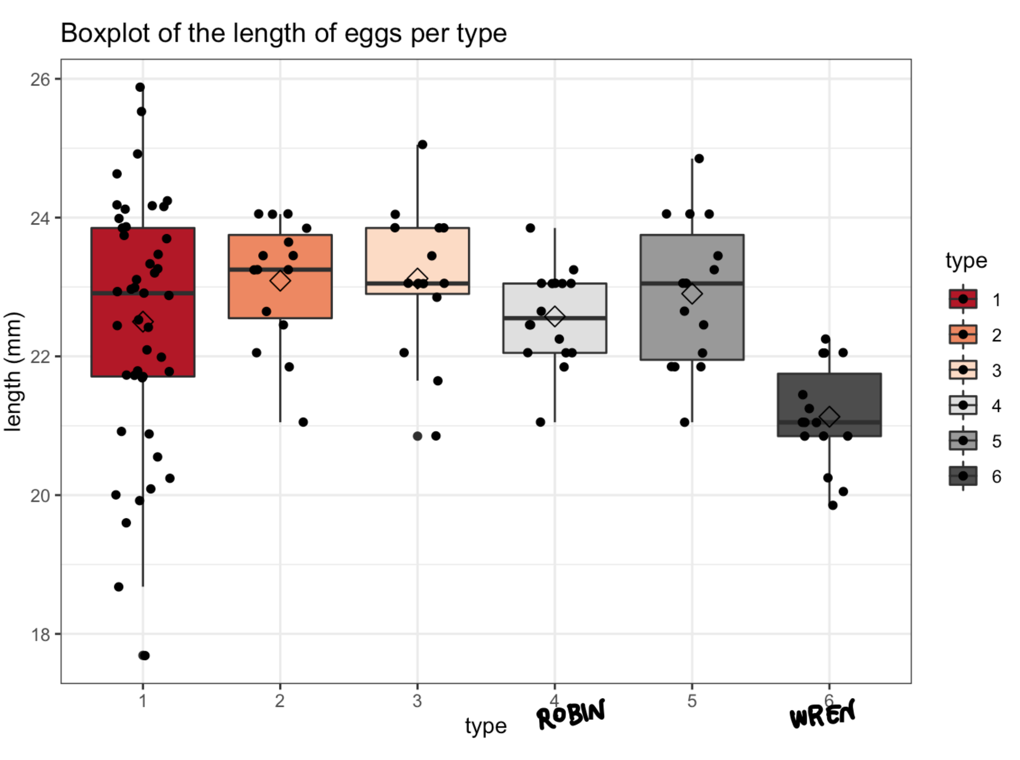
07 Nov - LEC 8

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.

Boxplot of the length of eggs per type



EXERCISE

We are interested in understanding if the LENGTH OF THE EGGS IN THE ROBIN'S NEST IS DIFFERENT FROM THE LENGTH OF THE EGGS IN THE WREN'S NEST

Data: two independent samples of the eggs' length

ROBIN: $(y_1^R, \dots, y_n^R)$ n independent observations of lengths from robins' nests

WREN: $(y_1^W, \dots, y_m^W)$ m independent observations of lengths from wrens' nests

Distributive assumptions
$$Y_i^R \sim N(\mu^R, \sigma^2) \quad iid. \quad i = 1, \dots, n$$
$$Y_i^W \sim N(\mu^W, \sigma^2) \quad iid. \quad i = 1, \dots, m$$

assuming common variances $var(Y_i^R) = var(Y_i^W) = \sigma^2$

We have two normal samples with equal variance (and different means)
The ML estimates of the group-specific means in this case are simply
$$\hat{\mu}^R = \bar{y}^R = \frac{1}{n}\sum_{i=1}^n y_i^R$$
$$\hat{\mu}^W = \bar{y}^W = \frac{1}{m}\sum_{i=1}^m y_i^W$$

Since we assume common variances, the ML estimate of σ^2 is
$$\hat{\sigma}^2 = \frac{1}{m+n}\left(\sum_{i=1}^n (y_i^R - \bar{y}^R)^2 + \sum_{i=1}^m (y_i^W - \bar{y}^W)^2\right)$$

and the unbiased estimate is
$$S^2 = \frac{1}{m+n-2}\left(\sum_{i=1}^n (y_i^R - \bar{y}^R)^2 + \sum_{i=1}^m (y_i^W - \bar{y}^W)^2\right)$$
$$= \frac{(n-1)S_R^2 + (m-1)S_W^2}{n+m-2}$$

weighted average of the group-specific estimates

where
$$S_R^2 = \frac{1}{(n-1)}\sum_{i=1}^n (y_i^R - \bar{y}^R)^2$$
$$S_W^2 = \frac{1}{(m-1)}\sum_{i=1}^m (y_i^W - \bar{y}^W)^2$$

