Statistics, Fall 2025 - Practice Session 5
Asymptotics & Hypothesis Testing

TA: Stylianos Daskalinas

Email: stylianosdaskalinas@gmail.com

Office hours by appointment. Office D3, 3rd floor , Bld. B

23/10/2025

Problem 1

Consider a random sample $\{X_i\}_{i=1}^n$ where X_i follows a distribution with density

$$f(x;a) = (1+a)x^a \mathbb{1}_{[0,1]}(x)$$

Given some significance level α, find the best critical region for testing the hypotheses

$$\mathcal{H}_0 : a = 1 \ , \quad \mathcal{H}_1 : a = 2$$

using the Likelihood Ratio method.

Solution

The Likelihood function is

$$L(a|\{x_i\}_{i=1}^n) = \prod_{i=1}^n (1+a)x_i^a = (1+a)^n \Big[\prod_{i=1}^n x_i\Big]^a$$

Then, according to the Neyman-Pearson lemma, the form of the best critical region is

$$C = \Big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \Big| \frac{L(a=1|\{x_i\}_{i=1}^n)}{L(a=2|\{x_i\}_{i=1}^n)} \leq k\Big\} = \Big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \Big| \Big(\frac{2}{3}\Big)^n \Big[\prod_{i=1}^n x_i\Big]^{-1} \leq k\Big\}$$

$$= \Big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \Big| \prod_{i=1}^n x_i \geq \frac{1}{k}\Big(\frac{2}{3}\Big)^n\Big\} = \Big\{(x_1, \ldots, x_n) \in \mathbb{R}^n \Big| -\sum_{i=1}^n \log(x_i) \leq c\Big\}$$

Then, we reject the null if $-\sum_{i=1}^n \log(x_i) \leq c$, and we know from the previous practice session that $-\sum_{i=1}^n \log(x_i) \sim \chi^2_{2n}$. The best test will be given by the critical value c such that

$$\alpha = \Pr(\sum_{i=1}^{n} \log(x_i) \leq c) = \Pr(\chi^2_{2n} \leq 2)$$

Finally, we can recover the critical value for the rejection region by calculating the α-quantile of the chi^2_{2n} distribution.

Problem 2

Consider two independent random samples $\{X_i\}_{i=1}^n$ and $\{Y_j\}_{j=1}^m$, where

$$X_i \sim N(a, c) \quad , \quad Y_j \sim N(b, c)$$

where all three parameters are unknown. For a significance level α, construct the rejection region for the null hypothesis $\mathcal{H}_0 : a = b$.

Solution

First, since the two samples are independent, their joint likelihood is simply the product of the two individual samples' likelihoods:

$$L(a, b, c|\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^m) = \prod_{i=1}^{n} f(x_i; a, c) \prod_{j=1}^{m} f(y_j; b, c)$$

$$= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi c}} e^{-\frac{1}{2c}\sum_{i=1}^{n}(x_i-a)^2} \prod_{j=1}^{m} \frac{1}{\sqrt{2\pi c}} e^{-\frac{1}{2c}\sum_{j=1}^{m}(y_j-b)^2}$$

$$= (2\pi c)^{-\frac{(n+m)}{2}} \exp\Big\{ -\frac{1}{2c}\Big[\sum_{i=1}^{n}(x_i - a)^2 + \sum_{j=1}^{m}(y_j - b)^2\Big]\Big\}$$

For the Likelihood under the null, we need to substitute $a = b$ and maximize with respect to the two parameters. We can use the log-likelihood to simplify calculations:

$$\log\Big(L(a, c|\{x_i\}_{i=1}^n, \{y_j\}_{j=1}^m)\Big) = \log\Big((2\pi c)^{-\frac{(n+m)}{2}} \exp\Big\{ -\frac{1}{2c}\Big[\sum_{i=1}^{n}(x_i - a)^2 + \sum_{j=1}^{m}(y_j - a)^2\Big]\Big\}\Big)$$

$$= -\frac{n+m}{2}\log(2\pi) - \frac{n+m}{2}\log(c) - \frac{1}{2c}\Big[\sum_{i=1}^{n}(x_i - a)^2 + \sum_{j=1}^{m}(y_j - a)^2\Big]$$

