Advanced Statistical Inference

Assesment exercises Unit 2

Laura Arribas and Arnau Garcia

An assessment presented for the course of
Advanced Statistical Inference 2023/24 (MESIO)

Facultat de Matemàtiques i Estadística
Universitat Politècnica de Catalunya
Date of submission: 16/Nov/2023
Teachers: Guadalupe Gómez and Alex Sànchez

Contents

Chapter 1

Problem 1

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables having a two-parameter exponential distribution, i.e., $X \sim \mu + exp(\lambda), -\infty < \mu < \infty, 0 < \lambda < \infty$. Let $X_{(1)} \leq \cdots \leq X_{(n)}$ be the order statistic.

(a) Apply the Factorization Theorem to prove that $X_{(1)}$ and $S = \sum_{i=2}^{n} \left(X_{(i)} - X_{(1)}\right)$ are sufficient statistics.

(b) Derive the conditional p.d.f. of $\mathbf{X}$ given $\left(X_{(1)}, S\right)$.

(c) How would you generate an equivalent sample $\mathbf{X}'$ (by simulation) when the value of $\left(X_{(1)}, S\right)$ are given?

Solution:

(a): We firstly find the pdf of the random variable X using the change of variable Theorem. Let $Y \sim exp(\lambda)$ be a random variable following a exponential distribution with parameter λ (the one on the hypothesis). Then, we know that the pdf of Y is

$$f_Y(y) = \lambda \exp(-\lambda y) \mathbb{1}_{\{x \geq 0\}}. \tag{1.1}$$

Now, let $\mu \in \mathbb{R}$ be the one of the hypothesis, the function $g(z) = \mu + z$ is clearly continuous and invertible, with inverse $g^{-1}(z) = z - \mu$. Thus, one can apply the change of variable Theorem and we have that the pdf of the random variable $X = g(Y) = \mu + Y$ is

$$f_X(x) = f_Y(g^{-1}(x)) \left| \frac{\partial (g^{-1}(x))}{\partial x} \right| = f_Y(x - \mu) = \lambda \exp(-\lambda(x - \mu)) \mathbb{1}_{\{x > \mu\}}. \tag{1.2}$$

Where the indicator is added in order to keep the negativity of the quantity inside the exponential function. Now, using the pdf of X we compute the joint density:

$$f(x_1,\dots,x_n;\lambda,\mu) = \prod_{i=1}^{n} \lambda \exp\left(-\lambda(x_i-\mu)\right)\mathbb{1}_{\{x_i>\mu\}}$$
$$= \lambda^n \exp\big(-\lambda \sum_{i=1}^{n}(x_i-\mu)\big)\big(\prod_{i=1}^{n}\mathbb{1}_{\{x_i>\mu\}}\big)$$

And now, we will derive the expressions for the indicator term and for the sum inside the exponential. We start with the indicator:

$$\prod_{i=1}^{n}\mathbb{1}_{\{x_i>\mu\}} = \prod_{i=1}^{n}\mathbb{1}_{(-\infty,x_i)}(\mu) = \mathbb{1}_{(-\infty,x_{(1)})}(\mu).$$

Now, we can rewrite the sum inside the exponential as follows:

$$\sum_{i=1}^{n}\left(x_i-\mu\right) = \sum_{i=1}^{n}\left(x_{(i)}-\mu\right) = \sum_{i=1}^{n}\left[(x_{(i)}-x_{(1)})+(x_{(1)}-\mu)\right]$$
$$= \sum_{i=1}^{n}(x_{(i)}-x_{(1)}) + n(x_{(1)}-\mu) = \sum_{i=2}^{n}(x_{(i)}-x_{(1)}) + n(x_{(1)}-\mu).$$

In the last equalities we have used that sum n elements is the same that sum the n elements reordered. In addition, in the last equality we have removed the sum for the index $i=1$, because for this case we have the term $x_{(1)}-x_{(1)}=0$.

Then, using the notation of the hypothesis, one can write the joint density of X as

$$f(x_1,\dots,x_n;\lambda,\mu) = \lambda^n \exp\Big(-\lambda\big(S+n(x_{(1)})\big)\Big)\mathbb{1}_{(-\infty,x_{(1)})}(\mu)$$
$$= \lambda^n \exp\left(-\lambda S\right)\exp\big(-\lambda n(x_{(1)}-\mu)\big)\mathbb{1}_{(-\infty,x_{(1)})}(\mu)$$
$$= h(x_1,\dots,x_n)g(T(x_1,\dots,x_n);\lambda,\mu).$$

Where $h(x_1,\dots,x_n)=1$, and $g(T(x_1,\dots,x_n);\lambda,\mu)$ is the whole joint density. And, $T(X_1,\dots,X_n)=\big(X_{(1)},S\big)$. Thus, applying the factorization theorem we have that $T(X_1,\dots,X_n)=\big(X_{(1)},S\big)$ is a sufficient statistic.

(b): Our objective in this section will be to find the pdf of $f_{X_{(1)},S}(x,z)$ and use this function to compute $f_{X|X_{(1)},S}$.

