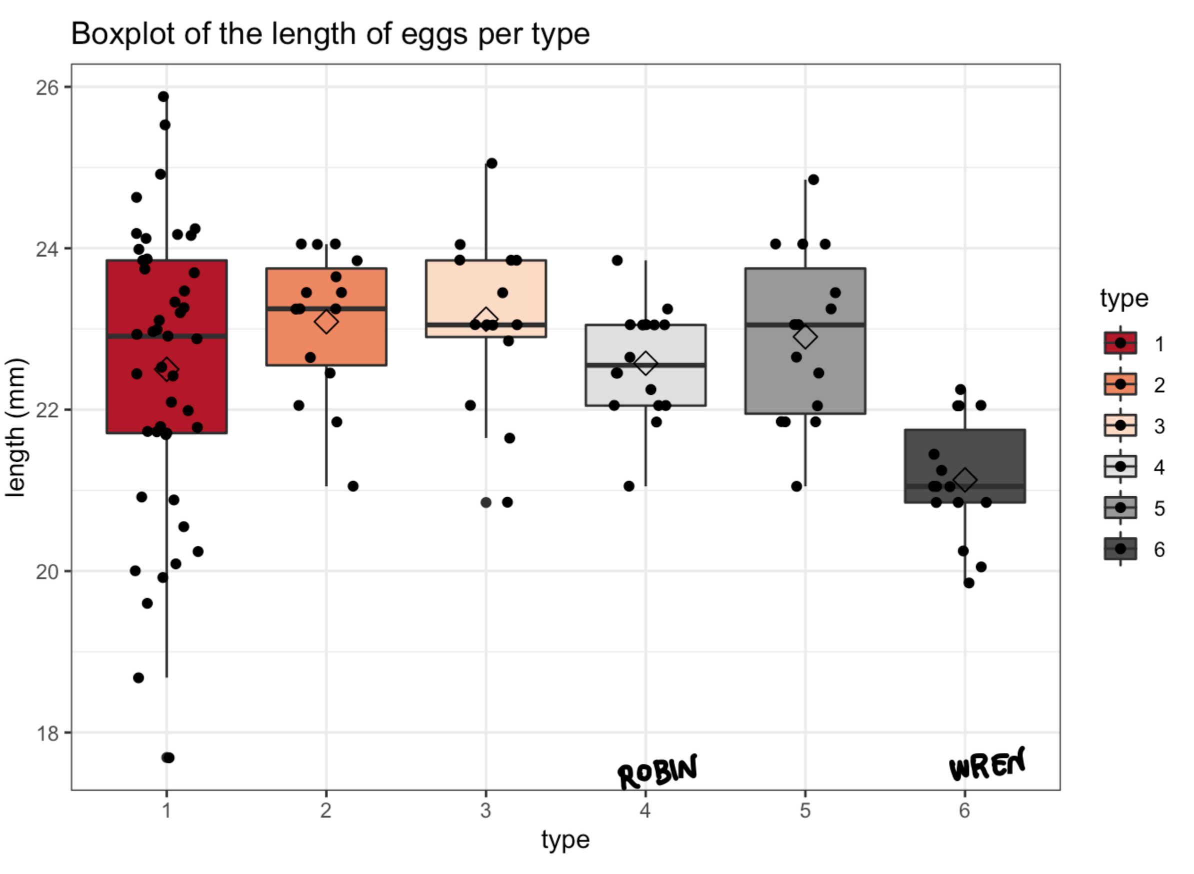
Lecture 6

https://gtpb.github.io/PSLS20/pages/07-ANOVA/ANOVA_cuckoo.html

1 The cuckoo dataset

The common cuckoo does not build its own nest: it prefers to lay its eggs in another birds' nest. It is known, since 1892, that the type of cuckoo bird eggs are different between different locations. In a study from 1940, it was shown that cuckoos return to the same nesting area each year, and that they always pick the same bird species to be a "foster parent" for their eggs.
Over the years, this has lead to the development of geographically determined subspecies of cuckoos. These subspecies have evolved in such a way that their eggs look as similar as possible as those of their foster parents.
The cuckoo dataset contains information on 120 Cuckoo eggs, obtained from randomly selected "foster" nests. For these eggs, researchers have measured the `length` (in mm) and established the `type` (species) of foster parent. The type column is coded as follows:

