# Solution of exercise 4 of PC8

*Octobre 16 2018*

## Likelihood ratio test and Neyman-Pearson

In PC7, we saw that the power function was a good way to compare two hypothesis tests with the same significance level. Here we present a test, namely the likelihood ratio test, which is automatically the best, i.e. with the smallest type II error for a fixed significance level when the hypotheses to test are simple, i.e. $\text{Card}(\Theta_1) = \text{Card}(\Theta_0) = 1$ (Neyman-Pearson Lemma).

**Exercise 4: Swimming pool**

A swimming pool seller wants to compare two products which kills bacteria. Both products guarantee that 95% of bacteria are eliminated. Yet the pH may become more basic so more harmful for users. Thus he does an experiment with two swimming pools in his shop. He puts product $A$ in pool 1 and product $B$ in pool 2. He has a pH meter but he doubts its reliability. Therefore he does 10 measurements in each swimming pool.

In pool 1, he observes $x_1, ..., x_{10}$ with respective values

$$7.33\ ;\ 6.17\ ;\ 7.46\ ;\ 8.13\ ;\ 6.68\ ;\ 6.76\ ;\ 7.97\ ;\ 6.76\ \ 6.81\ ;\ 8.40\ .$$

The empirical mean is $\bar{x}_{10} = 7.247$ and the empirical variance (the one divided by $n-1$) is $\sqrt{\sigma_x^2} = 0.73$.

In pool 2, he observes $y_1, ..., y_{10}$ with respective values

$$10.40\ ;\ 7.27\ ;\ 8.99\ ;\ 7.28\ ;\ 9.18\ ;\ 9.10\ ;\ 7.96\ ;\ 7.71\ ;\ 9.59\ ;\ 9.61\ .$$

The empirical mean is $\bar{y}_{10} = 8.709$ and the empirical variance (the one divided by $n-1$) is $\sqrt{\sigma_y^2} = 1.08$.

The seller does not have any preference beforehand and wants to give the best advice to his customers. He wants to clearly say "prefer A", "prefer B" or "do as you wish". We know that $pH = 7$ is neutral for the skin while $pH = 9$ is basic and harmful.

1. Assume that the observations are realizations of independent Gaussian random variables with variance $\sigma^2$ and mean $m_1$ for pool 1, and $m_2$ for pool 2. Give the statistical model considering a Gaussian vector. Compute the mean and the covariance matrix.

We observe $x_i$ the i-th measured pH in swimming pool 1 and $y_i$ the i-th measured pH in swimming pool 2 for $i \leqslant n = 10$. We assume that $(x_1, \dots, x_n)$ and $(y_1, \dots, y_n)$ are realizations of $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ respectively where $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ are two independent random vectors, $X_i$ are i.i.d. $\mathcal{N}(m_1, \sigma^2)$ and $Y_i$ are i.i.d. $\mathcal{N}(m_2, \sigma^2)$. Then $(X_1, \dots, X_n, Y_1, \dots, Y_n)$ is a Gaussian vector with mean vector $(\underbrace{m_1, \dots, m_1}_{10 \text{ times}}, \underbrace{m_2, \dots, m_2}_{10 \text{ times}})$ and covariance matrix $\sigma^2 I$, where $I$ is the identity matrix of size $20 \times 20$.
Then the statistical model is

$$\left(\mathbb{R}^{2n}, \mathcal{B}(\mathbb{R}^{2n}), \{\mathcal{N}((m_1, \dots, m_1, m_2, \dots, m_2), \sigma^2 I), m_1 \in \mathbb{R}, m_2 \in \mathbb{R}, \sigma > 0\}\right).$$

Note that the likelihood is

$$\mathcal{L}(X_1, \dots, Y_n; (m_1, m_2)) = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{20}} \exp\left(-\frac{\sum_{i=1}^n (X_i - m_1)^2 + \sum_{i=1}^n (Y_i - m_2)^2}{2\sigma^2}\right).$$

2. Assume that $\sigma^2 = 1$ is known.

a. Give a likelihood ratio test $H_0\ :\ m_1 = m_2 = 7$ against $H_1\ :\ m_1 = 7,\ m_2 = 9$ at level $\alpha$. Simplify the result as much as possible. Compute the p-value.