Let us observe that

$$n\lambda(\overline{X}-\mu) = n\lambda(X_{(1)}-\mu)+\lambda S, \tag{1.3}$$

given that we can rewrite S as $n(\overline{X}-X_{(1)})$.

The distribution of $X_{(1)}$ can be derived as follows:

$$\begin{aligned} F_{X_{(1)}}(y) &= P(X_{(1)} \leq y) = 1 - P(X_{(1)} > y) \\ &= 1 - P((X_1 > y) \cap (X_2 > y) \cap \cdots \cap (X_n > y)) \\ &= 1 - \prod_{i=1}^{n}(1 - F_{X_1}(y)) = 1 - (1 - F_X(y))^n \\ &= 1 \left[1 - (1 - e^{-\lambda(y-\mu)})\right]^n = 1 - e^{-n\lambda(y-\mu)}, \end{aligned}$$

where we have used that the X_i are i.i.d.. Thus, we conclude that $X_{(1)} - \mu \sim exp(n\lambda)$, and, therefore, $n\left(X_{(1)} - \mu\right) \sim E(\lambda)$. From this, we can obtain the distribution of $2n\lambda\left(X_{(1)} - \mu\right)$, that is

$$2n\lambda\left(X_{(1)} - \mu\right) \sim E(\frac{1}{2}) \equiv \chi^2_2. \tag{1.4}$$

Furthermore, from the hypothesis we know that $X_i - \mu \sim exp(\lambda)$. Let us denote $Y_i := X_i - \mu$. To find the distribution of $n\lambda\left(\overline{X} - \mu\right)$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}(Y_i + \mu) = \mu + \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Since $Y_i \sim exp(\lambda)$, then $\sum_{i=1}^{n} Y_i \sim \frac{1}{2\lambda}\mathcal{X}^2_{2n}$, so

$$2n\lambda(\overline{X} - \mu) \sim \mathcal{X}^2_{2n}. \tag{1.5}$$

We derive from equations 1.3, 1.4 and 1.5 that

$$2\lambda S \sim \mathcal{X}^2_{2(n-1)},$$

where we have used that the sum of random variables that follow a Xi-squared distribution is equal to a Xi-squared with degrees of freedom equal to the sum.

Knowing this relation we can easily compute the density function of S:

$$\begin{aligned} F_S(s) &= P(S \leq s) = P(2\lambda S \leq 2\lambda s) = F_{2\lambda S}(2\lambda s) \\ &= \frac{1}{\Gamma\left(\frac{2(n-1)}{2}\right)}\gamma\left(\frac{2(n-1)}{2}, \frac{2\lambda s}{2}\right) = \frac{1}{\Gamma(n-1)}\gamma(n-1, \lambda s). \end{aligned}$$

We can now calculate the value of the second term as an integral:

$$\gamma(n-1, \lambda z) = \int_0^{\lambda s} t^{n-2}e^{-1}dt = \Gamma(n-1) - \Gamma(n-1, \lambda s).$$

Inserting this into the previous equation, we derive

$$F_S(s) = 1 - \frac{\Gamma(n-1, \lambda s)}{\Gamma(n-1)}.$$

Note that the the negative term is the cdf of an Inverse Gamma with parameters $(n-1, \lambda)$ and variable $\frac{1}{s}$. Then,

$$F_S(s) = 1 - F_{InvGamma(n-1,\lambda)}(\frac{1}{s}).$$

To obtain the pdf of S we just need to derive this expression.

$$\begin{aligned} f_S(s) &= \frac{d}{ds}(1 - F_{InvGamma(n-1,\lambda)}(\frac{1}{s}) = -f_{InvGamma(n-1,\lambda)}(-\frac{1}{s^2}) \\ &= \frac{\lambda^{n-1}}{\Gamma(n-1)} \left(\frac{1}{s}\right)^{-n+2} e^{-\lambda s} = \frac{\lambda^{n-1}}{\Gamma(n-1)} s^{n-2} e^{-\lambda s} \end{aligned}$$

Next, let us compute the pdf of $X_{(1)}$. We have already seen that $X_{(1)} \sim exp(n\lambda) + \mu$, thus:

$$f_{X_{(1)}}(y) = n\lambda e^{-n\lambda(y-\mu)}.$$

Since $X_{(1)}$ and S are independent random variables, the joint distribution function can be expressed as the product of the two pdfs:

$$f_{X_{(1)},S}(y, s) = f_{X_{(1)}}(y) \cdot f_S(s) = \frac{n\lambda^n}{\Gamma(n-1)} s^{n-2} e^{-\lambda(n(y-\mu)+s)}$$

Given that in (a) we have proved that $X_{(1)}$ and S are sufficient statistics, we can finally compute $f_{X|(X_{(1)},S)}(x|y, s)$:

$$\begin{aligned} f_{X|(X_{(1)},S)}(x|y, s) &= \frac{f_X(x)}{f_{X_{(1)},S}(y, s)} \\ &= \frac{\Gamma(n-1)\lambda^n \exp\left(-\lambda(s + n(x_{(1)}))\right)}{n\lambda^n s^{n-2} e^{-\lambda(n(y-\mu)+s)}} \mathbb{1}_{(-\infty, x_{(1)})}(\mu) \\ &= \frac{\Gamma(n-1)}{n s^{n-2}} \mathbb{1}_{(-\infty, x_{(1)})}(\mu) \end{aligned}$$

That is what we wanted to find.

