Ozonmasters

Statistics

Home work

Kozhemyak Vitaly

November 13, 2020

Contents

Task 1

T1 (i)

Statement 1. *Consider* $X_1, \dots, X_n$ *— i.i.d. random variables. Let* $F(x)$ *be a CDF of random variable* $X_i, i = \{1, \dots, n\}$. *Then*

$$-\sum_{i=1}^{n} \log(1 - F(X_i)) \sim \Gamma(n, 1).$$

Proof. The CDF of random variable $1 - F(X_i)$ is

$$\mathbb{P}(1-F(X_i) \leqslant x) = \mathbb{P}(1-x \leqslant F(X_i)) = 1-\mathbb{P}(1-x > F(X_i)) = 1-F(F^{-1}(1-x)) = x.$$

Thus,

$$1 - F(X_i) \sim U[0,1] \Rightarrow -\log(1 - F(X_i)) \sim Exp(1).$$

Now we are going to use a moment-generating function $M_X(t)$ which uniquely determines a distribution. If we have $Y_1, \dots, Y_n \sim Exp(1)$ and $S_n = Y_1 + \dots + Y_n$ then

$$M_{S_n}(t) = \mathbb{E}[e^{tS_n}] = \mathbb{E}[e^{t(Y_1+\dots+Y_n)}] = \{i.i.d.\} = \mathbb{E}[e^{tY_1}] \cdot \ldots \cdot \mathbb{E}[e^{tY_n}] =$$

$$= M_{Y_1}(t) \cdot \ldots \cdot M_{Y_n}(t) = \frac{1}{(1-t)^n}, \; t < 1.$$

This is a moment-generating function of Gamma distribution $\Gamma(n, 1)$. If we put $Y_i = -\log(1 - F(X_i))$ we will get the statement.

□

<u>*Plotting exact confidence intervals*</u>

Consider $X_1, \dots, X_n$ — i.i.d. random variables with Weibull CDF

$$F(x) = 1 - e^{-(x/\lambda)^\tau}, \; x > 0.$$

Choose the next central statistic $Z = Z(X_1, \dots, X_n; \lambda) = \sum\limits_{i=1}^{n} \left(\frac{X_i}{\lambda}\right)^\tau$. A central statistics satisfies the next rules:

1. $F_Z(z)$ does not depend on λ. Denote $Z_i = \left(\frac{X_i}{\lambda}\right)^\tau$. Then

$$F_{Z_i}(z) = \mathbb{P}(Z_i \leqslant z) = \mathbb{P}(X_i \leqslant \lambda x^{1/\tau}) = 1 - e^{-x} \Rightarrow Z_i \sim Exp(1).$$

   Moreover, the result above tells us $Z = \sum_{i=1}^{n} Z_i \sim \Gamma(n, 1)$.

2. Z is continuous and strictly monotone in λ.

According to the definition of the exact confidence interval

$$\mathbb{P}(q_{\alpha/2} \leqslant Z \leqslant q_{1-\alpha/2}) = 1 - \alpha,$$

$$\mathbb{P}\left(\frac{q_{\alpha/2}}{\sum_{i=1}^{n} X_i^\tau} \leqslant \frac{1}{\lambda^\tau} \leqslant \frac{q_{1-\alpha/2}}{\sum_{i=1}^{n} X_i^\tau}\right) = 1 - \alpha,$$

$$\boxed{\mathbb{P}\left(\left(\frac{q_{\alpha/2}}{\sum_{i=1}^{n} X_i^\tau}\right)^{1/\tau} \geqslant \lambda \geqslant \left(\frac{q_{1-\alpha/2}}{\sum_{i=1}^{n} X_i^\tau}\right)^{1/\tau}\right) = 1 - \alpha}$$

where $q_{\alpha/2}$ and $q_{1-\alpha/2}$ are left and right $\frac{\alpha}{2}$ – quantiles of Gamma distribution $\Gamma(n, 1)$ with the significance level α respectively.

