DATA2002

ANOVA

Garth Tarr

THE UNIVERSITY OF SYDNEY

What is ANOVA?

t-test revision

The general ANOVA decomposition

What is ANOVA?

What does ANOVA stand for?

- The term **ANOVA** is an acronym/abbreviation for/of the term "**Analysis of Variance**".
- The term "variance", as well as the ANOVA procedure, is mainly due to Fisher from the 1920's, in particular the book "Statistical Methods for Research Workers" (something of a classic text, Fisher (1925)).

Yeah, but what is "Analysis of Variance"?

- In its (perhaps) "simplest" form, Analysis of Variance is a generalisation of a *two-sided* two-sample t-test to 3 or more samples.
- Which two-sample t-test though?

t-test revision

Two-sample t-tests

- There are (at least) 3 different procedures which might be referred to as a "two-sample t-test":
- the *Paired* (two-sample) t-test;
- the *Welch* test (unequal variances two-independent-sample t-test).
- the *Classical* or *Pooled* two-(independent)-sample (equal variances) t-test;
- They all take the form

$$\frac{\bar{X} - \bar{Y}}{\mathrm{SE}(\bar{X} - \bar{Y})} \; .$$

- They only differ in how the standard error is computed.
- We briefly review these.

Paired (two-sample) t-test

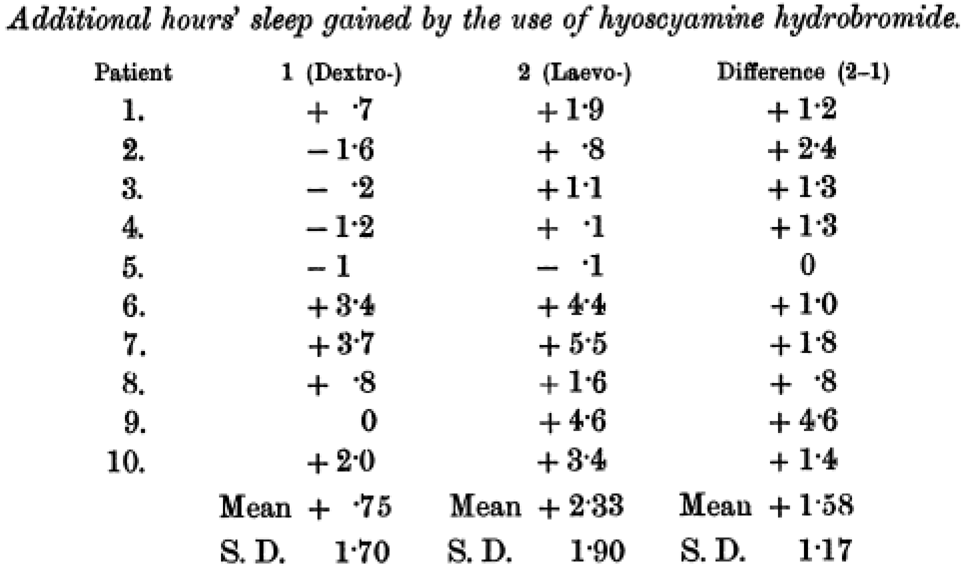
- For the **paired** (two-sample) t-test, it is assumed the differences $D_1 = X_1 - Y_1, \ldots, D_n = X_n - Y_n$ are iid normal with variance σ_D^2.
- Under these conditions $\bar{D} = \bar{X} - \bar{Y}$ is normal with variance $\frac{\sigma_D^2}{n}$ where n is the common sample size.
- σ_D^2 is estimated using S_D^2, the *sample variance of the differences*, giving a standard error of

$$\mathrm{SE}(\bar{X} - \bar{Y}) = \frac{S_D}{\sqrt{n}} .$$

- The test statistic is **exactly** distributed as t_{n-1} under H_0.
- This is just a **one-sample t-test** applied to the differences.

Sleep data

- The "classic" example where the t-test was "invented", from "Student's" 1908 *Biometrika* paper "The probable error of a mean":



Additional hours' sleep gained by the use of hyoscyamine hydrobromide.

Patient	1 (Dextro-)	2 (Laevo-)	Difference (2–1)
1.	+ ·7	+1·9	+1·2
2.	−1·6	+ ·8	+2·4
3.	− ·2	+1·1	+1·3
4.	−1·2	+ ·1	+1·3
5.	−1	− ·1	0
6.	+3·4	+4·4	+1·0
7.	+3·7	+5·5	+1·8
8.	+ ·8	+1·6	+ ·8
9.	0	+4·6	+4·6
10.	+2·0	+3·4	+1·4
	Mean + ·75	Mean +2·33	Mean +1·58
	S. D. 1·70	S. D. 1·90	S. D. 1·17

- It is available in R as the object `sleep`

```
sleep
```

```
##    extra group ID
## 1    0.7     1  1
## 2   -1.6     1  2
## 3   -0.2     1  3
## 4   -1.2     1  4
## 5   -0.1     1  5
## 6    3.4     1  6
## 7    3.7     1  7
## 8    0.8     1  8
## 9    0.0     1  9
## 10   2.0     1 10
## 11   1.9     2  1
## 12   0.8     2  2
## 13   1.1     2  3
## 14   0.1     2  4
## 15  -0.1     2  5
## 16   4.4     2  6
## 17   5.5     2  7
## 18   1.6     2  8
## 19   4.6     2  9
## 20   3.4     2 10
```

Source: Student (1908). Can you see the typo in the original paper?