(c): To generate an equivalent sample I would use the so-called *Inversion Method*, which is a technique used to generate random samples from a probability distribution. This method is particularly powerful when the CDF of the distribution is invertible. Notwithstanding, if this is not the case we can use the empirical CDF and achieve also good results. Let X be the random variable which distribution is known, and let $F = cdf(X)$ be the CDF of X.

What we do in the Inversion Method is:

1. Compute the inverse of the known CDF F. If we can compute this inverse analytically, we should do it (for instance in distributions as the exponential we ca do it). If not, we use the empirical CDF $\hat{F}_n(x)$.

2. We generate a random variable $U \sim U(0,1)$.

3. Let F^{-1} the inverse computed in the first step. We use U, F^{-1} to generate the desired sample, doing $F^{-1}(U)$. If we have used the empirical CDF then we must use the Empirical Inverse CDF. This is, we must find the smallest x such that $\hat{F}_n(x) \geq U$. And the value x obtained is a random sample from the empirical distribution.

Now, in order to apply the inversion method in our specific case we should use the results obtained in **(b)**. As we saw in **(b)**, the distribution of **X** given $(X_{(1)}, S)$ is known. And also using what we saw in **(b)** we can deduce that

$$f_X(x) = f_{X_{(1)},S}(x', s) f_{X|X_{(1)},S}(x|x', s),$$

where the two pdf in the right part of the equality are well-known. Thus, given $(X_{(1)}, S)$ we know the pdf of X, and in addition the CDF. And then we can apply the *Inversion Method* to generate an equivalent sample by simulation.

Chapter 2

Problem 2

Consider the trinomial distribution $M(n, p_1, p_2), 0 < p_1, p_2, p_1 + p_2 < 1$.
(a) Show that the Fisher Information Matrix is

$$I(p_1, p_2) = \frac{n}{1 - p_1 - p_2} \begin{bmatrix} \frac{1-p_2}{p_1} & 1 \\ 1 & \frac{1-p_1}{p_2} \end{bmatrix}$$

(b) For the Hardy-Weinberg mode $1, p_1(\theta) = \theta^2, p_2(\theta) = 2\theta(1-\theta)$, derive the Fisher information function

$$I(\theta) = \frac{2n}{\theta(1-\theta)}$$

Solution:

(a): Following the definition given in Pennsylvania State University we have that if we repeat an experiment n independent times, with each experiment ending in one of three mutually exclusive and exhaustive ways (success, first kind of failure, second kind of failure). If X denote the number of times the experiment results in a success, Y the number of first kind failure and Z the number of second kind failure, then we can write the joint probability mass function of X, Y as

$$f(x, y; p_1, p_2) = P(X = x, Y = y) = \frac{n!}{x!y!(n-x-y)!} p_1^x p_2^y (1 - p_1 - p_2)^{n-x-y} \quad (2.1)$$

Where $x, y = 0, 1, \ldots, n$, $x + y \leq n$. And in addition p_1 is the probability of success, p_2 the probability of failure of type 1 and p_3 the probability of failure of type 2. But using the relation $p_1 + p_2 + p_3 = 1$ we can isolate $p_3 = 1 - p_1 - p_2$ and write the probability mass function as a function of p_1, p_2 only (as we did previously).

Now, we will derive the Fisher information matrix from this probability mass function. We will compute the components of the matrix separately. First we take the logarithm of the probability mass function:

$$\log\left(f(x,y;p_1,p_2)\right) = \log\left(\frac{n!}{x!y!(n-x-y)!}\right) + x\log\left(p_1\right) + y\log\left(p_2\right) + \\ + (n-x-y)\log\left(1-p_1-p_2\right).$$

Now we compute the derivative of with respect p_1:

$$\frac{\partial \log\left(f(x,y;p_1,p_2)\right)}{\partial p_1} = \frac{x}{p_1} - \frac{n-x-y}{1-p_1-p_2}.$$

The second derivative with respect p_1:

$$\frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_1^2} = -\frac{x}{p_1^2} - \frac{n-x-y}{(1-p_1-p_2)^2}$$

Notice now that:

- $E(X) = np_1$.
- $E(Y) = np_2$.

These two results are direct if one observe that the marginals distributions X and Y are a binomial distribution. Then, using the well-known expectation of a binomial distribution the result holds.

Now, we use this and the previous calculus for compute the first component (i.e. first row first column) of the Fisher Information Matrix:

$$E\left(-\frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_1^2})\right) = E\left(\frac{x}{p_1^2} + \frac{n-x-y}{(1-p_1-p_2)^2}\right) \\ = \frac{n}{p_1} + \frac{n-np_1-np_2}{(1-p_1-p_2)^2} = n\left(\frac{1}{p_1} + \frac{1}{1-p_1-p_2}\right) = \frac{n}{1-p_1-p_2}\left(\frac{1-p_2}{p_1}\right)$$

And this is exactly the first component (ie first row first column) of the Fisher Information Matrix of the statements.