We want to test $H_0 \ : \ m_1 = m_2 = 7$ against $H_1 \ : \ m_1 = 7, \ m_2 = 9$ at level $\alpha$. The likelihood ratio test enables to automatically determine the statistic of the test. The likelihood ratio test function is by definition

$$\phi_\alpha^*(X_1, \dots, Y_n) = \begin{cases} 1 & \text{if } \frac{\mathcal{L}(X_1,\dots,Y_n;(m_1=7,m_2=9))}{\mathcal{L}(X_1,\dots,Y_n;(m_1=7,m_2=7))} > c_\alpha \\ \gamma_\alpha & \text{if } \frac{\mathcal{L}(X_1,\dots,Y_n;(m_1=7,m_2=9))}{\mathcal{L}(X_1,\dots,Y_n;(m_1=7,m_2=7))} = c_\alpha \\ 0 & \text{otherwise} \end{cases}$$

for some $c_\alpha$ and $\gamma_\alpha$. **Here are the steps to follow to make the likelihood ratio test explicit**:

1) We first translate the conditions in the definition of the test function on the likelihood ratio thanks to a more interpretable test statistic by studying the **monotonicity** of the likelihood ratio. Here we study the log-likelihood ratio since the likelihoods are always positive:

$$\begin{aligned} \log\left(\frac{\mathcal{L}(X_1, \dots, Y_n; (m_1 = 7, m_2 = 9))}{\mathcal{L}(X_1, \dots, Y_n; (m_1 = 7, m_2 = 7))}\right) &= \log\left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}^{20}} \exp\left(-\frac{\sum_{i=1}^n (X_i-7)^2 + \sum_{i=1}^n (Y_i-9)^2}{2\sigma^2}\right)}{\frac{1}{\sqrt{2\pi\sigma^2}^{20}} \exp\left(-\frac{\sum_{i=1}^n (X_i-7)^2 + \sum_{i=1}^n (Y_i-7)^2}{2\sigma^2}\right)}\right) \\ &= \frac{\sum_{i=1}^n (Y_i-7)^2 - \sum_{i=1}^n (Y_i-9)^2}{2\sigma^2} \\ &= \frac{2n}{\sigma^2}\frac{1}{n}\sum_{i=1}^n Y_i - \frac{16n}{\sigma^2}. \end{aligned}$$

Then The likelihood ratio is an increasing function of $\frac{1}{n}\sum_{i=1}^n Y_i$. Then the previous test function can be written as follows:

$$\phi_\alpha^*(X_1, \dots, Y_n) = \begin{cases} 1 & \text{if } \frac{1}{n}\sum_{i=1}^n Y_i > k_\alpha \\ \gamma_\alpha & \text{if } \frac{1}{n}\sum_{i=1}^n Y_i = k_\alpha \\ 0 & \text{otherwise} \end{cases}$$

for some other constant $k_\alpha$. This step enables to choose a test statistic, namely here $\frac{1}{n}\sum_{i=1}^n Y_i$. We can then answer the usual steps 3 and 4.

2) Step 5: we then **choose $\gamma_\alpha$ and $k_\alpha$ such that the significance level equals to $\alpha$.**

$$\alpha = \mathbb{E}_{H_0}(\phi_\alpha^*(X_1, \dots, Y_n)) = \gamma_\alpha \mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n Y_i = k_\alpha\right) + 1\mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n Y_i > k_\alpha\right).$$

The value of $\gamma_\alpha$ can be chosen equals to 0 since $\mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n Y_i = k_\alpha\right) = 0$ ($\gamma_\alpha$ only has a role when the observations are discrete as in Exercise 3, and can be deleted in the case of continuous observations). Then the likelihood ratio test becomes a pure test. Finally we search for $k_\alpha$ such that

$$\begin{aligned} \alpha = \mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n Y_i > k_\alpha\right) &= \mathbb{P}_{Y_i \sim \mathcal{N}(7,\sigma^2)}\left(\frac{1}{n}\sum_{i=1}^n Y_i > k_\alpha\right) = \mathbb{P}_{\frac{1}{n}\sum_{i=1}^n Y_i \sim \mathcal{N}(7,\sigma^2/n)}\left(\frac{1}{n}\sum_{i=1}^n Y_i > k_\alpha\right) \\ &= \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\left(Z > \sqrt{n}\frac{k_\alpha - 7}{\sigma}\right) = 1 - F_{\mathcal{N}(0,1)}\left(\sqrt{n}\frac{k_\alpha - 7}{\sigma}\right) \end{aligned}$$

i.e. $\sqrt{n}\frac{k_\alpha - 7}{\sigma} = q_{1-\alpha}$, where $q_{1-\alpha}$ is the $1-\alpha$-quantile of the standard Normal distribution. Finally we choose $k_\alpha = 7 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}$.

Step 6: We reject $H_0$ when $\frac{1}{n}\sum_{i=1}^n Y_i > 7 + \frac{\sigma}{\sqrt{n}} q_{1-\alpha}$.