- Let's try the "default" t-test command:

```
t.test(extra ~ group, data = sleep)
```

```
##
##      Welch Two Sample t-test
##
## data:  extra by group
## t = -1.8608, df = 17.776, p-value = 0.07939
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.3654832  0.2054832
## sample estimates:
## mean in group 1 mean in group 2
##            0.75            2.33
```

- What? Welch test?

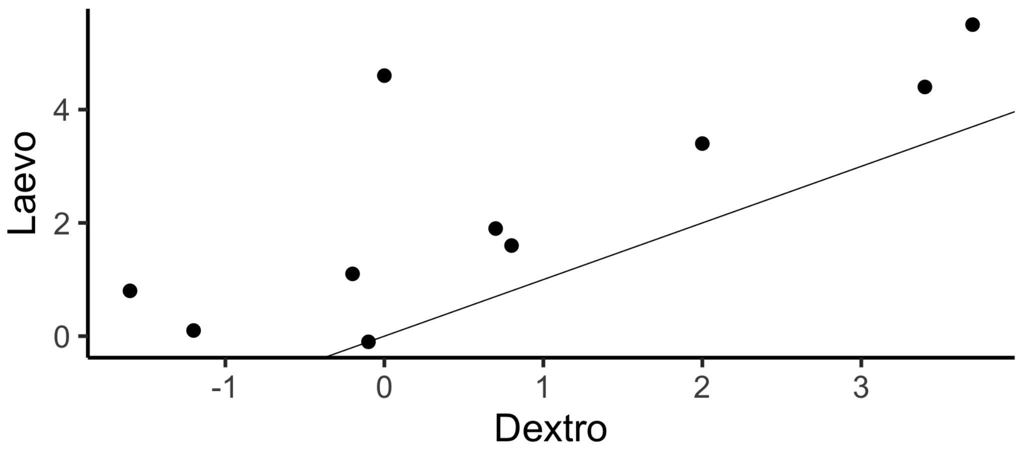
```r
library(tidyverse)
sleep_recode = sleep %>% dplyr::mutate(
  group = forcats::fct_recode(group,
                              Dextro = "1",
                              Laevo = "2"
  )
)
head(sleep_recode, n = 3)
```

```
##   extra  group ID
## 1   0.7 Dextro  1
## 2  -1.6 Dextro  2
## 3  -0.2 Dextro  3
```

```r
sleep_wide = tidyr::spread(sleep_recode,
                           key = group,
                           value = extra)
head(sleep_wide, n = 3)
```

```
##   ID Dextro Laevo
## 1  1    0.7   1.9
## 2  2   -1.6   0.8
## 3  3   -0.2   1.1
```

```r
ggplot(sleep_wide,
       aes(x = Dextro, y = Laevo)) +
  geom_point(size = 5) +
  geom_abline(slope = 1, intercept = 0) +
  theme_classic(base_size = 32)
```



- There is a clear trend: *samples are not independent*
- Most points "above" the $y = x$ line: suggests the y's are bigger than the x's.

- The **paired** t-test (two-sided, unless a direction was anticipated beforehand) gives:

```
t.test(extra ~ group, data = sleep, paired = TRUE)
```

```
##
##      Paired t-test
##
## data:  extra by group
## t = -4.0621, df = 9, p-value = 0.002833
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -2.4598858 -0.7001142
## sample estimates:
## mean of the differences
##                   -1.58
```

- Note the much smaller p-value!

Welch (unequal variances two-independent-sample) t-test

- For the Welch test it is only assumed that each sample is normal, with possibly different variances σ_X^2 and σ_Y^2 and different means, and all random variables are independent.
- Under these conditions $\bar{X} - \bar{Y}$ is normal with variance

$$\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\,.$$

    - The standard error is obtained by simply plugging in sample variances as estimators of population variances:

$$\mathrm{SE}(\bar{X} - \bar{Y}) = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}\,.$$

    - The test statistic is **approximately** $t_{d^*(m,n,\sigma_X,\sigma_Y)}$ under H_0, for a known function $d^*(\dots)$.
    - p-value is computed by plugging sample sd's into $d^*(\dots)$.

