Second assessment

Víctor Ballester Ribó
NIU: 1570866

Statistics
Degree in Mathematics
Universitat Autònoma de Barcelona
May 2022

Exercise 10. *Let* $X_1, \dots X_n \overset{i.i.d.}{\sim} N(\mu_X, {\sigma_X}^2)$*, and* $Y_1, \dots Y_m \overset{i.i.d.}{\sim} N(\mu_Y, {\sigma_Y}^2)$*. We are interested in testing*

$$H_0 : \mu_X = \mu_Y \quad \textit{versus} \quad H_0 : \mu_X \neq \mu_Y$$

with the assumption that ${\sigma_X}^2 = {\sigma_Y}^2 =: \sigma^2$*.*

1. *Derive the LRT for these hypotheses. Show that the LRT can be based on the statistic:*

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{{S_p}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} \tag{1}$$

where

$${S_p}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}(X_i - \overline{X})^2 + \sum_{j=1}^{m}(Y_j - \overline{Y})^2\right)$$

is the so-called pooled variance estimate.

Resolution. First of all, to simplify the notation, let's denote $\mathbf{X} := (X_1, \dots, X_n, Y_1, \dots, Y_m)$ and $\boldsymbol{\theta} := (\mu_X, \mu_Y, \sigma^2)$. Let's write first the likelihood $L(\boldsymbol{\theta}; \mathbf{X})$ of the experiment. We have that:

$$L(\boldsymbol{\theta}; \mathbf{X}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(X_i - \mu_X)^2}{2\sigma^2}} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(Y_j - \mu_Y)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n+m}{2}} \mathrm{e}^{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i - \mu_X)^2 + \sum_{j=1}^{m}(Y_j - \mu_Y)^2\right]}$$

since all the random variables $X_1, \dots, X_n, Y_1, \dots, Y_m$ are pairwise independent. Under H_0 we have a sample $X_1, \dots, X_n, Y_1, \dots, Y_m$ of size $n+m$ from a normal distribution $N(\mu_X, \sigma^2)$. But we already now that the MLE $\hat{\mu}_0$ for μ_X in this case is:

$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} X_i + \sum_{j=1}^{m} Y_j}{n+m} = \frac{n\overline{X} + m\overline{Y}}{n+m}$$

Under H_1, notice that due to the independence of the samples the two MLEs $\hat{\mu}_X$ and $\hat{\mu}_Y$ of μ_X and μ_Y, respectively, must be the usuals ones:

$$\hat{\mu}_X = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \quad \text{and} \quad \hat{\mu}_Y = \frac{\sum_{j=1}^{m} Y_i}{m} = \overline{Y}$$

Finally, since σ^2 is unknown, we must find the MLE $\hat{\sigma}^2$ of σ^2. Let's calculate the log-likelihood $\ell(\boldsymbol{\theta}; \mathbf{X})$ and $\frac{\partial \ell}{\partial \sigma^2}(\boldsymbol{\theta}; \mathbf{X})$. We have that:

$$\ell(\boldsymbol{\theta}; \mathbf{X}) = \log L(\boldsymbol{\theta}; \mathbf{X}) = -\frac{n+m}{2}\log(2\pi) - \frac{n+m}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(X_i - \mu_X)^2 + \sum_{j=1}^{m}(Y_j - \mu_Y)^2}{2\sigma^2}$$
$$\frac{\partial \ell}{\partial \sigma^2}(\boldsymbol{\theta}; \mathbf{X}) = -\frac{n+m}{2\sigma^2} + \frac{\sum_{i=1}^{n}(X_i - \mu_X)^2 + \sum_{j=1}^{m}(Y_j - \mu_Y)^2}{2\sigma^4}$$

Equating the last equation to zero, we get:

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma^2}(\boldsymbol{\theta};\mathbf{X}) = 0 &\iff -\frac{n+m}{2\sigma^2} + \frac{\sum_{i=1}^n (X_i-\mu_X)^2 + \sum_{j=1}^m (Y_j-\mu_Y)^2}{2\sigma^4} = 0\\ &\iff -(n+m) + \frac{\sum_{i=1}^n (X_i-\mu_X)^2 + \sum_{j=1}^m (Y_j-\mu_Y)^2}{\sigma^2} = 0\\ &\iff \sigma^2 = \frac{\sum_{i=1}^n (X_i-\mu_X)^2 + \sum_{j=1}^m (Y_j-\mu_Y)^2}{n+m} =: \hat{\sigma}^2\end{aligned}$$