Step 7 (p-value): The p-value is

$$\hat{\alpha}(x_1, \dots, y_n) = \mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n Y_i > \frac{1}{n}\sum_{i=1}^n y_i\right) = 1 - F_{\mathcal{N}(0,1)}\left(\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n y_i - 7}{\sigma}\right) \simeq 3e^{-8}.$$

```r
y= c(10.40 , 7.27, 8.99, 7.28, 9.18, 9.10,  7.96, 7.71,  9.59, 9.61)
n=10
sigma=1
c=sqrt(n)*(mean(y)-7)/sigma
1-pnorm(c)
```

```
## [1] 3.252509e-08
```

Looking at the p-value, we reject $H_0$ unless we choose a significance level smaller than $3e^{-8}$.

b. Explain why this test is uniformly most powerful among the hypothesis tests at level $\alpha$ which test the same hypotheses.

Using Neyman-Pearson Theorem, since the hypotheses are simple, the likelihood ratio test is uniformly most powerful among the hypothesis tests at level $\alpha$. It means that for any other hypothesis tests $\psi$ at most level $\alpha$ ( i.e. such that $\mathbb{E}_{H_0}(\psi) \leqslant \alpha$) then its power is smaller or equal to the power of the likelihood ratio test:

$$\mathbb{E}_{H_1}(\psi) \leqslant \mathbb{E}_{H_1}(\phi^*_\alpha).$$

3. We still assume that $\sigma^2 = 1$ is known, test $H_0 \;\; m_1 = m_2$ against $H_1 \;:\; m_1 < m_2$ at level $\alpha$.

We can choose the following test statistic: $\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n Y_i$, under $H_0$, this statistic should be around 0, and is distributed as $\mathcal{N}\left(0, 2\frac{\sigma^2}{n}\right)$; and under $H_1$ it should be small (negative). Then we reject $H_0$ when $\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n Y_i \leqslant c$.

The critical value $c$ is chosen such that

$$\alpha = \mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n Y_i \leqslant c\right) = \mathbb{P}_{Z\sim\mathcal{N}\left(0,2\frac{\sigma^2}{n}\right)}\left(Z \leqslant c\right) = \mathbb{P}_{Z\sim\mathcal{N}(0,1)}\left(Z \leqslant \frac{\sqrt{n}c}{\sqrt{2}\sigma}\right) = F_{\mathcal{N}(0,1)}\left(\frac{\sqrt{n}c}{\sqrt{2}\sigma}\right)$$

i.e. $c = \frac{\sqrt{2}\sigma}{\sqrt{n}} q_\alpha$.

Finally we reject $H_0$ if $\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n Y_i \leqslant \frac{\sqrt{2}\sigma}{\sqrt{n}} q_\alpha$.

The p-value is

$$\begin{aligned}\hat{\alpha}(x_1, \ldots, y_n) &= \mathbb{P}_{H_0}\left(\frac{1}{n}\sum_{i=1}^n X_i - \frac{1}{n}\sum_{i=1}^n Y_i \leqslant \frac{1}{n}\sum_{i=1}^n x_i - \frac{1}{n}\sum_{i=1}^n y_i\right) \\ &= F_{\mathcal{N}(0,1)}\left(\sqrt{n}\frac{\frac{1}{n}\sum_{i=1}^n x_i - \frac{1}{n}\sum_{i=1}^n y_i}{\sqrt{2}\sigma}\right) \simeq 0.0005.\end{aligned}$$

```r
x=c(7.33,6.17,  7.46, 8.13,  6.68, 6.76, 7.97 , 6.76 , 6.81 , 8.40)
y= c(10.40 , 7.27, 8.99, 7.28, 9.18, 9.10,  7.96, 7.71,  9.59, 9.61)
n=10
sigma=1
c=sqrt(n)*(mean(x)-mean(y))/(sqrt(2)*sigma)
c
```

```
## [1] -3.269131
```

```r
pnorm(c)
```

```
## [1] 0.000539391
```

Looking at the p-value, we reject $H_0$ unless we choose a significance level smaller than 0.0005.

4. Now, we assume that $\sigma^2$ is unknown (more realistic for this problem statement).

a. Give the distribution of $Z = \bar{Y}_n - \bar{X}_n$ and of $W = 9\sigma_X^2 + 9\sigma_Y^2$.

Using the Student/Gosset's Theorem to $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$:

- $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and $\sigma_X^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ are independent,

- $\bar{X}_n \sim \mathcal{N}(m_1, \sigma^2/n)$,

- $(n-1)\frac{\sigma_X^2}{\sigma^2} \sim \chi^2(n-1)$,

- $\bar{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\sigma_Y^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$ are independent,

- $\bar{Y}_n \sim \mathcal{N}(m_2, \sigma^2/n)$,

- $(n-1)\frac{\sigma_Y^2}{\sigma^2} \sim \chi^2(n-1)$.