Lengths of New Zealand rivers

- The file `nzrivers.txt` has lengths (in km) of rivers on the South Island of New Zealand

```
nzrivers = read_tsv("http://www.statsci.org/data/oz/nzrivers.txt")
glimpse(nzrivers)
```

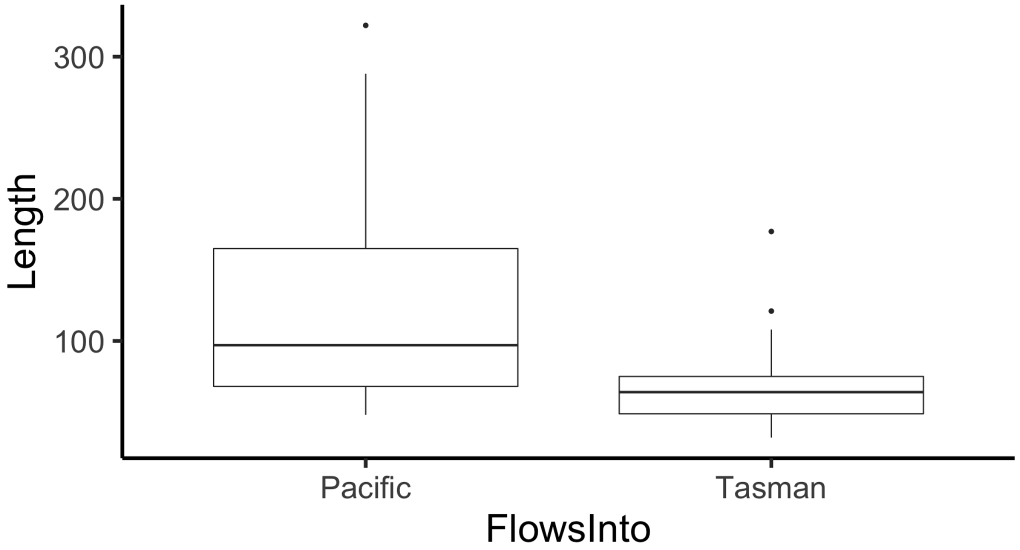
```
## Rows: 41
## Columns: 3
## $ River     <chr> "Clarence", "Conway", "Waiau", "Hurunui"…
## $ Length    <dbl> 209, 48, 169, 138, 64, 97, 161, 95, 145,…
## $ FlowsInto <chr> "Pacific", "Pacific", "Pacific", "Pacifi…
```

```
nzrivers %>%
  group_by(FlowsInto) %>%
  summarise(xbar = mean(Length),
            med = median(Length))
```

```
## # A tibble: 2 × 3
##   FlowsInto  xbar   med
##   <chr>     <dbl> <dbl>
## 1 Pacific   131.     97
## 2 Tasman     67.7    64
```

- If we wanted to test that the mean difference here was signficant, we see there is a big difference in variability between the two (and possibly skewness!)

```
ggplot(nzrivers,
       aes(x = FlowsInto, y = Length)) +
  geom_boxplot() +
  theme_classic(base_size = 32)
```



```r
welch = t.test(Length ~ FlowsInto, data = nzrivers)
welch
```

```
##
##     Welch Two Sample t-test
##
## data:  Length by FlowsInto
## t = 3.2632, df = 23.477, p-value = 0.003358
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   23.28176 103.67039
## sample estimates:
## mean in group Pacific  mean in group Tasman
##             131.15789              67.68182
```

- Ths is the default two-sample t-test in R.
- **Note** the "degrees of freedom", here, roughly 23.5.

Classical two-(independent)-sample (equal variances) t-test

- The "Classical" two-sample t-test assumes the same as the Welch test with the *extra* assumption that the two population variances $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ **are equal**.
- Under these conditions $\bar{X} - \bar{Y}$ is normal with variance $\sigma^2 \left(\frac{1}{m} + \frac{1}{n}\right)$ (for possibly different sample sizes m and n).
- σ^2 is estimated using the **pooled variance estimator**

  $$S_p^2 = \frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}$$

  (a **weighted average** of the two sample variances) giving a standard error of

  $$\mathrm{SE}(\bar{X} - \bar{Y}) = S_p\sqrt{\frac{1}{m} + \frac{1}{n}}\,.$$

- The test statistic is **exactly** distributed as t_{m+n-2} under H_0.

```r
classical = t.test(Length ~ FlowsInto, data = nzrivers, var.equal = TRUE)
classical
```

```
## 
##     Two Sample t-test
## 
## data:  Length by FlowsInto
## t = 3.4391, df = 39, p-value = 0.001403
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   26.14338 100.80878
## sample estimates:
## mean in group Pacific  mean in group Tasman 
##             131.15789              67.68182 
```

- Gives similar results, but look at the **degrees of freedom**; much bigger than the Welch test, and gives a bigger statistic, smaller p-value and narrower confidence interval
  - **underestimates the standard error of the mean difference**.

How serious is the equal variance assumption?