Now, we find the last component (second row second column) of the Fisher information Matrix. This is the component associated with the second derivative of the log-likelihood with respect of p_2. Thus, we will be doing very similar calculus than the ones done before. Since the calculus are very similar, we will summarise it.

The first derivative of the log-probability mass function with respect p_2:

$$\frac{\partial \log\left(f(x,y;p_1,p_2)\right)}{\partial p_2} = \frac{y}{p_2} - \frac{n-x-y}{1-p_1-p_2}.$$

The second derivative with respect p_2:

$$\frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_2^2} = -\frac{y}{p_2^2} - \frac{n-x-y}{(1-p_1-p_2)^2}$$

And then, computing the corresponding expectation as we did before we obtain the following:

$$E\left(-\frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_2^2})\right) = E\left(\frac{y}{p_2^2} + \frac{n-x-y}{(1-p_1-p_2)^2}\right)$$
$$= \frac{n}{p_2} + \frac{n-np_1-np_2}{(1-p_1-p_2)^2} = n\left(\frac{1}{p_2} + \frac{1}{1-p_1-p_2}\right) = \frac{n}{1-p_1-p_2}\left(\frac{1-p_1}{p_2}\right).$$

And then we have obtained the last component (second row, second column) of the matrix shown in the hypothesis.

Finally, we will find the two remaining components. These components are the associated with the cross derivatives. Since

$$\frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_1 \partial p_2} = \frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_2 \partial p_1}$$

we only need to compute one of the two components, because the other one will be identical.

We take the first derivative of the log-pmf with respect p_1 and we compute the derivative with respect p_2:

$$\frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_1 \partial p_2} = -\frac{n-x-y}{(1-p_1-p_2)^2}.$$

And then

$$E\left(-\frac{\partial^2 \log\left(f(x,y;p_1,p_2)\right)}{\partial p_1 \partial p_2} =\right) = \frac{n-np_1-np_2}{(1-p_1-p_2)^2} = \frac{n}{1-p_1-p_2}.$$

Then, we have obtained the two components in the statements. And so, finally, we have we have demonstrated that the Fisher Information Matrix of the trinomial distribution is the one shown in the statements.

(b): Let us return to the log-pmf used before in the previous section, which was:

$$\log\left(f(x,y;p_1,p_2)\right) = \log\left(\frac{n!}{x!y!(n-x-y)!}\right) + x\log\left(p_1\right) + y\log\left(p_2\right) + \\ + (n-x-y)\log\left(1-p_1-p_2\right)$$

as seen in section **(a)**. In our case, $p_1(\theta) = \theta^2$ and $p_2(\theta) = 2\theta(1-\theta)$, therefore,

$$\begin{aligned}\log\left(f(x,y;p_1,p_2)\right) &= \log\left(\frac{n!}{x!y!(n-x-y)!}\right) + x\log\left(\theta^2\right) + y\log\left(2\theta(1-\theta)\right) + \\ &\quad + (n-x-y)\log\left(1-\theta^2-2\theta(1-\theta)\right) \\ &= \log\left(\frac{n!}{x!y!(n-x-y)!}\right) + x\log(\theta^2) + y\log(2\theta) + y\log(1-\theta) + \\ &\quad + (n-x-y)\log\left((1-\theta)^2\right).\end{aligned}$$

Now we can compute the first and second derivatives of l_X with respect to θ:

$$\begin{aligned}\frac{\partial\log\left(f(x,y;p_1,p_2)\right)}{\partial\theta} &= \frac{2x}{\theta} + \frac{y}{\theta} - \frac{y}{1-\theta} - \frac{2(n-x-y)}{1-\theta} \\ \frac{\partial^2\log\left(f(x,y;p_1,p_2)\right)}{\partial^2\theta} &= \frac{-2x}{\theta^2} - \frac{-y}{\theta^2} - \frac{y}{(1-\theta)^2} - \frac{2(n-x-y)}{(1-\theta)^2}.\end{aligned}$$

Finally, we only need to determine the value of $E\left[-\frac{\partial^2\log\left(f(x,y;\theta)\right)}{\partial^2\theta}\right]$.

$$\begin{aligned}E\left[-\frac{\partial^2\log\left(f(x,y;\theta)\right)}{\partial^2\theta}\right] &= \frac{2}{\theta^2}E[x] + \frac{1}{\theta^2}E[y] + \frac{n}{(1-\theta)^2}E[y] + \frac{2}{(1-\theta)^2}(n-E[x]-E[y]) \\ &= \frac{2}{\theta^2}n\theta^2 + \frac{1}{\theta^2}2n\theta(1-\theta) + \frac{1}{(1-\theta)^2}2n\theta(1-\theta) + \frac{2n}{(1-\theta)^2}n(1-\theta)^2 \\ &= 2n + \frac{2n(1-\theta)}{\theta} + \frac{2n\theta}{(1-\theta)} + 2n \\ &= 2n\left(2 + \frac{1-\theta}{\theta} + \frac{\theta}{1-\theta}\right) \\ &= 2n\frac{2\theta(1-\theta) + (1-\theta)^2 + \theta^2}{\theta(1-\theta)} \\ &= \frac{2n}{\theta(1-\theta)}.\end{aligned}$$

Thus, we have derived that $I_x(\theta) = E\left[-\frac{\partial^2\log\left(f(x,y;\theta)\right)}{\partial^2\theta}\right] = \frac{2n}{\theta(1-\theta)}$.