- `type=1`: Meadow pipit
- `type=2`: Tree pipit
- `type=3`: Dunnock
- `type=4`: European robin
- `type=5`: White wagtail
- `type=6`: Eurasian wren



EXERCISE

we consider the length of the eggs for the robin and the wren
we want to understand if the length of the eggs of the wren is different from the length of the eggs of the robin.

ROBIN: $(y_1^R, \dots, y_n^R)$ n independent observations from $Y^R \sim N(\mu^R, \sigma^2)$

WREN: $(y_1^W, \dots, y_m^W)$ m independent observations from $Y^W \sim N(\mu^W, \sigma^2)$

assuming common variances

we want to test the hypothesis $H_0: \mu^R = \mu^W$
$H_1: \mu^R \neq \mu^W$

Two-sample T-test assuming equal variance $(\mathrm{var}(Y^R) = \mathrm{var}(Y^W) = \sigma^2)$

From the data, we can easily compute

$$\hat\mu^R = \bar y^R = n^{-1}\sum_{i=1}^n y_i^R \qquad s_R^2 = (n-1)^{-1}\sum_{i=1}^n (y_i^R - \bar y^R)^2$$
$$\hat\mu^W = \bar y^W = m^{-1}\sum_{i=1}^m y_i^W \qquad s_W^2 = (m-1)^{-1}\sum_{i=1}^m (y_i^W - \bar y^W)^2$$

since we assume $\sigma_R^2 = \sigma_W^2 = \sigma^2$ we can use as an estimate of the overall variance the quantity $s^2 = \dfrac{(n-1)s_R^2 + (m-1)s_W}{n-1-m-1}$ (weighted average)

$H_0: \mu^R = \mu^W \iff H_0: \mu^R - \mu^W = 0$

$\bar Y^R \sim N(\mu^R, \frac{\sigma^2}{n})$
$\bar Y^W \sim N(\mu^W, \frac{\sigma^2}{m})$ independent $\Rightarrow \bar Y^R - \bar Y^W \sim N(\mu^R - \mu^W, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$

$$\Rightarrow T = \frac{\bar Y^R - \bar Y^W - 0}{\sqrt{S^2\left(\frac{1}{n}+\frac{1}{m}\right)}} = \frac{\bar Y^R - \bar Y^W}{\sqrt{S^2\left(\frac{m+n}{mn}\right)}} \overset{H_0}{\sim} t_{n+m-2}$$

and we reject H_0 at level α if $|t^{obs}| > t_{n+m-2;\,1-\frac{\alpha}{2}}$

Correspondence between t-test for comparing the means of two independent samples with equal variances and test on the regression coefficient of a simple lm.

We can reformulate the test using a simple linear model
Write the full vector of the response as $\underline{y} = (\underline{y}^R, \underline{y}^W) = (\underbrace{y_1, \dots, y_n}_{\text{robin}}, \underbrace{y_{n+1}, \dots, y_{n+m}}_{\text{wren}})$

$Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma^2)$ iid $\qquad i = 1, \dots, n+m$

x_i is a DUMMY variable (indicator variable)

$$x_i = \begin{cases} 0 & \text{if the bird is a robin} \\ 1 & \text{if the bird is a wren} \end{cases} \longrightarrow X = [\underline{1}\ \underline{x}] = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\ \vdots\\1\end{matrix}}\right\} n \\ \left.\vphantom{\begin{matrix}1\\ \vdots\\1\end{matrix}}\right\} m \end{matrix}$$

