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 Gaussian LM.

We can reformulate the test using a Gaussian linear model.
Write the full vector of the response as:

$$\underline{y} = \begin{bmatrix} \underline{y}^R \\ \underline{y}^W \end{bmatrix} = (\underbrace{y_1, \dots, y_n}_{\text{Robin}}, \underbrace{y_{n+1}, \dots, y_{n+m}}_{\text{Wren}})^T \qquad (n+m)\text{- dimensional vector}$$

MODEL FORMULATION

$Y_i = \mu_i + \varepsilon_i$

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

$x_i$ is a DUMMY variable (indicator variable)

$x_i = \begin{cases} 0 & \text{if the } i\text{-th egg is in a ROBIN's nest} \\ 1 & \text{if the } i\text{-th egg is in a WREN's nest} \end{cases} \longrightarrow X = [\underline{1} \; \underline{x}] = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix} \begin{matrix} \} n \\ \\ \\ \\ \} m \\ \\ \end{matrix}$

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 $(\mu^R, \mu^W)$ and $(\beta_1, \beta_2)$
>
> $$\begin{cases} \mu^R = \beta_1 \\ \mu^W = \beta_1 + \beta_2 \end{cases} \iff \begin{cases} \beta_1 = \mu^R \\ \beta_2 = \mu^W - \mu^R \end{cases}$$
>
> 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 $H_0: \mu^R = \mu^W \iff H_0: \mu^W - \mu^R = 0$
$\iff H_0: \beta_2 = 0$

To test this hypothesis using the linear model
$H_0: \beta_2 = 0$ we have seen the test on individual coefficients → t-test
in particular

$$T = \frac{\hat\beta_2 - 0}{\sqrt{\frac{S^2}{\sum_{i=1}^{n+m}(x_i - \bar x)^2}}} \overset{H_0}{\sim} t_{n+m-2}$$

↳ expression of $\widehat{Var}(\hat\beta_2)$ 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 $\beta_2$ 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 $\bar x, \bar y, \sum_{i=1}^{n+m} x_i y_i, \sum_{i=1}^{n+m}(x_i - \bar x)^2$

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

Hence

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

The estimate of $\beta_1$ 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

$$\hat\beta_1 = \bar y^W - \bar y^R$$
$$\hat\beta_2 = \bar y^R$$
$$s^2 = \frac{(n-1)s_R^2 + (m-1)s_W^2}{n+m-2}$$
$$\sum_{i=1}^{n+m}(x_i - \bar x)^2 = \frac{nm}{n+m}$$

Going back to the test,

$$T = \frac{\hat\beta_2}{\sqrt{\frac{S^2}{\sum_{i=1}^{n+m}(x_i - \bar x)^2}}} = \frac{\bar y^W - \bar y^R}{\sqrt{\frac{(n-1)s_R^2 + (m-1)s_W^2}{n+m-2}\cdot\left(\frac{nm}{n+m}\right)^{-1}}}$$

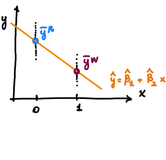
$$\Rightarrow T = \frac{\bar y^W - \bar y^R}{\sqrt{\frac{(n-1)s_R^2 + (m-1)s_W^2}{n+m-2}\cdot\frac{n+m}{nm}}} \overset{H_0}{\sim} t_{n+m-2}$$

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

Hence we have proven the correspondence of the two procedures.

If we plot the estimated model

$$\hat y_i = \hat\beta_1 + \hat\beta_2 x_i$$
$$= \bar y^R + (\bar y^W - \bar y^R)x_i$$




Remark:
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 results of inference is the same