- The assumption of equal variance is quite crucial for the validity of the Classical test.
- In particular, quite strange things can happen if
  - the population variances are different;
  - the sample sizes are very different;
- Consider the simulation on the next slide where the **smaller sample** has a **bigger variance**

Simulation

```
B = 10000
pval.Classical = pval.Welch = vector(length = B)
set.seed(123)
for(i in 1:B){
  x = rnorm(100, sd = 1)          # both samples have the same mean
  y = rnorm(20, sd = 3)           # smaller sample has bigger variance
  pval.Welch[i] = t.test(x, y)$p.val
  pval.Classical[i] = t.test(x, y, var.equal = TRUE)$p.val
}
mean(pval.Welch < .05)        # Rejects about 5% of the time.
```

```
## [1] 0.0487
```

```
mean(pval.Classical < .05)    # Rejects far too often!!
```

```
## [1] 0.2887
```

Some comments

- Of the 3 different two-sample t-tests, the Classical test is the one that requires the *most assumptions*:
- One could almost "do away" with it:
    - a Welch test could always be used instead (Welch test is the default in R);
    - the paired test can also be used if the sample sizes are equal!
        - In that case the differences are still iid normal!
        - The paired test suffers a *minor* loss of power (due to the lower degrees of freedom only) but is robust against positive correlation.
- **But** the Classical test is the one that generalises to ANOVA.
- We must always be aware of these key *assumptions*:
    - independence between samples;
    - equal variance.

The general ANOVA decomposition

ANOVA (in the case of g groups)

1. **Hypotheses:** $H_0\colon\ \mu_1 = \mu_2 = \ldots = \mu_g$ vs H_1: at least one $\mu_i \neq \mu_j$.
2. **Assumptions:** Observations are independent within each of the g samples. Each of the g populations have the same variance, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_g^2 = \sigma$. Each of the g populations are normally distributed (or the sample sizes are large enough such that you can rely on the central limit theorem).
3. **Test statistic:** $T = \frac{\text{Treatment Mean Sq}}{\text{Residual Mean Sq}}$. Under H_0, $T \sim F_{g-1,\, N-g}$ where g is the number of groups.
4. **Observed test statistic:** t_0.
5. **p-value:** $P(T \geq t_0) = P(F_{g-1,\, N-g} \geq t_0)$. Note: always looking in the upper tail.
6. **Decision:** If the p-value is less than α we reject the null hypothesis and conclude that the population mean of at least one group is significantly different to the others. If the p-value is larger than α we do not reject the null hypothesis and conclude that there is no significant difference between the population means.

The normal model

- We model y_{ij} (for each $j = 1, 2, \ldots, n_i$ and $i = 1, 2, \ldots, g$) as the value taken by a random variable

  $$Y_{ij} \sim N(\mu_i, \sigma^2)\,,$$

  and that all random variables are independent.
- Thus we have g different iid samples, the sample for group i (of size n_i) being iid $N(\mu_i, \sigma^2)$.
  - In other words, for each $i = 1, 2, \ldots, g$, $Y_{i1}, \ldots, Y_{in_i}$ are iid $N(\mu_i, \sigma^2)$ random variables.

The dreaded "dot" notation

- When working with double subscripts it is convenient to introduce the **dot** notation:
  - replacing either (or both) subscript(s) with a dot means **adding** over that/those subscript(s);
  - replacing either (or both) subscript(s) with a dot **and writing a bar over the letter** means **averaging** over that/those subscript(s).
- For example:
  - total for sample i is $\sum_{j=1}^{n_i} y_{ij} = y_{i\bullet}$
  - average for sample i is $\frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_{i\bullet}$
  - grand total of all observations is $\sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} = y_{\bullet\bullet}$
  - overall average of all observations is $\frac{1}{N} \sum_{i=1}^{g} \sum_{j=1}^{n_i} y_{ij} = \bar{y}_{\bullet\bullet}$
    - here $N = n_1 + \ldots + n_g$ is the total number of observations.
  - Also, $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i\bullet})^2$ is the i-th group's sample variance.

The general ANOVA decomposition

- The "weighted average" decomposition introduced earlier for the two-sample t-test is a special case of a more general decomposition.
- It is most easily explained by considering the so-called **Total Sum of Squares**:

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\bullet\bullet})^2$$

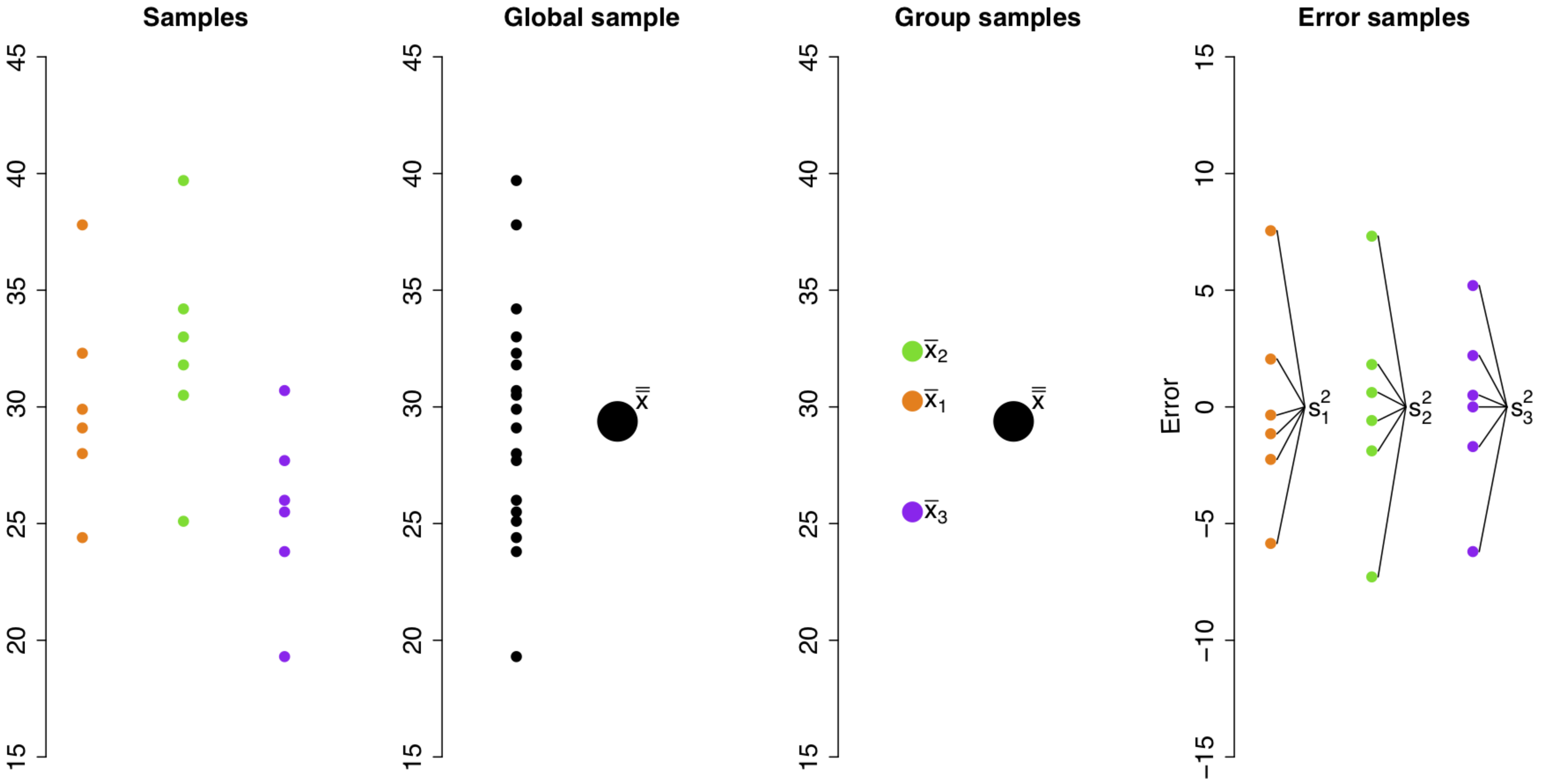
which is precisely $(N-1)$ times the *combined* sample variance of all the observations,

$$\hat{\sigma}_0^2 = \frac{\text{Total SS}}{N-1} = \frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\bullet\bullet})^2}{N-1}$$

- We start by adding and subtracting the group means inside the square, grouping and expanding:

$$
\begin{aligned}
\sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{\bullet\bullet})^2 &= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left[(y_{ij}-\bar{y}_{i\bullet})+(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})\right]^2 \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}\left[(y_{ij}-\bar{y}_{i\bullet})^2+2(y_{ij}-\bar{y}_{i\bullet})(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})+(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2\right] \\
&= \sum_{i=1}^{g}\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})^2+2\sum_{i=1}^{g}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})}_{=0}+\sum_{i=1}^{g}(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2\underbrace{\sum_{j=1}^{n_i}1}_{=n_i} \\
&= \underbrace{\sum_{i=1}^{g}\underbrace{\sum_{j=1}^{n_i}(y_{ij}-\bar{y}_{i\bullet})^2}_{=(n_i-1)s_i^2}}_{\text{sample variances}}+\underbrace{\sum_{i=1}^{g}n_i(\bar{y}_{i\bullet}-\bar{y}_{\bullet\bullet})^2}_{\text{sample means}} \\
&= \text{Residual SS}+\text{Treatment SS}
\end{aligned}
$$

The steps are not examinable.


Samples
Global sample
Group samples
Error samples
15
20
25
30
35
40
45
Error
−15
−10
−5
0
5
10
15
$\bar{\bar{x}}$
$\bar{x}_2$
$\bar{x}_1$
$\bar{x}_3$
s_1^2
s_2^2
s_3^2

Residual Sum of Squares; Residual Mean Square

- The first term, viewed as a random variable under the normal model, can be written as

$$\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2=\sum_{i=1}^{g}\underbrace{(n_i-1)S_i^2}_{\sim\sigma^2\chi^2_{n_i-1}}\sim\sigma^2\chi^2_{N-g}$$

  noting that $\sum_{i=1}^{g}(n_i-1)=N-g$. This is called the **Residual Sum of Squares**.