Chapter 3

Problem 3

Consider n identical systems that operate independently. It is assumed that the time till failure of a system has a $G\left(\frac{1}{\theta}, 1\right)$ distribution. Let $Y_1, Y_2, \ldots, Y_r$ be the failure times until the r th failure.

(a) Show that the total life $T_{n,r} = \sum_{i=1}^{r} Y_i + (n-r)Y_r$ is distributed like $\frac{\theta}{2}\chi^2[2r]$

(b) Construct the α-level UMP test of $H_0 : \theta \leq \theta_0$ against $H_1 : \theta > \theta_0$ based on $T_{n,r}$.

(c) What is the power function of the UMP test?

Solution:

(a): In this section of the problem we will be using several properties from different distributions. We firstly expose these properties and then we proof the statement.

1. If $X \sim exp(1/2)$ then $X \sim \mathcal{X}^2_2$. This property is seen in the problem 10 of the Unit 1 part 2 list of problems.

2. If $X_1, \ldots, X_n$ is a iid sample following $\mathcal{X}^2_r$, then $\sum_{i=1}^{n} X_i \sim \mathcal{X}^2_{rn}$.

3. If $X \sim exp(\lambda)$ then $kX \sim exp(\lambda/k)$ where $k \in \mathbb{R}$.

Now, we are able to start the proof of this section. After different attempts we have followed the next approach: we understand that the $Y_1, Y_2, \ldots, Y_r$ failure times until the rth failure are $T_{(1)}, \ldots, T_{(r)}$. This is, the order statistics of $T_1, \ldots, T_r$, where these random variables are the times till failure of the systems. These random variables are iid following $T \sim G(1/\theta, 1)$ by hypothesis. So, we have that $Y_i = T_{(i)}$.

Now, notice that the pdf of T is

$$f_T(x; \frac{1}{\theta}, 1) = \frac{(1/\theta)^1}{\Gamma(1)} x^{1-1} \exp\left(\frac{x}{\theta}\right) = \frac{1}{\theta} \exp\left(\frac{x}{\theta}\right).$$

This is, we can see that the pdf of T is the pdf of an exponential distribution with parameter $1/\theta$. Then, indeed, $T \sim G(1/\theta, 1)$ is equivalent to $T \sim exp(1/\theta)$. And so, from now on, we will use that $T \sim exp(1/\theta)$.

Now, we will see how the behaviour of the order statistics of T is:

- The first time failure, $T_{(1)}$, corresponds to the probability of failure of the whole n identical and independent systems. Then, it corresponds with the sum of n random variables following $exp(1/\theta)$. Thus, $T_{(1)} \sim n \exp(1/\theta)$, which is (using the property 3 exposed previously) $T_{(1)} \sim \exp(n/\theta)$.
- $T_{(2)} - T_{(1)}$ corresponds to the probability that one of the $n-1$ identical independent systems fails. Then $T_{(2)} - T_{(1)} \sim \exp(\frac{n-1}{\theta})$.
- In general, $T_{(i)} - T_{(i-1)}$ corresponds to the probability that some of the $n - i + 1$ remaining systems fail. Thus, $T_{(i)} - T_{(i-1)} \sim \exp(\frac{n-i+1}{\theta})$.

In addition, we can observe that $\frac{2}{\theta}(T_{(i)} - T_{(i-1)}) \sim \exp(\frac{n-i+1}{2})$ (again using property 3). Now, we will deal with the $T_{n,r}$ statistic introduced in the hypothesis:

$$\begin{aligned}\frac{2}{\theta}T_{n,r} &= \sum_{i=1}^{r}\frac{2}{\theta}T_{(i)} + (n-r)\frac{2}{\theta}T_{(r)} = \sum_{i=1}^{r-1}\frac{2}{\theta}T_{(i)} + (n-r+1)\frac{2}{\theta}T_{(r)} \\ &= \sum_{i=1}^{r-1}\frac{2}{\theta}T_{(i)} + (n-r+1)\frac{2}{\theta}T_{(r-1)} + X_1\end{aligned}$$

Where in the last equality X_1 is a random variable following an $exp(1/2)$. And this random variable appears because

$$\begin{aligned}(n-r+1)\frac{2}{\theta}T_{(r)} &= (n-r+1)\frac{2}{\theta}T_{(r-1)} + (n-r+1)\exp\left(\frac{n-r+1}{2}\right) \\ &= (n-r+1)\frac{2}{\theta}T_{(r-1)} + \exp\left(\frac{1}{2}\right)\end{aligned}$$