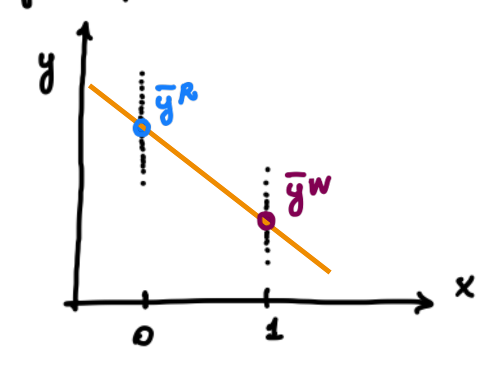
model matrix

Let's see what happens to Y_i depending on the value of x_i

- if $x_i = 0$ $\quad Y_i \sim N(\beta_1, \sigma^2) \Rightarrow \mu_i = \beta_1 = \mu^R$
- if $x_i = 1$ $\quad Y_i \sim N(\beta_1 + \beta_2, \sigma^2) \Rightarrow \mu_i = \beta_1 + \beta_2 = \mu^W$

So if we want to test $H_0: \mu^R = \mu^W \iff H_0: \beta_1 = \beta_1 + \beta_2$
$\iff H_0: \beta_2 = 0$

if we plot this model



moreover $\beta_2 = \mu^W - \mu^R$

To test this hypotesis using the linear model
$H_0: \beta_2 = 0$ we have seen the test on the coefficients $\rightarrow$ test t
in particular

$$T = \frac{\hat\beta_2 - 0}{\sqrt{\frac{S^2}{\Sigma(x_i - \bar x)^2}}} \overset{H_0}{\sim} t_{m+n-2}$$

From the previous lectures we know that

$$\hat\beta_2 = \frac{\sum_{i=1}^{n+m}(x_i - \bar x)(y_i - \bar y)}{\sum_{i=1}^{n+m}(x_i - \bar x)^2} = \frac{\sum_{i=1}^{n+m} x_i y_i - (n+m)\bar x \bar y}{\sum_{i=1}^{n+m}(x_i - \bar x)^2}$$

we need to compute $\bar x, \bar y, \Sigma x_i y_i, \Sigma (x_i - \bar x)^2$

- $\bar x = \frac{1}{n+m}\sum_{i=1}^{n+m} x_i = \frac{m}{n+m}$
- $\bar y = \frac{1}{n+m}\sum_{i=1}^{n+m} y_i = \frac{1}{n+m}\left(\sum_{i=1}^n y_i + \sum_{i=n+1}^{m+n} y_i\right) = \frac{1}{n+m}(n\bar y^R + m\bar y^W)$
- $\sum_{i=1}^{m+n} x_i y_i = \sum_{i=n+1}^{m} y_i = m\bar y^W$
- $\sum_{i=1}^{n+m}(x_i - \bar x)^2 = \sum_{i=1}^{n}(\overset{0}{x_i} - \bar x)^2 + \sum_{i=n+1}^{m}(\overset{1}{x_i} - \bar x)^2 = \sum_{i=1}^{n}(-\bar x)^2 + \sum_{i=n+1}^{m}(1-\bar x)^2 =$
$= n\cdot\left(\frac{m}{n+m}\right)^2 + \sum_{i=n+1}^{m}\left(1 - \frac{m}{n+m}\right)^2 =$
$= \frac{nm^2}{(n+m)^2} + m\cdot\frac{n^2}{(n+m)^2} = \frac{nm\cancel{(n+m)}}{(n+m)^{\cancel{2}}} = \frac{nm}{n+m}$

Hence

$$\hat\beta_2 = \frac{\cancel{m}\bar y^W - \cancel{(n+m)}\cdot\frac{\cancel{m}}{\cancel{n+m}}\cdot\frac{1}{n+m}(n\bar y^R + m\bar y^W)}{\frac{n\cancel{m}}{n+m}}$$
$$= \frac{\bar y^W - \frac{1}{n+m}(n\bar y^R + m\bar y^W)}{\frac{n}{n+m}} =$$
$$= \frac{\cancel{\frac{1}{n+m}}(n\bar y^W + \cancel{m\bar y^W} - n\bar y^R - \cancel{m\bar y^W})}{\frac{n}{\cancel{n+m}}} = \bar y^W - \bar y^R$$