The estimators of the means are
$$\bar{Y}^R = \frac{1}{n}\sum_{i=1}^n Y_i^R \quad \text{and} \quad \bar{Y}^W = \frac{1}{m}\sum_{i=1}^m Y_i^W$$
with
$$\bar{Y}^R \sim N\left(\mu^R, \frac{\sigma^2}{n}\right) \quad \text{and} \quad \bar{Y}^W \sim N\left(\mu^W, \frac{\sigma^2}{m}\right) \quad \text{independent}$$

We want to test the hypothesis $\begin{cases} H_0: \mu^R = \mu^W \\ H_1: \mu^R \neq \mu^W \end{cases}$

The procedure to perform this test is a two-sample T-test assuming equal variances

Notice that $H_0: \mu^R = \mu^W \Rightarrow H_0: \mu^W - \mu^R = 0$

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

Under H_0, $\mu^W - \mu^R = 0$. Hence,
$$\bar{Y}^W - \bar{Y}^R \overset{H_0}{\sim} N\left(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$$
$$\Rightarrow \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{\sigma^2\left(\frac{1}{m} + \frac{1}{n}\right)}} \overset{H_0}{\sim} N(0,1)$$

but σ^2 is unknown. We substitute it with S^2:
$$\Rightarrow T = \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{S^2\left(\frac{1}{m} + \frac{1}{n}\right)}} = \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{S^2\left(\frac{m+n}{mn}\right)}} = \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{\frac{(n-1)S_R^2 + (m-1)S_W^2}{n+m-2} \cdot \frac{m+n}{mn}}} \overset{H_0}{\sim} t_{n+m-2}$$

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

We will now see how to perform this type of analysis using the Gaussian linear model

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

We can reformulate the test using a simple linear model
Write the full vector of the response as
$$\underline{y} = \begin{bmatrix} \underline{y}^R \\ \underline{y}^W \end{bmatrix} = (\underbrace{y_1, \dots, y_n}_{Robin}, \underbrace{y_{n+1}, \dots, y_{n+m}}_{Wren})^T \qquad (n+m)\text{-dimensional vector}$$

MODEL FORMULATION

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

x_i is a DUMMY variable (indicator variable)
$$x_i = \begin{cases} 0 & \text{if the } i\text{-th egg is in a ROBIN's nest} \\ 1 & \text{if the } i\text{-th egg is in a WREN's nest} \end{cases} \longrightarrow X = [\underline{1}\ \underline{x}] = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\\ \vdots\\1\end{matrix}}\right\} n \\ \left.\vphantom{\begin{matrix}1\\ \vdots\\1\end{matrix}}\right\} m \end{matrix}$$

Let's see what happens to Y_i depending on the bird species:

- if egg_i is in a ROBIN's nest

  $x_i = 0 \Rightarrow \mu_i = \beta_1 + \beta_2 \cdot 0 = \beta_1 \Rightarrow Y_i \sim N(\beta_1, \sigma^2)$ for $i = 1, \dots, n$

  This is the group of eggs from robins $\Rightarrow Y_i \sim N(\mu^R, \sigma^2) \Rightarrow \beta_1 = \mu^R$

- if egg_i is in a WREN's nest

  $x_i = 1 \Rightarrow \mu_i = \beta_1 + \beta_2 \cdot 1 = \beta_1 + \beta_2 \Rightarrow Y_i \sim N(\beta_1 + \beta_2, \sigma^2)$ for $i = n+1, \dots, n+m$

  This is the group of eggs from wrens $\Rightarrow Y_i \sim N(\mu^W, \sigma^2) \Rightarrow \beta_1 + \beta_2 = \mu^W$

Remark: this is a reparameterization:
a one-to-one correspondence between (μ^R, μ^W) and (β_1, β_2)
$$\begin{cases} \mu^R = \beta_1 \\ \mu^W = \beta_1 + \beta_2 \end{cases} \Longleftrightarrow \begin{cases} \beta_1 = \mu^R \\ \beta_2 = \mu^W - \mu^R \end{cases}$$

The correspondence also holds for the ML estimates: $\begin{cases} \hat{\beta}_1 = \hat{\mu}^R \\ \hat{\beta}_2 = \hat{\mu}^W - \hat{\mu}^R \end{cases}$

So if we want to test $H_0: \mu^R = \mu^W \Leftrightarrow H_0: \mu^W - \mu^R = 0$
$\Leftrightarrow H_0: \beta_2 = 0$

To test this hypothesis using the linear model
$H_0: \beta_2 = 0$ we have seen the test on individual coefficients $\rightarrow$ t-test
in particular
$$T = \frac{\hat{\beta}_2 - 0}{\sqrt{\frac{S^2}{\sum_{i=1}^{n+m}(x_i - \bar{x})^2}}} \overset{H_0}{\sim} t_{m+n-2}$$

expression of $\hat{var}(\hat{\beta}_2)$ in the simple LM.

We now compute the estimated regression model and show the equivalence with the previous procedure.