Where in the previous equalities $\exp(\cdot)$ means the exponential distribution. It is, we are committing an abuse of notation. Returning now to the equality of $(2/\theta)T_{n,r}$, what we will do is repeat the argument used:

$$\begin{aligned}\frac{2}{\theta}T_{n,r} &= \sum_{i=1}^{r-1}\frac{2}{\theta}T_{(i)} + (n-r+1)\frac{2}{\theta}T_{(r-1)} + X = \sum_{i=1}^{r-2}\frac{2}{\theta}T_{(i)} + (n-r+2)\frac{2}{\theta}T_{(r-1)} + X \\ &= \sum_{i=1}^{r-2}\frac{2}{\theta}T_{(i)} + (n-r+2)\frac{2}{\theta}T_{(r-2)} + X_1 + X_2 = \cdots = \sum_{j=1}^{r}X_j\end{aligned}$$

Where $X_1, \dots X_r$ are iid following an exponential distribution with parameter $1/2$. And now, using the property 1, the random variable $X \sim \exp\left(\frac{1}{2}\right)$ is also following a $\mathcal{X}^2_2$. And, applying the property 2 we have that $\sum_{j=1}^{r} X_j \sim \mathcal{X}^2_{2r}$. Hence, $T_{n,r}$ is distributed like $\frac{\theta}{2}\mathcal{X}^2_{2r}$.

(b): Using the previous section we know that $T_{n,r} \sim \frac{\theta}{2}\mathcal{X}^2_{2r}$. Notice that, the pdf for a random variable following a $\mathcal{X}^2_{2r}$ is

$$f(x) = \frac{x^{\frac{2r}{2}-1}\exp\left(-\frac{x}{2}\right)}{2^{\frac{2r}{2}}\Gamma\left(\frac{2r}{2}\right)} = \frac{x^{r-1}\exp\left(-\frac{x}{2}\right)(1/2)^r}{\Gamma(r)},$$

which is, exactly, the pdf of a random variable following $G(r, 1/2)$. And now, using the scaling properties of the Gamma distribution, we have that, if Z is a random variable following a $G(r, 1/2)$, then $\frac{\theta}{2}Z \sim G(r, 1/\theta)$. And then the pdf of the random variable $T_{n,r}$ is

$$f_{T_{n,r}}(t) = \frac{(1/\theta)^r}{\Gamma(r)} t^{r-1}\exp\left(-\frac{t}{\theta}\right).$$

Notice that this distribution belongs to the exponential family, because we can write the pdf as

$$f_{T_{n,r}}(t) = h(t)c(\theta)\exp\left(w(\theta)g(t)\right),$$

where $h(t) = t^{r-1}$, $c(\theta) = \frac{(1/\theta)^r}{\Gamma(r)}$, $g(t) = t$ and $w(\theta) = -1/\theta$. And $w(\theta)$ is increasing in θ. If we take a sample of m observations (let us denote these observations as $T^1_{n,r}, \dots, T^m_{n,r}$), then the sufficient and complete statistic (the privileged one) is $\sum_{i=1}^{m} T^i_{n,r}$. Now, we use the Slide 135 of the theory class, where we have the theoretical result about one sided UMP test for exponential families. We have by hypothesis that $\theta_1 > \theta_0$, and then we have that the rejection region

$$R = \{\sum_{i=1}^{m} t_i > k_\alpha\}$$

is the rejection region for the UMP test. Notice that $\sum_{i=1}^{m} T^i_{n,r}$ follows a $G(rm, 1/\theta)$, because the m random variables $T_{n,r}$ taken are independent identically distributed, following $G(r, 1/\theta)$. Thus, we have that $\alpha = P_{H_0}(x \in R)$, which is equivalent to

$$k_\alpha = F^{-1}_{\theta_0}(1-\alpha),$$

where $F_{\theta_0} = cdf(G(mr, 1/\theta_0))$.

(c): Using the definition of the power function seen in class we have that, under H_0, it is for θ such that $\theta \leq \theta_0$, the power function is

$$\eta(\theta) = \text{type 1 error} = \alpha.$$

And, under H_1, it is values of θ such that $\theta > \theta_0$, the power function corresponds to

$$\eta(\theta) = P_{\theta_1}(\sum_{i=1}^{m} t_i > k_\alpha).$$

And we can compute this, using that k_α is known, and computed in the previous section:

$$\eta(\theta) = P_{\theta_1}(\sum_{i=1}^{m} t_i > k_\alpha) = 1 - P_{\theta_1}(\sum_{i=1}^{m} t_i \leq k_\alpha) = 1 - F_{\theta_1}(k_\alpha).$$

Where $F_{\theta_1} = cdf(G(mr, 1/\theta_1))$, and θ_1 is a value of θ under H_1.

Finally, notice that, in particular, if we take $m = 1$, we have only one process of n individual and independent systems. And in this case we only have to substitute m for 1 in this section and the previous one.

Problem 4

Laura Arribas and Arnau Garcia

2023-11-14

The data in the flightdelays dataset, available from the resample package contains flight delays for two airlines, American and United.

(a) Compute the proportion of times that each carrier's flights was delayed more than 20 min. Conduct a two-sided permutation test to see if the difference in these proportions is statistically significant.