T1 (ii)

Consider $X_1, \ldots, X_n \sim Exp(\lambda)$. As we know $\mathbb{E}[X_1] = \frac{1}{\lambda}$, $\mathrm{Var}[X_1] = \frac{1}{\lambda^2}$. According to the CLT we have

$$\lambda\sqrt{n}\left(\bar{X} - \frac{1}{\lambda}\right) \to \mathcal{N}(0, 1), \ n \to +\infty.$$

Consider the next two statistics

$$S_1(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i,$$

$$S_2(X_1, \ldots, X_n) = \sqrt{\frac{1}{2n}\sum_{i=1}^{n} X_i^2}.$$

Plotting asymptotic confidence intervals using S_1 statistics

Using the S_1 statistics we get

$$\mathbb{P}\left(-z_{1-\alpha/2} \leqslant \lambda\sqrt{n}\left(\bar{X}-\frac{1}{\lambda}\right) \leqslant z_{1-\alpha/2}\right) = 1-\alpha,$$

$$\boxed{\mathbb{P}\left(\frac{1}{\bar{X}}\left(1-\frac{z_{1-\alpha/2}}{\sqrt{n}}\right) \leqslant \lambda \leqslant \frac{1}{\bar{X}}\left(1+\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)\right) = 1-\alpha.}$$

The length of the confidence interval is

$$\boxed{l_1 = \frac{2}{\bar{X}}\frac{z_{1-\alpha/2}}{\sqrt{n}}.}$$

Plotting asymptotic confidence intervals using S_2 statistics

Using the next formula $\left(\sigma^2 = \frac{1}{n}\sum\limits_{i=1}^{n}(X_i-\mu)^2\right)$ we can derive the next equation

$$\frac{1}{\lambda^2} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\frac{1}{\lambda}\right)^2 \Rightarrow \bar{X} = \frac{\lambda}{2}\frac{1}{n}\sum_{i=1}^{n}X_i^2.$$

Then

$$\mathbb{P}\left(-z_{1-\alpha/2} \leqslant \lambda\sqrt{n}\left(\bar{X}-\frac{1}{\lambda}\right) \leqslant z_{1-\alpha/2}\right) = 1-\alpha,$$

$$\boxed{\mathbb{P}\left(\sqrt{\frac{2n}{\sum\limits_{i=1}^{n}X_i^2}\left(1-\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)} \leqslant \lambda \leqslant \sqrt{\frac{2n}{\sum\limits_{i=1}^{n}X_i^2}\left(1+\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)}\right) = 1-\alpha.}$$

The length of the confidence interval is

$$\boxed{l_2 = \sqrt{\frac{2n}{\sum\limits_{i=1}^{n}X_i^2}}\cdot\left(\sqrt{\left(1+\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)}-\sqrt{\left(1-\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)}\right).}$$