Let's verify now that ℓ has a maximum attained at $\sigma^2 = \hat{\sigma}^2$:

$$\begin{aligned}\left.\frac{\partial^2 \ell}{\partial (\sigma^2)^2}(\boldsymbol{\theta};\mathbf{X})\right|_{\sigma^2=\hat{\sigma}^2} &= \left.\frac{n+m}{2\sigma^4} - \frac{\sum_{i=1}^n (X_i-\mu_X)^2 + \sum_{j=1}^m (Y_j-\mu_Y)^2}{\sigma^6}\right|_{\sigma^2=\hat{\sigma}^2}\\ &= \frac{n+m}{2\hat{\sigma}^4} - \frac{n+m}{\hat{\sigma}^4}\\ &= -\frac{n+m}{2\hat{\sigma}^4}\\ &< 0\end{aligned}$$

because $\frac{n+m}{2\hat{\sigma}^4}$ is always positive. Hence $\hat{\sigma}^2$ is definitely the MLE of σ^2. But we have to substitute the MLEs of μ_X and μ_Y, so under H_0, $\hat{\sigma}_0^2 := \hat{\sigma}^2$ is:

$$\hat{\sigma}_0^2 = \frac{\sum_{i=1}^n (X_i-\hat{\mu}_0)^2 + \sum_{j=1}^m (Y_j-\hat{\mu}_0)^2}{n+m}$$

And in general (among all the parametric space of μ_X and μ_Y) we have:

$$\hat{\sigma}_1^2 = \frac{\sum_{i=1}^n (X_i-\hat{\mu}_X)^2 + \sum_{j=1}^m (Y_j-\hat{\mu}_Y)^2}{n+m}$$

Putting all of this together on the LRT statistic $\lambda := \dfrac{\sup\{L(\theta,\mathbf{X}) : \mu_X = \mu_Y \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{\geq 0}\}}{\sup\{L(\theta,\mathbf{X}) : \mu_X, \mu_Y \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{\geq 0}\}}$, we have:

$$\begin{aligned}\lambda &= \frac{\sup\{L(\boldsymbol{\theta};\mathbf{X}) : \mu_X = \mu_Y \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{\geq 0}\}}{\sup\{L(\boldsymbol{\theta};\mathbf{X}) : \mu_X, \mu_Y \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{\geq 0}\}}\\ &= \frac{(2\pi\hat{\sigma}_0^2)^{-\frac{n+m}{2}} \mathrm{e}^{-\frac{1}{2\hat{\sigma}_0^2}\left[\sum_{i=1}^n (X_i-\hat{\mu}_0)^2 + \sum_{j=1}^m (Y_j-\hat{\mu}_0)^2\right]}}{(2\pi\hat{\sigma}_1^2)^{-\frac{n+m}{2}} \mathrm{e}^{-\frac{1}{2\hat{\sigma}_1^2}\left[\sum_{i=1}^n (X_i-\hat{\mu}_X)^2 + \sum_{j=1}^m (Y_j-\hat{\mu}_Y)^2\right]}}\\ &= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}_1^2}\right)^{-\frac{n+m}{2}} \cdot \frac{\mathrm{e}^{-\frac{n+m}{2}}}{\mathrm{e}^{-\frac{n+m}{2}}}\\ &= \left(\frac{\sum_{i=1}^n (X_i-\hat{\mu}_0)^2 + \sum_{j=1}^m (Y_j-\hat{\mu}_0)^2}{\sum_{i=1}^n (X_i-\hat{\mu}_X)^2 + \sum_{j=1}^m (Y_j-\hat{\mu}_Y)^2}\right)^{-\frac{n+m}{2}} \qquad (2)\end{aligned}$$