(b) Compute the variance in the flight delay lengths for each carrier. Conduct a permutation test to see if the variance for United Airlines differs from that of American Airlines.

(a)

Let us take a look at our dataset.

```
df <- FlightDelays
head(df)
```

```
##   ID Carrier FlightNo Destination DepartTime Day Month FlightLength Delay
## 1  1      UA      403         DEN      4-8am Fri   May          281    -1
## 2  2      UA      405         DEN     8-Noon Fri   May          277   102
## 3  3      UA      409         DEN      4-8pm Fri   May          279     4
## 4  4      UA      511         ORD     8-Noon Fri   May          158    -2
## 5  5      UA      667         ORD      4-8am Fri   May          143    -3
## 6  6      UA      669         ORD      4-8am Fri   May          150     0
##   Delayed30
## 1        No
## 2       Yes
## 3        No
## 4        No
## 5        No
## 6        No
```

First, we will compute the proportion of times that each carrier's flights were delayed more than 20 min.

```
delayAA <- subset(df, select = Delay, subset = Carrier == "AA", drop = TRUE)
delayUA <- subset(df, select = Delay, subset = Carrier == "UA", drop = TRUE)
mean(delayAA>20)
```

```
## [1] 0.1693049
```

```r
mean(delayUA>20)
```

```
## [1] 0.2128228
```

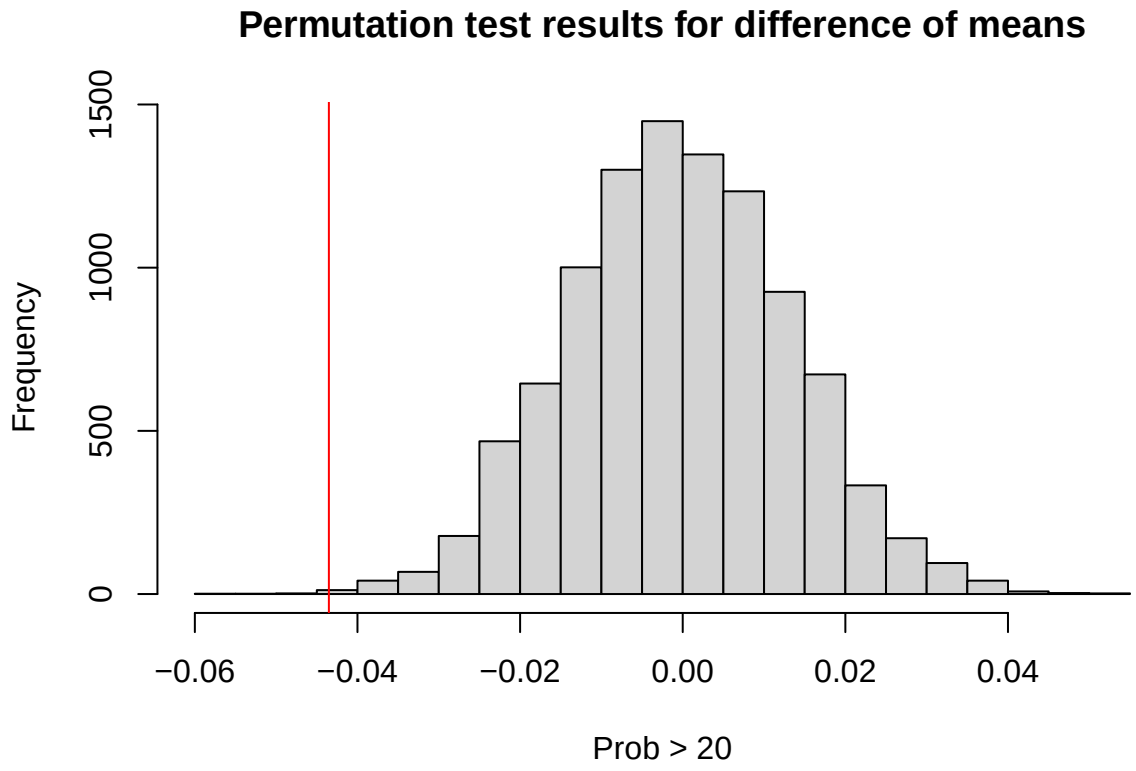
American Airlines: 16.9%. United Airlines: 21.3%.

```r
set.seed(101) ## For reproducibility

l1 <- length(delayAA)
l2 <- length(delayUA)
nsim <- 9999 ## Number of simulations
res <- numeric(nsim) ## Set aside space for the results

for(i in 1:nsim) {
  ## Compute random indexes for the permutation
  index <- sample(l1 + l2, size = l1, replace = FALSE)
  ## Compute and store difference in means
  res[i] <- mean(df$Delay[index]>20) - mean(df$Delay[-index]>20)
}
obs <- mean(delayAA > 20) - mean(delayUA > 20)

hist(res, breaks=20, main="Permutation test results for difference of means",
     xlab="Prob > 20", ylab="Frequency")
abline(v=obs, col="red")
```



Let us observe that our observed difference of means is negative. Taking this into account, we can compute the p-value as follows:

```
obs
```

```
## [1] -0.04351791
```

```
(sum(res <= obs) + 1)/(nsim + 1) + (sum(res >= -obs) + 1)/(nsim + 1)
```

```
## [1] 0.0018
```

Taking a significance level of 0.05, we can conclude that the difference of means between the two airlines is not 0. In other words, there is a significant difference between the proportion of delays.

