# DATA2002

## ANOVA contrasts

Garth Tarr

THE UNIVERSITY OF SYDNEY

Contrasts

Confidence intervals

# Contrasts

# Beyond ANOVA: contrasts

- Recall our model: for $i = 1, 2, \ldots, g$ and $j = 1, 2, \ldots, n_i$,
- the $j$-th observation in the $i$-th sample is modelled as the value taken by

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

  and all random variables are assumed independent.

- We can rewrite this as:

$$Y_{ij} = \mu_i + \varepsilon_{ij},$$

  where $\varepsilon_{ij} \sim N(0, \sigma^2)$ is the error term.

- The ANOVA $F$-test is a test of the hypothesis

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

- If this hypothesis is "rejected", then what?
- Further analysis reduces to the study of **contrasts**

# Contrasts

- A **contrast** is a **linear combination** where the coefficients **add to zero**.
- In an ANOVA context, a contrast is a linear combination of **means**.
- We make the distinction between two kinds of contrast:
- **population contrasts**: contrasts involving the *population* group means i.e. the $\mu_i$'s;
- **sample contrasts**: contrasts involving the *sample* group means i.e. the $\bar{y}_{i\bullet}$'s and $\bar{Y}_{i\bullet}$'s.

For example, we might consider the *population* contrast

$$\mu_1 - \mu_2,$$

whose corresponding *observed* sample version is

$$\bar{y}_{1\bullet} - \bar{y}_{2\bullet},$$

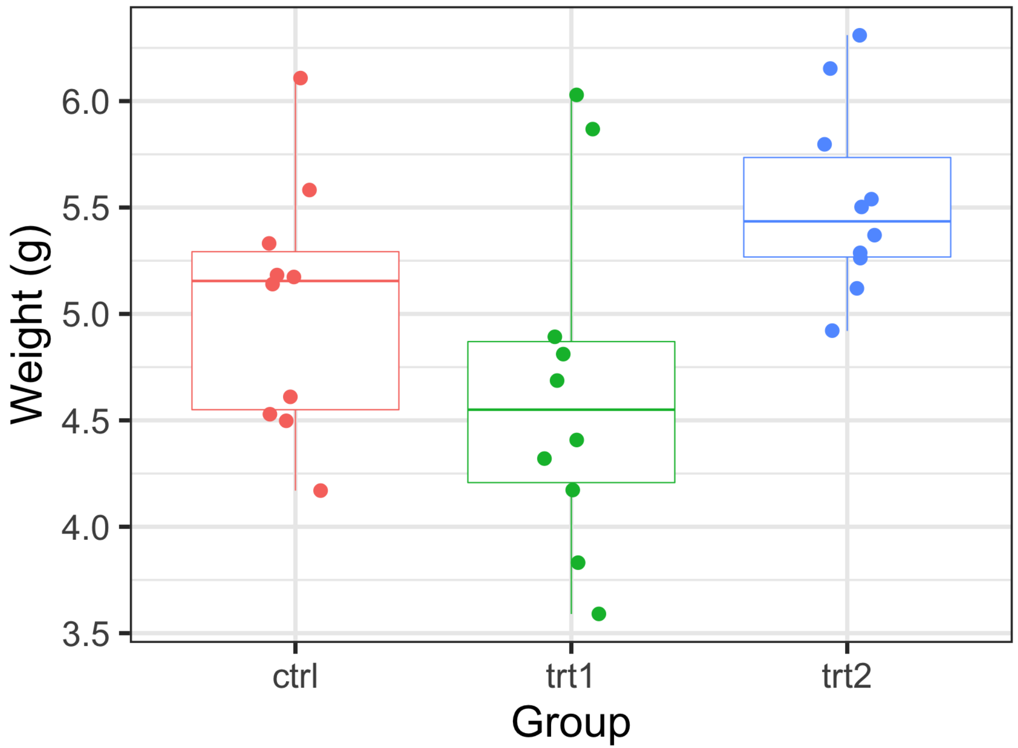
which is the observed value of the random variable

$$\bar{Y}_{1\bullet} - \bar{Y}_{2\bullet}.$$

# Plant growth data

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(tidyverse)
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: $H_0\colon\ \mu_1 = \mu_2 = \mu_3$.

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

The p-value, $P(F_{2,27} \geq 4.846) = 0.0159$ is less than 0.05, so we reject the null hypothesis at the 5% level of significance and conclude there is evidence to suggest that at least one of the groups has a significantly different mean yeild to the others.