- Dividing by $N-g$ we obtain an unbiased estimator of σ^2, the generalisation of the pooled estimate of the variance, known as the **Residual Mean Square**:

$$\hat{\sigma}^2=\frac{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2}{N-g}\sim\left(\frac{\sigma^2}{N-g}\right)\chi^2_{N-g}\,.$$

Treatment Sum of Squares

- The full "random variable" version of the decomposition looks like

$$\underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{\bullet\bullet})^2}_{\sim\sigma^2\chi^2_{N-1}\ \text{under } H_0} = \underbrace{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_{i\bullet})^2}_{\sim\sigma^2\chi^2_{N-g}\ \text{always}} + \underbrace{\sum_{i=1}^{g} n_i(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2}_{\sim\ ????}$$

- When H_0 is true, the final term **must** have a $\sigma^2\chi^2_{g-1}$ distribution;
    - when the sample sizes $n_1 = \ldots = n_g = n$ are equal this is just $(g-1)$ times the sample variance of the iid normals $\sqrt{n}\bar{Y}_{1\bullet}, \ldots, \sqrt{n}\bar{Y}_{g\bullet}$ with variance σ^2, so this is correct in that case; in general this is a bit more complicated though.
- If the true group means are not all equal, this will tend to get bigger.
- This is the **Treatment Sum of Squares**.
- The ratio $\dfrac{\sum_{i=1}^{g} n_i(\bar{Y}_{i\bullet}-\bar{Y}_{\bullet\bullet})^2}{g-1}$ is the **Treatment Mean Square**.

Treatment? Huh?

- The term "Treatment" dates back to the beginnings of Analysis of Variance, where R.A. Fisher applied these techniques to agricultural trials, notably concerning fertiliser treatments.
- The **Treatment Sum of Squares** is the generalisation of the term $\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{1}{m}+\frac{1}{n}}}\right)^2$ in the analysis of the two-combined-sample variance.
    - It measures the variability of the sample means in a certain sense.

$$\text{Treatment Mean Square} = \frac{\text{Treatment Sum of Squares}}{g-1} = \frac{\sum_{i=1}^{g} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{g-1}$$

The "ratio of variances" test

- Continuing the analogy to the two-sample t-test, we can consider the ratio of variance estimates as a test statistic to test the null hypothesis

  $$H_0\colon \mu_1 = \mu_2 = \ldots = \mu_g$$

  against the alternative that they are not all equal.

- The estimate under the null hypothesis is just the "combined" sample variance

  $$\hat{\sigma}_0^2 = \frac{1}{N-1} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2 \,.$$

- The estimate under the alternative or "full model" is just the **Residual Mean Square**:

  $$\hat{\sigma}^2 = \frac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2}{N-g} \,.$$

The F statistic

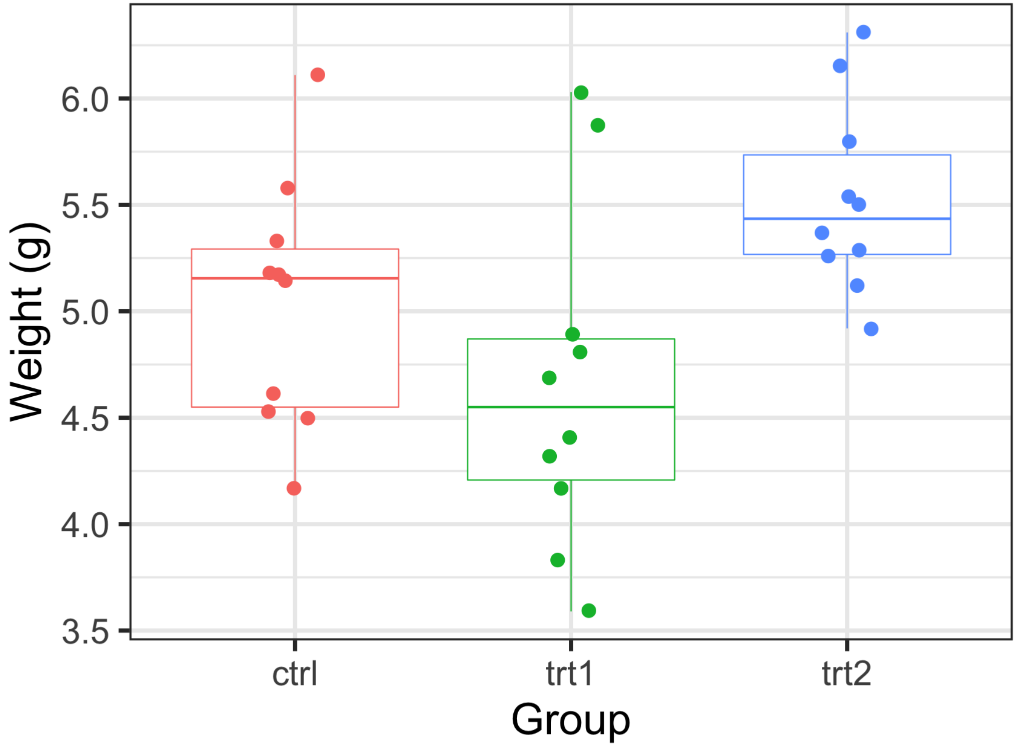
- It turns out that, a sensible test statistic considers the ratio of these two ways of estimating σ^2:

$$\begin{aligned}\frac{\text{Treatment Mean Square}}{\text{Residual Mean Square}} &= \frac{\sum_{i=1}^{g} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2/(g-1)}{\hat{\sigma}^2} \\ &= \frac{\sum_{i=1}^{g} n_i(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2/(g-1)}{\sum_{i=1}^{g}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_{i\bullet})^2/(N-g)} \\ &\sim \frac{\chi^2_{g-1}/(g-1)}{\chi^2_{N-g}/(N-g)} \quad \text{(both independent)} \\ &\sim F_{g-1,N-g} \quad \text{under } H_0.\end{aligned}$$

- the denominator is **always** an unbiased estimator of σ^2 regardless of whether H_0 is true or not
- the numerator is only an unbiased estimator of σ^2 if H_0 is true, otherwise **it tends to get bigger**.

The `PlantGrowth` data has results from an experiment to compare yields (as measured by dried weight of plants) obtained under a control and two different treatment conditions Dobson (1983; Table 7.1).

```r
# built into R, load it into the environment
data("PlantGrowth")
library(ggplot2)
ggplot(PlantGrowth,
       aes(y = weight, x = group,
           colour = group)) +
  geom_boxplot(coef = 10) +
  geom_jitter(width=0.1, size = 5) +
  theme_bw(base_size = 36) +
  theme(legend.position = "none") +
  labs(y = "Weight (g)",
       x = "Group")
```



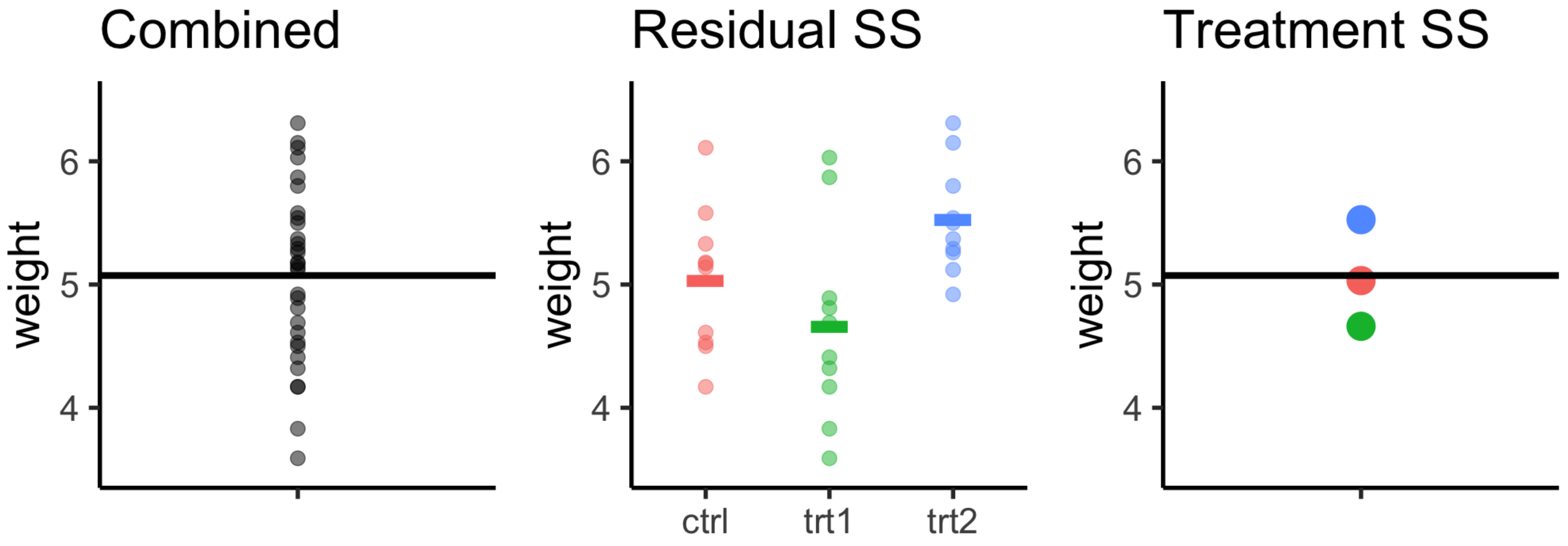
We want to compare the means of the **three** groups.

```r
p0 = ggplot(PlantGrowth, aes(y = weight, x = "")) +
  geom_point(alpha = 0.5, size = 4) +
  geom_hline(yintercept = mean(PlantGrowth$weight), lwd = 2) +
  theme_classic(base_size = 28) + theme(legend.position = "none") +
  coord_cartesian(ylim = c(3.5,6.5)) + labs(title = "Combined", x="")

p1 = ggplot(PlantGrowth, aes(y = weight, x = group, colour = group)) +
  geom_point(alpha = 0.5, size = 4) +
  stat_summary(aes(colour = group), fun = mean, geom = "point",
               size = 30, pch="-") +
  theme_classic(base_size = 28) + theme(legend.position = "none") +
  coord_cartesian(ylim = c(3.5,6.5)) + labs(title = "Residual SS", x = "")

p2 = ggplot(PlantGrowth, aes(y = weight, x = "")) +
  stat_summary(aes(colour = group), fun = mean, geom = "point",
               size = 8) +
  geom_hline(yintercept = mean(PlantGrowth$weight), lwd = 2) +
  theme_classic(base_size = 28) + theme(legend.position = "none") +
  coord_cartesian(ylim = c(3.5,6.5)) + labs(title = "Treatment SS", x = "")

gridExtra::grid.arrange(p0, p1, p2, nrow = 1)
```