In class, we have seen the following web with a useful example of a permutation test https://thomasleeper.com/Rcourse/Tutorials/permutationtests.html. Following this tutorial steps now we can compare our results to the ones obtained using the library **coin**.

```
library(coin)
```

```
## Loading required package: survival
```

```
independence_test(df$Delay>20 ~ df$Carrier , alternative = "two.sided")
```

```
##
##  Asymptotic General Independence Test
##
## data:  df$Delay > 20 by df$Carrier (AA, UA)
## Z = 3.2134, p-value = 0.001312
## alternative hypothesis: two.sided
```

AS you can see, our permutation test provided the same inference and a nearly identical p-value.

(b)

We can easily compute the variance in the flight delay lengths for each carrier, as follows:

```
var(delayAA)
```

```
## [1] 1606.457
```

```
var(delayUA)
```

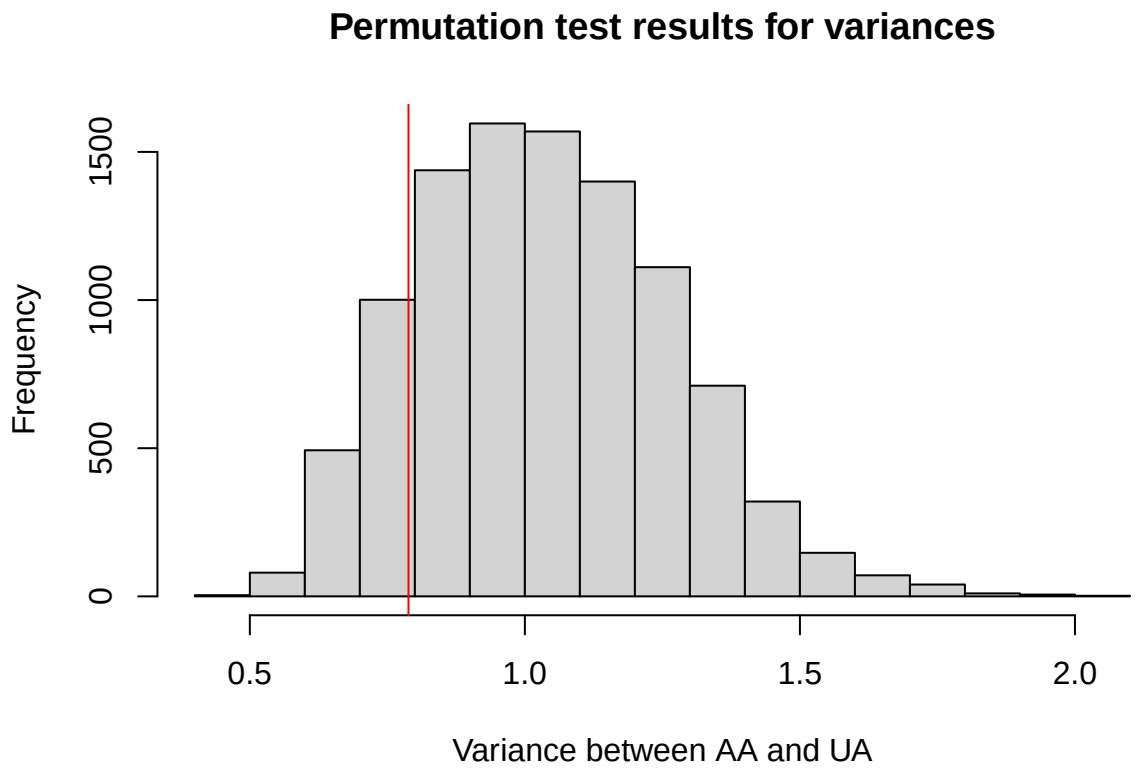
```
## [1] 2037.525
```

```r
varRes <- numeric(nsim) ## Set aside space for the results

for(i in 1:nsim) {
  ## Compute random indexes for the permutation
  index <- sample(l1 + l2, size = l1, replace = FALSE)
  ## Compute and store the division of variances
  varRes[i] <- var(df$Delay[index])/var(df$Delay[-index])
}

varObs<- var(delayAA)/var(delayUA)

hist(varRes, breaks=20, main="Permutation test results for variances",
     xlab="Variance between AA and UA", ylab="Frequency")
abline(v=varObs, col="red")
```



Since our observation is lower than 1 (center of the distribution), we can calculate the p-value of our test as follows:

```r
varObs
```

```
## [1] 0.7884356
```

```r
(sum(varRes <= varObs) + 1)/(nsim + 1) + (sum(varRes >= 2-varObs) + 1)/(nsim + 1)
```

```
## [1] 0.3699
```

Since the p-value is bigger than 0.05, we can conclude that there isn't a significance difference between the two airlines' variances.