We have a SIMPLE LINEAR MODEL $Y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$

From the previous lectures we know that the estimate of β_2 in the simple lm is:
$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n+m}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n+m}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n+m} x_i y_i - (n+m)\bar{x}\bar{y}}{\sum_{i=1}^{n+m}(x_i - \bar{x})^2}$$

we need to compute $\bar{x}, \bar{y}, \sum_{i=1}^{n+m} x_i y_i, \sum_{i=1}^{n+m}(x_i - \bar{x})^2$

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

  $= n \cdot \left(\frac{m}{n+m}\right)^2 + \sum_{i=n+1}^m \left(1 - \frac{m}{n+m}\right)^2 =$

  $= \frac{nm^2}{(n+m)^2} + m \cdot \frac{n^2}{(n+m)^2} = \frac{nm\cancel{(n+m)}}{(n+m)^{\cancel{2}}} = \frac{nm}{n+m}$

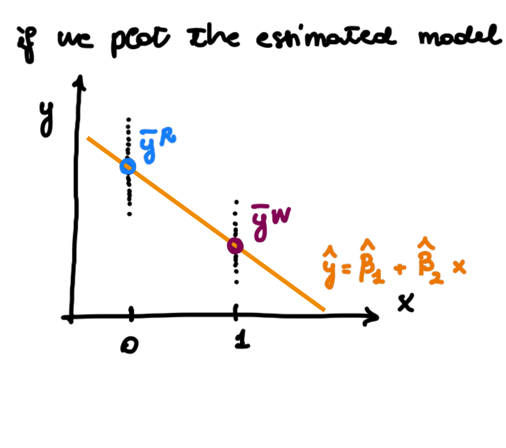
Hence
$$\hat{\beta}_2 = \frac{\cancel{m}\bar{y}^W - \cancel{(n+m)} \cdot \frac{\cancel{m}}{\cancel{n+m}} \cdot \frac{1}{n+m}(n\bar{y}^R + m\bar{y}^W)}{\frac{n\cancel{m}}{n+m}}$$
$$= \frac{\bar{y}^W - \frac{1}{n+m}(n\bar{y}^R + m\bar{y}^W)}{\frac{n}{n+m}} =$$
$$= \frac{\cancel{\frac{1}{n+m}}(n\bar{y}^W + \cancel{m\bar{y}^W} - n\bar{y}^R - \cancel{m\bar{y}^W})}{\frac{n}{\cancel{n+m}}} = \bar{y}^W - \bar{y}^R$$

The estimate of β_1 instead is $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}$
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
$$s^2 = \frac{1}{n+m-2}\sum_{i=1}^{n+m}(y_i - \hat{y}_i)^2 = \frac{1}{n+m-2}\sum_{i=1}^{n+m}(y_i - \hat{\beta}_1 - \hat{\beta}_2 x_i)^2 =$$
$$= \frac{1}{n+m-2}\sum_{i=1}^{n+m}(y_i - \bar{y}^R - (\bar{y}^W - \bar{y}^R)x_i)^2 =$$
$$= \frac{1}{n+m-2}\left[\sum_{i=1}^n (y_i - \bar{y}^R)^2 + \sum_{i=n+1}^{n+m}(y_i - \cancel{\bar{y}^R} - \bar{y}^W + \cancel{\bar{y}^R})^2\right] =$$
$$= \frac{1}{n+m-2}\left[\underbrace{\sum_{i=1}^n (y_i - \bar{y}^R)^2}_{(n-1)s_R^2} + \underbrace{\sum_{i=n+1}^{n+m}(y_i - \bar{y}^W)^2}_{(m-1)s_W^2}\right] = \frac{1}{n+m-2}\left[(n-1)s_R^2 + (m-1)s_W^2\right]$$

Hence we obtain
$$\hat{\beta}_2 = \bar{y}^W - \bar{y}^R$$
$$\hat{\beta}_1 = \bar{y}^R$$
$$s^2 = \frac{(n-1)s_R^2 + (m-1)s_W^2}{n+m-2}$$
$$\sum_{i=1}^{n+m}(x_i - \bar{x})^2 = \frac{nm}{n+m}$$

if we plot the estimated model



$$\hat{y}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$$
$$= \bar{y}^R + (\bar{y}^W - \bar{y}^R)x_i$$

Going back to the test,
$$T = \frac{\hat{\beta}_2}{\sqrt{\frac{S^2}{\sum_{i=1}^{n+m}(x_i - \bar{x})^2}}} = \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{\frac{(n-1)S_R^2 + (m-1)S_W^2}{n+m-2} \cdot \left(\frac{nm}{n+m}\right)^{-1}}}$$
$$\Rightarrow T = \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{\frac{(n-1)S_R^2 + (m-1)S_W^2}{n+m-2} \cdot \frac{n+m}{n \cdot m}}} \overset{H_0}{\sim} t_{n+m-2}$$

it's the same expression as the two-sample t-test

Hence we have proven the correspondence of the two procedures.

Remark:
Notice that if we consider instead a covariate
$$x_i = \begin{cases} 1 & \text{if the bird is a robin} \\ 0 & \text{if the bird is a wren} \end{cases}$$
then $\mu^W = \beta_1$ and $\mu^R = \beta_1 + \beta_2$
is a different model but the result of inference is the same