANOVA: Effect of `diet`

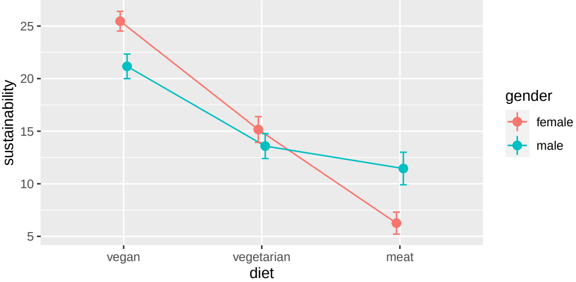
```r
m = aov_ez(id = "ID", dv = "sustainability",
           between = "diet", data = df)
summary(m)
```

```
## Anova Table (Type 3 tests)
##
## Response: sustainability
##      num Df den Df    MSE      F     ges    Pr(>F)
## diet      2    139 36.857 67.935 0.49431 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

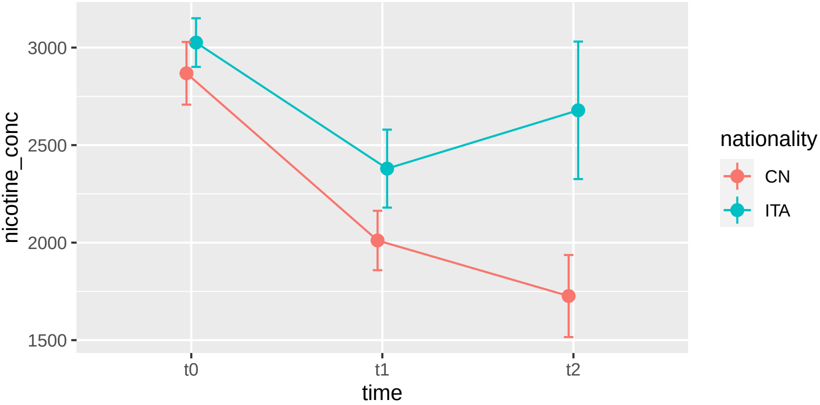
Two-way ANOVA: Effect of `gender` and `diet`

```
aov_ez(id = "ID", dv = "sustainability",
       between = c("gender", "diet"), data = df) %>% summary()
```

```
## Anova Table (Type 3 tests)
##
## Response: sustainability
##             num Df den Df    MSE       F     ges    Pr(>F)
## gender           1    136 33.591  0.0534 0.00039 0.8176617
## diet             2    136 33.591 74.0283 0.52122 < 2.2e-16 ***
## gender:diet      2    136 33.591  8.1932 0.10753 0.0004368 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```r
df_nicotine = read.csv("R05_notes_dataset_nicotine.csv", stringsAsFactors=TRUE)
```



Within-subject ANOVA for factor `time`:

```r
aov_ez(id = "ID", dv = "nicotine_conc",
       within = "time", data = df_nicotine) %>% summary()
```

```
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##                Sum Sq num Df Error SS den Df F value    Pr(>F)
## (Intercept) 683527262      1 38774274     39 687.506 < 2.2e-16 ***
## time         17628370      2 64973788     78  10.581 8.593e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
##      Test statistic   p-value
## time        0.72909 0.0024707
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
##
##       GG eps Pr(>F[GG])
## time 0.78684  0.0003502 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         HF eps   Pr(>F[HF])
## time 0.8142253 0.0002922006
```

```r
aov_ez(id = "ID", dv = "nicotine_conc", between="nationality",
       within = "time", data = df_nicotine) %>% summary()
```

```
##
## Univariate Type III Repeated-Measures ANOVA Assuming Sphericity
##
##                     Sum Sq num Df Error SS den Df  F value    Pr(>F)
## (Intercept)      674309848      1 31945156     38 802.1177 < 2.2e-16 ***
## nationality        6829118      1 31945156     38   8.1235 0.0070220 **
## time              13995826      2 61796338     76   8.6064 0.0004274 ***
## nationality:time   3177450      2 61796338     76   1.9539 0.1487760
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##
## Mauchly Tests for Sphericity
##
##                  Test statistic   p-value
## time                    0.75397 0.0053836
## nationality:time        0.75397 0.0053836
##
##
## Greenhouse-Geisser and Huynh-Feldt Corrections
##  for Departure from Sphericity
##
##                   GG eps Pr(>F[GG])
## time             0.80255   0.001178 **
## nationality:time 0.80255   0.158440
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                     HF eps   Pr(>F[HF])
```