From the simple lm: $\hat\beta_1 = \bar y - \hat\beta_1 \bar x$
in this case:

$$\hat\beta_1 = \frac{1}{n+m}(n\bar y^R + m\bar y^W) - \frac{m}{n+m}(\bar y^W - \bar y^R)$$
$$= \frac{1}{n+m}(n\bar y^R + \cancel{m\bar y^W} - \cancel{m\bar y^W} + m\bar y^R)$$
$$= \frac{n+m}{n+m}\bar y^R = \bar y^R$$

Finally

$$\hat s^2 = \frac{1}{n+m-2}\sum_{i=1}^{n+m}(y_i - \hat y_i)^2 = \frac{1}{n+m-2}\sum_{i=1}^{n+m}(y_i - \hat\beta_1 - \hat\beta_2 x_i)^2 =$$
$$= \frac{1}{n+m-2}\sum_{i=1}^{n+m}(y_i - \bar y^R - (\bar y^W - \bar y^R)x_i)^2 =$$
$$= \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(y_i - \bar y^R)^2 + \sum_{i=n+1}^{n+m}(y_i - \cancel{\bar y^R} - \bar y^W + \cancel{\bar y^R})^2\right] =$$
$$= \frac{1}{n+m-2}\left[\underbrace{\sum_{i=1}^{n}(y_i - \bar y^R)^2}_{(n-1)s_R^2} + \underbrace{\sum_{i=n+1}^{n+m}(y_i - \bar y^W)^2}_{(m-1)s_W^2}\right]$$

Finally, going back to the test,

$$T = \frac{\hat\beta_2}{\sqrt{\frac{S^2}{\Sigma(x_i - \bar x)^2}}} = \frac{\bar Y^W - \bar Y^R}{\sqrt{S^2\left(\frac{n+m}{nm}\right)}} \overset{H_0}{\sim} t_{n+m-2}$$

Notice that if we consider instead a covariate

$$z_i = \begin{cases} 1 & \text{if the bird is a robin} \\ 0 & \text{if the bird is a wren} \end{cases}$$

then $\mu^W = \beta_1$ and $\mu^R = \beta_1 + \beta_2$
is a different model but the result is the same

Until now, we only had 2 categories (bird species) $\rightarrow$ we only need 1 dummy
Let's consider now {robin, wren, pipit}
Now $\underline{y} = (\underline{y}^R, \underline{y}^W, \underline{y}^P)$
I need 2 indicator variables to encode 3 groups

$Y_i = \beta_1 + \beta_2 x_{i1} + \beta_3 x_{i2} + \varepsilon_i \qquad i = 1, \dots, N$

$$x_{i1} = \begin{cases} 0 & \text{if the bird is a robin or a pipit} \\ 1 & \text{if the bird is a wren} \end{cases}$$

$$x_{i2} = \begin{cases} 0 & \text{if the bird is a robin or a wren} \\ 1 & \text{if the bird is a pipit} \end{cases}$$

$$X = [\underline{1}\ \underline{x}_1\ \underline{x}_2] = \begin{bmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\ \vdots\\1\end{matrix}}\right\} \#\text{robins} \\ \left.\vphantom{\begin{matrix}1\\ \vdots\\1\end{matrix}}\right\} \#\text{wrens} \\ \left.\vphantom{\begin{matrix}1\\ \vdots\\1\end{matrix}}\right\} \#\text{pipits} \end{matrix}$$

ROBIN: $\mu^R = \beta_1$
WREN: $\mu^W = \beta_1 + \beta_2$
PIPIT: $\mu^P = \beta_1 + \beta_3$

multiple linear model
We can generalize the comparison of the means of 2 groups to $G \geq 2$ groups.
We do not need ad-hoc tests but only the general theory of the multiple lm.