The first order condition

$$\frac{\partial}{\partial a}\log\Big(L(a,c|\{x_i\}_{i=1}^n,\{y_j\}_{j=1}^m)\Big)=0\Rightarrow\frac{1}{c}\Big[\sum_{i=1}^n(x_i-a)+\sum_{j=1}^m(y_j-a)\Big]=0$$

$$\Rightarrow n\overline{x}+m\overline{y}-(n+m)a=0\Rightarrow\hat{a}_0=\frac{n}{n+m}\overline{x}+\frac{m}{n+m}\overline{y}$$

$$\frac{\partial}{\partial c}\log\Big(L(a,c|\{x_i\}_{i=1}^n,\{y_j\}_{j=1}^m)\Big)=0\Rightarrow-\frac{n+m}{2c}+\frac{1}{2c^2}\Big[\sum_{i=1}^n(x_i-a)^2+\sum_{j=1}^m(y_j-a)^2\Big]=0$$

$$\Rightarrow\hat{c}_0=\frac{1}{n+m}\Big[\sum_{i=1}^n(x_i-\hat{a}_0)^2+\sum_{j=1}^m(y_j-\hat{a}_0)^2\Big]$$

Substituting the maximizers in the Likelihood function, we get

$$L_0=(2\pi e\hat{c}_0)^{-\frac{n+m}{2}}$$

For the Likelihood under the alternative, we need to simply maximize the initial Likelihood function. Again, we can use the log-likelihood for ease of calculations

$$\log\Big(L(a,b,c|\{x_i\}_{i=1}^n,\{y_j\}_{j=1}^m)\Big)=\log\Big((2\pi c)^{-\frac{(n+m)}{2}}\exp\Big\{-\frac{1}{2c}\Big[\sum_{i=1}^n(x_i-a)^2+\sum_{j=1}^m(y_j-b)^2\Big]\Big\}\Big)$$

$$=-\frac{n+m}{2}\log(2\pi)-\frac{n+m}{2}\log(c)-\frac{1}{2c}\Big[\sum_{i=1}^n(x_i-a)^2+\sum_{j=1}^m(y_j-b)^2\Big]$$

The first order conditions are:

$$\frac{\partial}{\partial a}\log\Big(L(a,b,c|\{x_i\}_{i=1}^n,\{y_j\}_{j=1}^m)\Big)=0\Rightarrow\frac{1}{c}\sum_{i=1}^n(x_i-a)=0\Rightarrow\hat{a}_1=\overline{x}$$

$$\frac{\partial}{\partial b}\log\Big(L(a,b,c|\{x_i\}_{i=1}^n,\{y_j\}_{j=1}^m)\Big)=0\Rightarrow\frac{1}{c}\sum_{j=1}^m(y_j-b)=0\Rightarrow\hat{b}_1=\overline{y}$$

$$\frac{\partial}{\partial c}\log\Big(L(a,b,c|\{x_i\}_{i=1}^n,\{y_j\}_{j=1}^m)\Big)=0\Rightarrow-\frac{n+m}{2c}+\frac{1}{2c^2}\Big[\sum_{i=1}^n(x_i-a)^2+\sum_{j=1}^m(y_j-b)^2\Big]=0$$

$$\Rightarrow\hat{c}_1=\frac{1}{n+m}\Big[\sum_{i=1}^n(x_i-\hat{a}_1)^2+\sum_{j=1}^m(y_j-\hat{b}_1)^2\Big]$$

The maximized Likelihood under the alternative is

$$L_1=(2\pi e\hat{c}_1)^{-\frac{n+m}{2}}$$

Then, we can construct the Likelihood Ratio statistic as

$$\Lambda = \frac{L_0}{L_1} = \Big(\frac{\hat{c}_1}{\hat{c}_0}\Big)^{\frac{n+m}{2}}$$

and, for some $k \in [0,1]$ the critical region as

$$\begin{aligned} C &= \Big\{(x_1, \ldots, x_n, y_1, \ldots, y_m) \in \mathbb{R}^{n+m} \Big| \Big(\frac{\hat{c}_1}{\hat{c}_0}\Big)^{\frac{n+m}{2}} \leq k\Big\} \\ &= \Big\{(x_1, \ldots, x_n, y_1, \ldots, y_m) \in \mathbb{R}^{n+m} \Big| \frac{\hat{c}_1}{\hat{c}_0} \leq k^{\frac{2}{n+m}}\Big\} \end{aligned}$$

Now, notice that, denoting by S_x^2 the sample variance of $\{x_i\}_{i=1}^n$ and by S_y^2 the sample variance of $\{y_j\}_{j=1}^m$, we get

$$\begin{aligned} \hat{c}_1 &= \frac{1}{n+m}\Big[\sum_{i=1}^n (x_i - \hat{a}_1)^2 + \sum_{j=1}^m (y_j - \hat{b}_1)^2\Big] \\ &= \frac{1}{n+m}\Big[\sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{j=1}^m (y_j - \overline{y})^2\Big] = \frac{n}{n+m}S_x^2 + \frac{m}{n+m}S_y^2 \end{aligned}$$