**But which one?!?!**

- Is it `ctrl` vs `trt1`?
- Is it `ctrl` vs `trt2`?
- Is it `trt1` vs `trt2`?
- Or is it some other linear combination of the means that is different?

# Distribution of sample contrasts

- Any $c_1, \ldots, c_g$ with $c_\bullet = \sum_{i=1}^g c_i = 0$ defines a *sample contrast*

$$\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} \,.$$

- Under our *normal-with-equal-variances* model, this random variable has distribution given by

$$\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} \sim N\left(\sum_{i=1}^{g} c_i \mu_i,\ \sigma^2 \sum_{i=1}^{g} \frac{c_i^2}{n_i}\right) \,.$$

- The corresponding *population contrast* is the expected value of the (random) sample contrast.
- Conversely, the (observed) *sample contrast* $\displaystyle\sum_{i=1}^{g} c_i \bar{y}_{i\bullet}$ is an *estimate* of the corresponding *population contrast*;
- the (random) sample contrast $\displaystyle\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}$ is the corresponding *estimator*.

# Behaviour of contrasts under the null hypothesis

- Under the "ANOVA null hypothesis" $H_0\colon \mu_1 = \ldots = \mu_g (= \mu,\ \text{say})$,
- all population contrasts are zero:

$$\sum_{i=1}^{g} c_i \mu_i = \sum_{i=1}^{g} c_i \mu = \mu \sum_{i=1}^{g} c_i = 0\,;$$

- all (random) sample contrasts have *expectation* zero:

$$E\left(\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}\right) = \sum_{i=1}^{g} c_i \mu_i = 0\,.$$

- Therefore the "ANOVA null hypothesis" can be rephrased as "all population contrasts are zero".

# Maybe not what we want

- Thus in *some* examples, in a particular sense, the "ANOVA null hypothesis" may be "too strong":
- we may only wish to test one (or more) "special" population contrasts are zero.
- Also, the "ANOVA null hypothesis" *may not be rejected for the reason we want*.
- some contrasts may be non-zero, but are they the ones we are interested in?

# $t$-tests for individual contrasts

- Suppose we really only want to test that $H_0$: $\sum_{i=1}^{g} c_i\mu_i = 0$ for some "special contrast" given by $c_1, \dots, c_g$ (with $\sum_{i=1}^{g} c_i = 0$).
- We can of course perform the ANOVA Mean-Square Ratio $F$-test, but can we possibly do better?
- We can perform a more "targeted" $t$-test using the corresponding sample contrast *and* the **residual mean square**.

- The corresponding (random) sample contrast

$$\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} \sim N\left(\sum_{i=1}^{g} c_i \mu_i, \sigma^2 \sum_{i=1}^{g} \frac{c_i^2}{n_i}\right) .$$

- The *standardised version*

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} - \sum_{i=1}^{g} c_i \mu_i}{\sigma \sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}}$$

  thus has a **standard normal distribution**.

- Replacing $\sigma$ in the denominator with $\hat{\sigma} = \sqrt{\mathrm{ResMS}} \sim \sqrt{\chi^2_{N-g}/(N-g)}$ (indep. of the $\bar{Y}_{i\bullet}$'s) gives

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet} - \sum_{i=1}^{g} c_i \mu_i}{\hat{\sigma} \sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}} \sim t_{N-g} .$$

- Thus a $t$-statistic for testing the hypothesis that $\sum_{i=1}^{g} c_i \mu_i = 0$ is

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}}{\hat{\sigma}\sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}}$$

which has a $t_{N-g}$ distribution if $\sum_{i=1}^{g} c_i \mu_i = 0$.

- This *generalises* the two-sample $t$-statistic.

# Yield

Let $\mu_1, \mu_2$ and $\mu_3$ represent the population means of treatment 1, treatment 2 and the control group, respectively.