Then, since $(X_1, \dots, X_n)$ and $(Y_1, \dots, Y_n)$ are independent:

- $Z = \bar{Y}_n - \bar{X}_n$ and $W = 9\sigma_X^2 + 9\sigma_Y^2$ are independent,

- $Z \sim \mathcal{N}(m_2 - m_1, 2\sigma^2/n)$, and so under $H_0$, $U = -\frac{\sqrt{n}}{\sqrt{2}\sigma} Z \sim \mathcal{N}(0,1)$,

- $V = \frac{W}{\sigma^2} = 9\frac{\sigma_X^2}{\sigma^2} + 9\frac{\sigma_Y^2}{\sigma^2} = (n-1)\frac{\sigma_X^2}{\sigma^2} + (n-1)\frac{\sigma_Y^2}{\sigma^2} \sim \chi^2(2(n-1))$ (indeed if $A \sim \chi^2(d_A)$ and independently $B \sim \chi^2(d_B)$ then $A + B \sim \chi^2(d_A + d_B)$).

b. Give a hypothesis test $H_0 : m_1 = m_2$ against $H_1 : m_1 < m_2$ at level $\alpha$. Compute the p-value.

We cannot use the statistic $-Z$ we were using in question 3, because we don't know its distribution anymore under $H_0$ since $\sigma$ is not known. Instead we use the previous computations. We know that under $H_0$, $U$ and $V$ are independent, $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2(2(n-1))$, then $\frac{U}{\sqrt{V/(2(n-1))}} \sim \mathcal{T}(2(n-1))$. Then under $H_0$:

$$\frac{U}{\sqrt{V/(2(n-1))}} = \frac{\frac{\sqrt{n}}{\sqrt{2}\sigma} Z}{\sqrt{\frac{W}{2(n-1)\sigma^2}}} = \sqrt{n}\frac{\bar{X}_n - \bar{Y}_n}{\sqrt{\sigma_X^2 + \sigma_Y^2}} \sim \mathcal{T}(2(n-1)).$$

$S = \sqrt{n}\frac{\bar{X}_n - \bar{Y}_n}{\sqrt{\sigma_X^2 + \sigma_Y^2}}$ is the statistic of our test. And we reject $H_0$ when this statistic is small : $S < s_c$. $s_c$ is chosen such that

$$\alpha = \mathbb{P}_{H_0}(S < s_c) = F_{\mathcal{T}(2(n-1))}(s_c)$$

i.e. $s_c = t_{2(n-1),\alpha}$, where $t_{2(n-1),\alpha}$ is the $\alpha$ quantile of the $\mathcal{T}(2(n-1))$ distribution.

Finally we reject $H_0$ when $S = \sqrt{n}\frac{\bar{X}_n - \bar{Y}_n}{\sqrt{\sigma_X^2 + \sigma_Y^2}} < t_{2(n-1),\alpha}$. This is the famous (one-sided)-**two-sample t-test** with equality of standard deviation.

The p-value is

$$\hat{\alpha}(x_1, \dots, y_n) = \mathbb{P}_{H_0}\left(\sqrt{n}\frac{\bar{X}_n - \bar{Y}_n}{\sqrt{\sigma_X^2 + \sigma_Y^2}} < s_{obs}\right) = F_{\mathcal{T}(2(n-1))}(s_{obs}) \simeq 0.0012.$$

```
x=c(7.33,6.17,  7.46, 8.13,  6.68, 6.76, 7.97 , 6.76 , 6.81 , 8.40)
y= c(10.40 , 7.27, 8.99, 7.28, 9.18, 9.10,  7.96, 7.71,  9.59, 9.61)
n=10
mean(x)
```

```
## [1] 7.247
```

```
sigma_x=sd(x)
sigma_x
```

```
## [1] 0.7326368
```

```r
mean(y)
```

```
## [1] 8.709
```

```r
sigma_y=sd(y)
sigma_y
```

```
## [1] 1.083866
```

```r
s_obs=sqrt(n)*(mean(x)-mean(y))/(sqrt(sigma_x^2+sigma_y^2))
s_obs
```

```
## [1] -3.533915
```

```r
pt(s_obs,df=18)
```

```
## [1] 0.001185601
```

Finally we reject $H_0$ at level 0.05.

With the R function **t.test**, we obtain the same results

```r
t.test(x,y,alternative = "less",var.equal=T)
```

```
##
##  Two Sample t-test
##
## data:  x and y
## t = -3.5339, df = 18, p-value = 0.001186
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##        -Inf -0.7446084
## sample estimates:
## mean of x mean of y
##     7.247     8.709
```

c. What should be the conclusion of the seller?

The seller should recommend product $A$.