Combined sample variance (estimate under the null hypothesis): $\hat{\sigma}_0^2 = \dfrac{1}{N-1} \sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{\bullet\bullet})^2$

Residual mean square (estimate under the alternative hypothesis): $\hat{\sigma}^2 = \dfrac{\sum_{i=1}^{g} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_{i\bullet})^2}{N-g}$

Treatment mean square: $\dfrac{\sum_{i=1}^{g} n_i (\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2}{g-1}$

Decomposition

```
PlantGrowth = PlantGrowth %>%
  mutate(overall_mean = mean(weight)) %>%
  group_by(group) %>%
  mutate(group_mean = mean(weight))
PlantGrowth %>% slice(1:2)
```

```
## # A tibble: 6 × 4
## # Groups:   group [3]
##   weight group overall_mean group_mean
##    <dbl> <fct>        <dbl>      <dbl>
## 1   4.17 ctrl          5.07       5.03
## 2   5.58 ctrl          5.07       5.03
## 3   4.81 trt1          5.07       4.66
## 4   4.17 trt1          5.07       4.66
## 5   6.31 trt2          5.07       5.53
## 6   5.12 trt2          5.07       5.53
```

```
N = nrow(PlantGrowth)
g = 3
```

Treatment mean square

group means vs overall mean

```
treat_ss = sum((PlantGrowth$group_mean -
                 PlantGrowth$overall_mean)^2)
treat_ms = treat_ss/(g-1)
c(treat_ss, treat_ms)
```

```
## [1] 3.76634 1.88317
```

Residual mean square

observations vs their group means

```
resid_ss = sum((PlantGrowth$weight -
                 PlantGrowth$group_mean)^2)
resid_ms = resid_ss/(N-g)
c(resid_ss, resid_ms)
```

```
## [1] 10.4920900  0.3885959
```

```
plant_anova = aov(weight ~ group, data = PlantGrowth)
plant_anova
```

```
## Call:
##    aov(formula = weight ~ group, data = PlantGrowth)
##
## Terms:
##                    group Residuals
## Sum of Squares   3.76634  10.49209
## Deg. of Freedom        2        27
##
## Residual standard error: 0.6233746
## Estimated effects may be unbalanced
```

```
summary(plant_anova)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## group        2  3.766  1.8832   4.846 0.0159 *
## Residuals   27 10.492  0.3886
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1. **Hypotheses:** H_0: $\mu_1 = \mu_2 = \mu_3$ vs H_1: at least one $\mu_i \neq \mu_j$ for $i \neq j$.
2. **Assumptions:** Observations are independent within each of the 3 samples. Each of the 3 populations are normally distributed with the common variance σ.
3. **Test statistic:** $T = \frac{\text{Treatment Mean Sq}}{\text{Residual Mean Sq}}$. Under H_0, $T \sim F_{g-1,\, N-g}$ where $g = 3$ is the number of groups.
4. **Observed test statistic:** $t_0 = \frac{1.88}{0.39} = 4.8$.
5. **p-value:** $P(T \geq 4.8) = P(F_{2,\, 27} \geq 4.8) = 0.0159$. Manually in R: `1-pf(4.8, 2, 27)`
6. **Decision:** As the p-value is less than α we reject the null hypothesis and conclude that the population mean of at least one group is significantly different to the others.

💬 Which are different? Ctrl vs Trt1? Ctrl vs Trt2? Trt1 vs Trt2?

Further reading

Larsen and Marx (2012) sections 12.1 and 12.2.

References

Dobson, A. J. (1983). *An introduction to statistical modelling*. London: Chapman & Hall.

Fisher, R. (1925). *Statistical methods for research workers*. Edinburgh Oliver & Boyd.

Larsen, R. J. and M. L. Marx (2012). *An Introduction to Mathematical Statistics and its Applications*. 5th ed. Boston, MA: Prentice Hall. ISBN: 978-0-321-69394-5.

Student (1908). "The probable error of a mean". In: *Biometrika* 6.1, pp. 1-25. DOI: 10.1093/biomet/6.1.1. URL: http://biomet.oxfordjournals.org/content/6/1/1.short.