Let's consider if there is a difference between treatment 1, `trt1` and treatment 2, `trt2`, this corresponds to contrast coefficients $c_1 = 1$, $c_2 = -1$ and $c_3 = 0$.

```
plant_summary = PlantGrowth %>%
  mutate(group = factor(group,
      levels = c("trt1","trt2", "ctrl"))) %>%
  group_by(group) %>%
  summarise(n = n(),
            mean_weight = mean(weight)) %>%
  mutate(contrast_coefficients = c(1,-1,0))
plant_summary
```

```
## # A tibble: 3 × 4
##   group     n mean_weight contrast_coefficients
##   <fct> <int>       <dbl>                 <dbl>
## 1 trt1     10        4.66                     1
## 2 trt2     10        5.53                    -1
## 3 ctrl     10        5.03                     0
```

```
n_i = plant_summary %>% pull(n)
ybar_i = plant_summary %>% pull(mean_weight)
c_i =  plant_summary %>%
  pull(contrast_coefficients)
```

Sample contrast:

```
sum(c_i * ybar_i)
```

```
## [1] -0.865
```

# Residual standard error

Recall the ANOVA analysis from earlier:

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

- The "Residual standard error" is the estimate of $\sigma$, the (population) standard deviation (within each group)

We use the `tidy()` function from the **broom** package to help extract these terms from the anova object.

```
library(broom)
tidy(plant_anova)
```

```
## # A tibble: 2 × 6
##   term         df sumsq meansq statistic p.value
##   <chr>     <dbl> <dbl>  <dbl>     <dbl>   <dbl>
## 1 group         2  3.77  1.88       4.85  0.0159
## 2 Residuals    27 10.5   0.389     NA    NA
```

```
resid_ms = tidy(plant_anova)$meansq[2]
resid_se = sqrt(resid_ms)
c(resid_ms, resid_se)
```

```
## [1] 0.3885959 0.6233746
```

- The $t$-statistic is obtained as follows:

$$\frac{\sum_{i=1}^{g} c_i \bar{Y}_{i\bullet}}{\hat{\sigma}\sqrt{\sum_{i=1}^{g} \frac{c_i^2}{n_i}}}$$

```
se = sqrt(resid_ms * sum((c_i^2) / n_i))
se
```

```
## [1] 0.2787816
```

```
t_stat = sum(c_i * ybar_i)/se
t_stat
```

```
## [1] -3.102787
```

- The test statistic has a $t_{N-g}$ distribution if $\sum_{i=1}^{g} c_i \mu_i = 0$.
- This is the same degrees of freedom as the denominator (residual) degrees of freedom from the ANOVA!
- A (two-sided) p-value is obtained using

```
2*pt(abs(t_stat), df = plant_anova$df.res,
     lower.tail = FALSE)
```

```
## [1] 0.004459236
```

- Why is this better than an ordinary two-sample $t$-test?
    - (Potentially) a smaller standard error! Better estimate of $\sigma$.

# Confidence intervals

# Confidence interval

- A confidence interval for a population contrast can be obtained in the usual way, based on the $t$-statistic.
- Suppose the "multiplier", or critical value, $t^\star$ satisfies

$$P(-t^\star \le t_{N-g} \le t^\star) = 0.95\,.$$

- Then whatever be the "true" values of the $\mu_i$'s, since the quantity

$$\frac{\sum_{i=1}^g c_i \bar{Y}_{i\bullet} - \sum_{i=1}^g c_i \mu_i}{\hat{\sigma}\sqrt{\sum_{i=1}^g \frac{c_i^2}{n_i}}} \sim t_{N-g}$$

  we have, using the usual confidence interval-type manipulations,

$$P\left(\sum_i c_i \bar{Y}_{i\bullet} - t^\star \hat{\sigma}\sqrt{\sum_i \frac{c_i^2}{n_i}} \le \sum_i c_i \mu_i \le \sum_i c_i \bar{Y}_{i\bullet} + t^\star \hat{\sigma}\sqrt{\sum_i \frac{c_i^2}{n_i}}\right) = 0.95\,.$$

# "Observed value" of confidence interval

Therefore for observed sample means $\bar{y}_{1\bullet}, \ldots, \bar{y}_{g\bullet}$, a 95% confidence interval for the "true" population contrast $\sum_i c_i \mu_i$ is given by

$$\underbrace{\sum_i c_i \bar{y}_{i\bullet}}_{\text{estimate}} \pm t^\star \left( \underbrace{\hat{\sigma} \sqrt{\sum_i \frac{c_i^2}{n_i}}}_{\text{st.error}} \right)$$

where, as above, $\hat{\sigma}$ denotes the square root of the **residual mean square**

- A two-sided 95% confidence interval for the "special contrast" considered above is given as follows:
- the "multiplier" $t^\star$ is determined via:

```
t_star = qt(0.975, df = 969)
t_star
```

```
## [1] 1.962415
```

- The interval is then obtained using

```
sum(c_i * ybar_i) + c(-1,1) * t_star * se
```

```
## [1] -1.4120853 -0.3179147
```