1-way ANOVA

- deals w/ categorical explanatory variable (usually w/ more than 2 values – which group?) and quantitative response (enables our talking about means $\mu_1, \ldots$)
- key statistic F-statistic

ANOVA table

df1 = "degrees of freedom for groups" = (# of groups) − 1

df2 = "degrees of freedom for error/residual" = (total # of cases from all samples) − (# of groups)

If n = # of cases, drawn from I groups, then $df2 = n - I$.

Note: $(n - I) + (I - 1) = n - 1 = df_{Total}$

An example of ANOVA on a small dataset

Thomas Scofield

November 22, 2021

First, an example of ANOVA on a small data set

For the data given below, I want to create a data frame:

Group A: 15, 18, 17

Group B: 14, 11, 13

Group C: 16, 17, 19, 17

To do so, I'll combine the `c()` command with `data.frame()`:

```
myDat <- data.frame(grp = c("A","A","A","B","B","B","C","C","C","C"),
                    vals = c(15,18,17,14,11,13,16,17,19,17))
myDat
```

```
##    grp vals
## 1    A   15
## 2    A   18
## 3    A   17
## 4    B   14
## 5    B   11
## 6    B   13
## 7    C   16
## 8    C   17
## 9    C   19
## 10   C   17
```

Now, to carry out ANOVA calculations, can do as usual

```
anova(lm( vals ~ grp, data=myDat ))
```

```
## Analysis of Variance Table
##
## Response: vals
##           Df Sum Sq Mean Sq F value   Pr(>F)
## grp        2 40.017 20.0083   9.945 0.009001 **
## Residuals  7 14.083  2.0119
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Some reminders on R commands for bivariate quantitative data

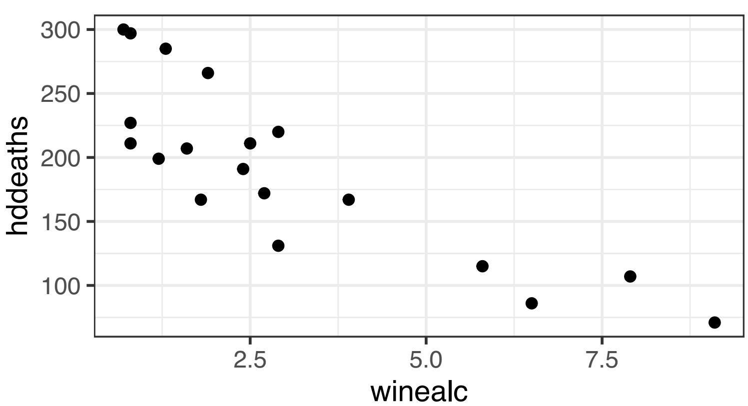
I need some data to use for demonstrations, and choose a data set I need to import using `read.csv()`:

```
hdAndWine <- read.csv("http://scofield.site/teaching/data/csv/heartDiseaseDeathsAndWine.csv")
hdAndWine
```

```
##           country winealc hddeaths
## 1       Australia     2.5      211
## 2     Netherlands     1.8      167
## 3         Austria     3.9      167
## 4     New Zealand     1.9      266
## 5         Belgium     2.9      131
## 6          Norway     0.8      227
## 7          Canada     2.4      191
## 8           Spain     6.5       86
## 9         Denmark     2.9      220
## 10         Sweden     1.6      207
## 11        Finland     0.8      297
## 12    Switzerland     5.8      115
## 13         France     9.1       71
## 14 United Kingdom     1.3      285
## 15        Iceland     0.8      211
## 16  United States     1.2      199
## 17        Ireland     0.7      300
## 18   West Germany     2.7      172
## 19          Italy     7.9      107
```

The `country` column isn't so much a variable as it is an identifier. (No variable can be very interesting if all its values occur with frequency 1, anyway.) That leaves us with two quantitative variables, `winealc` and `hdDeath`. Taking `winealc` as the explanatory variable, we can produce a scatterplot:

```
gf_point(hddeaths ~ winealc, data=hdAndWine)
```



It appears a negative linear relationship is appropriate. That makes the **correlation** meaningful:

```
gf_point(hddeaths ~ winealc, data=hdAndWine)
gf_point(winealc ~ hddeaths, data=hdAndWine)
    # try these, and note the same value either way
```

You can obtain the **slope** and **intercept** for the least-squares regression line:

```
lm(hddeaths ~ winealc, data=hdAndWine)
lm(winealc ~ hddeaths, data=hdAndWine)
```

```
    # Try these. They do NOT give the same slope/intercept
```

And, you can add the regression line to the scatterplot:

```
gf_point(hddeaths ~ winealc, data=hdAndWine) %>% gf_lm(type="lm")
```