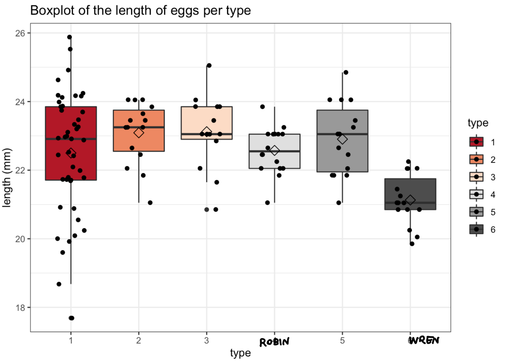
1 The cuckoo dataset

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.



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

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

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

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

We want to study whether the two groups are equal or different. We are assuming that observations in each group are normal with group-specific means and common variances $\Rightarrow$ the two distributions are equal iff their means are equal $\Leftrightarrow$ $\mu^W = \mu^R$.

Thus, we want to perform a test of equality of two means of two normal samples.

We have two normal samples with equal variance (and different means)

DATA DENSITY

$$f(\underline{y}^R, \underline{y}^W) = \prod_{i=1}^{n} \phi(y_i^R; \mu^R, \sigma^2) \cdot \prod_{j=1}^{m} \phi(y_j^W; \mu^W, \sigma^2)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i^R - \mu^R)^2\right\} \cdot \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_j^W - \mu^W)^2\right\}$$

$$= (2\pi)^{-\frac{n+m}{2}} (\sigma^2)^{-\frac{n+m}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i^R - \mu^R)^2 + \sum_{j=1}^{m}(y_j^W - \mu^W)^2\right]\right\} \propto L(\mu^R, \mu^W, \sigma^2)$$

Log-likelihood

$$\ell(\mu^W, \mu^R, \sigma^2) = -\frac{n+m}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_i^R - \mu^R)^2 + \sum_{j=1}^{m}(y_j^W - \mu^W)^2\right]$$

MAXIMUM LIKELIHOOD ESTIMATES

- The ML estimates of the group-specific means in this case are simply

$$\hat{\mu}^R = \bar{y}^R = \frac{1}{n}\sum_{i=1}^{n} y_i^R$$

$$\hat{\mu}^W = \bar{y}^W = \frac{1}{m}\sum_{i=1}^{m} y_i^W$$

- ML estimate of the (common) variance σ^2

we consider the log-likelihood $\ell(\sigma^2, \hat{\mu}^W, \hat{\mu}^R)$, which is a function only of σ^2.

we compute the score function and solve the likelihood equation

$$s_2(\sigma^2) = \frac{\partial}{\partial\sigma^2}\ell(\hat{\mu}^R, \hat{\mu}^W, \sigma^2)$$

$$= -\frac{n+m}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2(\sigma^2)^2}\left[\sum_{i=1}^{n}(y_i^R - \hat{\mu}^R)^2 + \sum_{j=1}^{m}(y_j^W - \hat{\mu}^W)^2\right]$$

$$s_2(\sigma^2) = 0 \Rightarrow \frac{-(n+m)\sigma^2}{2(\sigma^2)^2} + \frac{1}{2(\sigma^2)^2}\left[\sum_{i=1}^{n}(y_i^R - \hat{\mu}^R)^2 + \sum_{j=1}^{m}(y_j^W - \hat{\mu}^W)^2\right] = 0$$

$$\hat{\sigma}^2 = \frac{1}{n+m}\left[\sum_{i=1}^{n}(y_i^R - \hat{\mu}^R)^2 + \sum_{j=1}^{m}(y_j^W - \hat{\mu}^W)^2\right]$$

Notice that $\hat{\sigma}^2 = \dfrac{n\hat{\sigma}_R^2 + m\hat{\sigma}_W^2}{n+m}$ where $\hat{\sigma}_R^2$ and $\hat{\sigma}_W^2$ are the group-specific MLEs.

Moreover, the unbiased estimate is

$$S^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(y_i^R - \bar{y}^R)^2 + \sum_{j=1}^{m}(y_j^W - \bar{y}^W)^2\right)$$

$$= \frac{(n-1)S_R^2 + (m-1)S_W^2}{n+m-2}$$

weighted averages of the group-specific estimates

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

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

ESTIMATORS

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

$$\bar{Y}^R \sim N(\mu^R, \frac{\sigma^2}{n}) \quad \text{and} \quad \bar{Y}^W \sim N(\mu^W, \frac{\sigma^2}{m}) \quad \text{independent}$$

And the estimator of the variance is

$$S^2 = \frac{(n-1)S_R^2 + (m-1)S_W^2}{n+m-2}$$ where $S_R^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(Y_i^R - \bar{Y}^R)^2$, $S_W^2 = \frac{1}{(m-1)}\sum_{i=1}^{m}(Y_i^W - \bar{Y}^W)^2$

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

Under H_0, $\mu^W - \mu^R = 0$. Hence,

$$\bar{Y}^W - \bar{Y}^R \overset{H_0}{\sim} N(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$$

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

however, σ^2 is unknown. Thus we substitute it with an estimator.
This affects the distribution of the test statistic (similar to the transformation we have seen in the test about β_j in the LM).

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

$$= \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{S^2\left(\frac{n+m}{nm}\right)}} = \frac{\bar{Y}^W - \bar{Y}^R}{\sqrt{\frac{(n-1)S_R^2 + (m-1)S_W^2}{n+m-2}\cdot\frac{n+m}{nm}}} \overset{H_0}{\sim} t_{n+m-2}$$

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