Also, see that

$$\begin{aligned} &\sum_{i=1}^n (x_i - \hat{a}_0)^2 = \sum_{i=1}^n (x_i - \frac{n}{n+m}\overline{x} - \frac{m}{n+m}\overline{y})^2 = \sum_{i=1}^n (x_i - \overline{x} + \overline{x} - \frac{n}{n+m}\overline{x} - \frac{m}{n+m}\overline{y})^2 \\ &= \sum_{i=1}^n (x_i - \overline{x} + \frac{m}{n+m}(\overline{x} - \overline{y}))^2 = \sum_{i=1}^n (x_i - \overline{x})^2 + 2\sum_{i=1}^n (x_i - \overline{x})\frac{n}{n+m}(\overline{x} - \overline{y}) + n\Big(\frac{m}{n+m}\Big)^2(\overline{x} - \overline{y})^2 \\ &= nS_x^2 + n\Big(\frac{m}{n+m}\Big)^2(\overline{x} - \overline{y})^2 \end{aligned}$$

and

$$\begin{aligned} &\sum_{j=1}^m (y_j - \hat{a}_0)^2 = \sum_{j=1}^m (y_j - \frac{n}{n+m}\overline{x} - \frac{m}{n+m}\overline{y})^2 = \sum_{j=1}^m (y_j - \overline{y} + \overline{y} - \frac{n}{n+m}\overline{x} - \frac{m}{n+m}\overline{y})^2 \\ &= \sum_{j=1}^m (y_j - \overline{y} + \frac{n}{n+m}(\overline{y} - \overline{x}))^2 = \sum_{j=1}^m (y_j - \overline{y})^2 + 2\sum_{j=1}^m (y_j - \overline{y})\frac{n}{n+m}(\overline{y} - \overline{x}) + m\Big(\frac{n}{n+m}\Big)^2(\overline{y} - \overline{x})^2 \\ &= mS_y^2 + m\Big(\frac{n}{n+m}\Big)^2(\overline{y} - \overline{x})^2 \end{aligned}$$

Thus, for $\hat{c}_0$ we have

$$
\begin{aligned}
\hat{c}_0 &= \frac{1}{n+m}\Big[\sum_{i=1}^{n}(x_i-\hat{a}_0)^2+\sum_{j=1}^{m}(y_j-\hat{a}_0)^2\Big] \\
&= \frac{1}{n+m}\Big[mS_x^2+n\Big(\frac{m}{n+m}\Big)^2(\overline{x}-\overline{y})^2+mS_y^2+m\Big(\frac{n}{n+m}\Big)^2(\overline{y}-\overline{x})^2\Big] \\
&= \frac{n}{n+m}S_x^2+\frac{m}{n+m}S_y^2+\frac{nm}{(n+m)^2}(\overline{x}-\overline{y})^2
\end{aligned}
$$

Then, the final shape of the statistic is

$$
\begin{aligned}
\Lambda^{\frac{2}{n+m}} &= \frac{\frac{n}{n+m}S_x^2+\frac{m}{n+m}S_y^2}{\frac{n}{n+m}S_x^2+\frac{m}{n+m}S_y^2+\frac{nm}{(n+m)^2}(\overline{x}-\overline{y})^2} \\
&= \frac{nS_x^2+mS_y^2}{nS_x^2+mS_y^2+\frac{nm}{n+m}(\overline{x}-\overline{y})^2} \\
&= \left(1+\frac{\frac{nm}{n+m}(\overline{x}-\overline{y})^2}{nS_x^2+mS_y^2}\right)^{-1}
\end{aligned}
$$

It can be shown that, under the null hypothesis, the statistic

$$
T=\frac{\sqrt{\frac{nm}{n+m}}(\overline{x}-\overline{y})}{\sqrt{\frac{nS_x^2+mS_y^2}{n+m-2}}}
$$

follows the t distribution with $n+m-2$ degrees of freedom, therefore, the critical region becomes

$$
\Lambda\leq k \Rightarrow \frac{n+m-2}{n+m-2+T^2}\leq k^{\frac{2}{n+m}} \Rightarrow |T|\geq (n+m-2)(k^{-\frac{2}{n+m}}-1)=c
$$

Then, for a significance level a, according to the Neyman-Pearson lemma we can recover the critical values of the Uniformly-Most-Powerful test by finding a number c such that

$$
\alpha=\Pr(\Lambda\leq k)=\Pr(|T|\geq c)
$$

Since we know that $T\sim t_{n+m-2}$, c will simply be the $1-\frac{\alpha}{2}$ quantile of the t distribution, and the decision rule will be

$$
\begin{aligned}
&\text{Reject } \mathcal{H}_0 \text{ if } |\hat{T}|>t_{\frac{\alpha}{c},n+m-2} \\
&\text{Accept } \mathcal{H}_0 \text{ otherwise}
\end{aligned}
$$

Notes

- If $X \sim N$ and $Y \sim \chi_k^2$, and the two random variables are independent, then

$$T = \frac{X}{\sqrt{Y/k}} \sim t_k$$