Let's take a look at the plot below

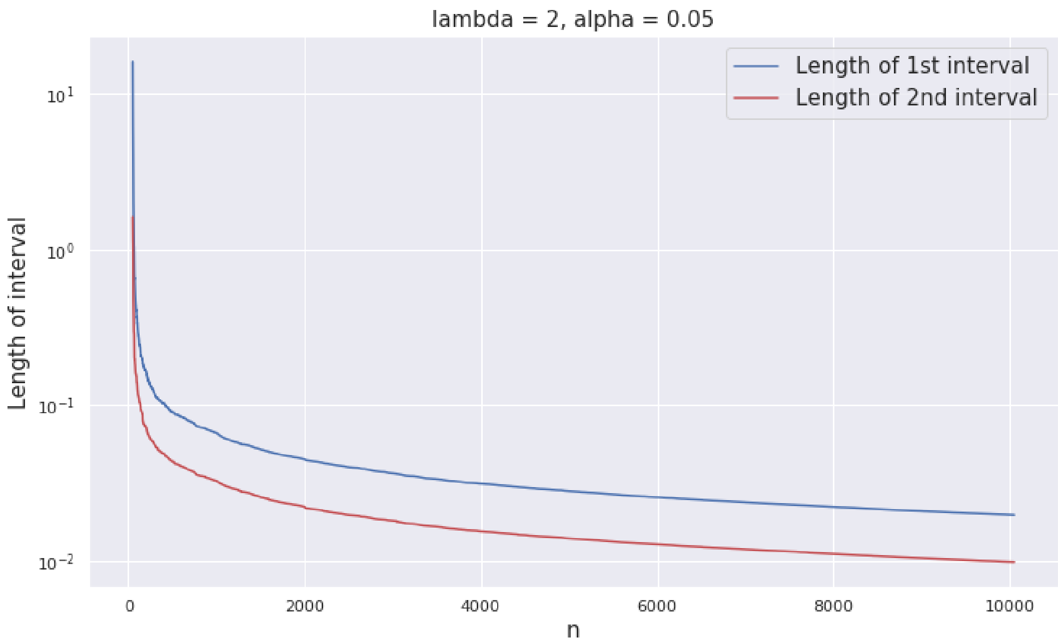

Figure 1: The length of the second interval is less than the length of the first interval ($l_2 < l_1$).

T1 (iii)

Consider $X_1, \dots, X_n$ — i.i.d. with a CDF $F(x) = 1 - e^{-x/\lambda}$. The order statistics $X_{(i)}, X_{(j)}, i < j$ satisfy the next equation

$$\mathbb{P}\left(X_{(i)} < m = F^{-1}\left(\frac{1}{2}\right) < X_{(j)}\right) = \mathbb{P}(m < X_{(j)}) - \mathbb{P}(m \leqslant X_{(i)}).$$

Using a CDF of Binomial distribution

$$F_{X_{(r)}}(m) = \mathbb{P}(X_{(r)} \leqslant m) = \sum_{k=r}^{n} C_n^k (1 - F(m))^k (F(m))^{n-k} =$$

$$= \sum_{k=r}^{n} C_n^k \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} = 2^{-n} \sum_{k=r}^{n} C_n^k.$$

Thus,

$$\mathbb{P}\left(X_{(i)} < m < X_{(j)}\right) = 2^{-n} \sum_{k=i}^{j-1} C_n^k.$$

Since

$$F(m) = 1 - e^{-m/\lambda} = \frac{1}{2}$$

then

$$m = \lambda \ln 2.$$

Finally,

$$\boxed{\begin{gathered}\mathbb{P}(X_{(i)} < m < X_{(j)}) = \mathbb{P}\left(\frac{X_{(i)}}{\ln 2} < \lambda < \frac{X_{(j)}}{\ln 2}\right) = \\ = 2^{-n}\sum_{k=i}^{j-1} C_n^k = 1 - \alpha,\ \ i < j.\end{gathered}}$$