Now, note that:

$$\begin{aligned}\sum_{i=1}^n (X_i-\hat{\mu}_0)^2 &= \sum_{i=1}^n (X_i-\overline{X}+\overline{X}-\hat{\mu}_0)^2\\ &= \sum_{i=1}^n (X_i-\overline{X})^2 + 2(\overline{X}-\hat{\mu}_0)\sum_{i=1}^n (X_i-\overline{X}) + \sum_{i=1}^n (\overline{X}-\hat{\mu}_0)^2\\ &= \sum_{i=1}^n (X_i-\overline{X})^2 + n(\overline{X}-\hat{\mu}_0)^2\end{aligned}$$

because $\sum_{i=1}^n (X_i-\overline{X}) = \sum_{i=1}^n X_i - n\overline{X} = 0$. Similarly, we have that

$$\sum_{j=1}^m (Y_i-\hat{\mu}_0)^2 = \sum_{j=1}^m (Y_i-\overline{Y})^2 + m(\overline{Y}-\hat{\mu}_0)^2$$

But $\hat{\mu}_0 = \frac{n\overline{X}+m\overline{Y}}{n+m}$, so

$$\overline{X} - \hat{\mu}_0 = \overline{X} - \frac{n\overline{X}+m\overline{Y}}{n+m} = \frac{n\overline{X}+m\overline{X}-n\overline{X}-m\overline{Y}}{n+m} = m\frac{\overline{X}-\overline{Y}}{n+m}$$
$$\overline{Y} - \hat{\mu}_0 = \overline{Y} - \frac{n\overline{X}+m\overline{Y}}{n+m} = \frac{n\overline{Y}+m\overline{Y}-n\overline{X}-m\overline{Y}}{n+m} = n\frac{\overline{Y}-\overline{X}}{n+m}$$

and the numerator of the fraction in Eq. (2) becomes:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\hat{\mu}_0)^2+\sum_{j=1}^{m}(Y_j-\hat{\mu}_0)^2 &= \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{m}\left(Y_i-\overline{Y}\right)^2+n\cdot m^2\frac{\left(\overline{X}-\overline{Y}\right)^2}{\left(n+m\right)^2}+m\cdot n^2\frac{\left(\overline{Y}-\overline{X}\right)^2}{\left(n+m\right)^2}\\ &= \sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{m}\left(Y_i-\overline{Y}\right)^2+\frac{nm}{n+m}\left(\overline{X}-\overline{Y}\right)^2\end{aligned}$$

So, recalling that $\hat{\mu}_X = \overline{X}$ and $\hat{\mu}_Y = \overline{Y}$, λ becomes:

$$\begin{aligned}\lambda &= \left(\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{m}\left(Y_i-\overline{Y}\right)^2+\frac{nm}{n+m}\left(\overline{X}-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{m}\left(Y_j-\overline{Y}\right)^2}\right)^{-\frac{n+m}{2}}\\ &= \left(1+\frac{\frac{nm}{n+m}\left(\overline{X}-\overline{Y}\right)^2}{\frac{n+m-2}{n+m-2}\left[\sum_{i=1}^{n}\left(X_i-\hat{\mu}_X\right)^2+\sum_{j=1}^{m}\left(Y_j-\hat{\mu}_Y\right)^2\right]}\right)^{-\frac{n+m}{2}}\\ &= \left(1+\frac{\frac{nm}{n+m}\left(\overline{X}-\overline{Y}\right)^2}{(n+m-2){S_p}^2}\right)^{-\frac{n+m}{2}}\\ &= \left(1+\frac{\left(\overline{X}-\overline{Y}\right)^2}{(\frac{1}{n}+\frac{1}{m})(n+m-2){S_p}^2}\right)^{-\frac{n+m}{2}}\end{aligned}$$

We will reject H_0 when $\lambda <$ const. So:

$$\begin{aligned}\lambda < \text{const.} &\iff \left(1+\frac{\left(\overline{X}-\overline{Y}\right)^2}{(\frac{1}{n}+\frac{1}{m})(n+m-2){S_p}^2}\right)^{-\frac{n+m}{2}} < \text{const.}\\ &\iff 1+\frac{\left(\overline{X}-\overline{Y}\right)^2}{(\frac{1}{n}+\frac{1}{m})(n+m-2){S_p}^2} > \text{const.}\\ &\iff \frac{\left(\overline{X}-\overline{Y}\right)^2}{(\frac{1}{n}+\frac{1}{m}){S_p}^2} > \text{const.}\\ &\iff T^2 > \text{const.}\\ &\iff |T| > \text{const.}\end{aligned}$$

because it goes without saying that $n+m-2>0$.

2. *Show that, under H_0, $T \sim t_{n+m-2}$ (this is know as the two-sample t-test).*
Resolution. Under H_0 we know that $\overline{X} \sim N(\mu_X, \sigma^2/n)$ and $\overline{Y} \sim N(\mu_X, \sigma^2/m)$. But since we know that $-\overline{Y} \sim N(-\mu_X, \sigma^2/m)$, we have that $\overline{X}-\overline{Y} \sim N(0, \sigma^2/n+\sigma^2/m) = N(0, \sigma^2(\frac{1}{n}+\frac{1}{m}))$ and so:

$$\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}} \sim N(0,1)$$

On the other hand:

$${S_p}^2 = \frac{1}{n+m-2}\left(\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2+\sum_{j=1}^{m}\left(Y_j-\overline{Y}\right)^2\right) = \frac{(n-1){S_x}^2+(m-1){S_y}^2}{n+m-2}$$

where ${S_x}^2$ and ${S_y}^2$ are the respective sample variances. By Fisher's theorem we know that ${S_x}^2 \sim \frac{\sigma^2}{n-1}{\chi_{n-1}}^2$ and ${S_y}^2 \sim \frac{\sigma^2}{m-1}{\chi_{m-1}}^2$. So, recalling that ${\chi_a}^2 + {\chi_b}^2 \sim {\chi_{a+b}}^2$ we have that:

$$
\begin{aligned}
{S_p}^2 = \frac{(n-1){S_x}^2 + (m-1){S_y}^2}{n+m-2} &\sim \frac{(n-1)\frac{\sigma^2}{n-1}{\chi_{n-1}}^2 + (m-1)\frac{\sigma^2}{m-1}{\chi_{m-1}}^2}{n+m-2} \\
&\sim \frac{\sigma^2{\chi_{n-1}}^2 + \sigma^2{\chi_{m-1}}^2}{n+m-2} \\
&\sim \frac{\sigma^2}{n+m-2}{\chi_{n+m-2}}^2
\end{aligned}
$$

Finally:

$$
T = \frac{\overline{X} - \overline{Y}}{\sqrt{{S_p}^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{\frac{\overline{X}-\overline{Y}}{\sqrt{\sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)}}}{\sqrt{\frac{{S_p}^2}{\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{{\chi_{n+m-2}}^2}{n+m-2}}} = t_{n+m-2}
$$

which follows from the definition of the Stundent's t-distribution (quotient of a standard normal distribution and the square root of a chi-square random variable divided by its degrees of freedom).

3. *Samples of wood were obtained from the core and periphery of a certain Byzantine church. The date of the wood was determined, giving the following data:*

Core	1294	1279	1274	1264	1263	1254	1251	1251	1248	1240	1232	1220	1218	1210
Periphery	1284	1272	1256	1254	1242	1274	1264	1256	1250					

Use the two-sample t-test to determine if the mean age of the core is the same as the mean age of the periphery.

Resolution. We will do the problem with a level of significance $\alpha = 0.05$. We assign the letter X to the data of Core and the letter Y to the Periphery's one. In this case, we have $n = 14$, $m = 9$, $\overline{X} = 1249.857$, $\overline{Y} = 1261.333$ and ${S_p}^2 = 433.129$. Therefore, the observed value t_0 of T is: $t_0 = -1.29066$. The p-value of the test will be given by:

$$
p := \mathbb{P}(|T| \geq |t_0|) = \mathbb{P}(T \geq |t_0|) + \mathbb{P}(T \leq -|t_0|) = 2\mathbb{P}(T \geq |t_0|) = 0.2109
$$

Since $p > \alpha = 0.05$, we can't reject H_0.