N1

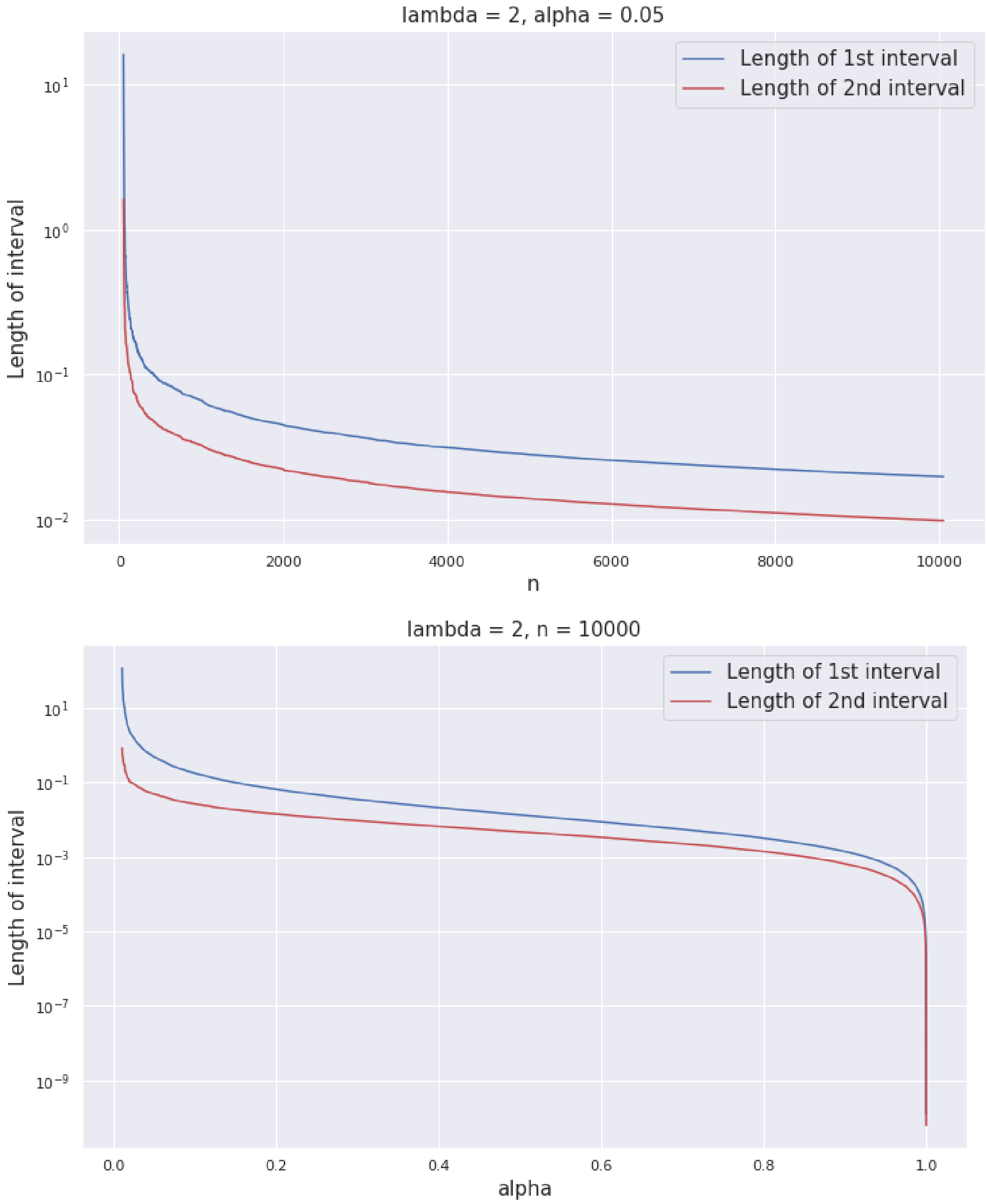

Figure 2: The length of the second interval is less than the length of the first interval ($l_2 < l_1$).

Task 2

T2

Let X_i, Y_i be random variables such that

$$X_i = \begin{cases} 1, & p, \\ 0, & 1-p, \end{cases} \quad Y_i = \begin{cases} 1, & q, \\ 0, & 1-q, \end{cases}$$

where p is probability to answer "YES" in a city, q is probability to answer "YES" in a village. Using Central Limit Theorem we get

$$\frac{\overline{X} - \overline{Y} - (p-q)}{\sqrt{\dfrac{p(1-p)}{n} + \dfrac{q(1-q)}{n}}} = \{\text{Law of Large Numbers}\} =$$

$$= \frac{\overline{X} - \overline{Y} - (p-q)}{\sqrt{\dfrac{\overline{X}(1-\overline{X})}{n} + \dfrac{\overline{Y}(1-\overline{Y})}{n}}} \to \mathcal{N}(0,1), \; n \to +\infty.$$

Denote

$$S_n = \sqrt{\frac{\overline{X}(1-\overline{X})}{n} + \frac{\overline{Y}(1-\overline{Y})}{n}}.$$

The confidence intervals can be calculated as following

$$\mathbb{P}\left(-z_{1-\alpha/2} \leqslant \frac{\overline{X} - \overline{Y} - (p-q)}{S_n} \leqslant z_{1-\alpha/2}\right) = 1 - \alpha,$$

$$\mathbb{P}\left(\overline{X} - \overline{Y} - z_{1-\alpha/2} S_n \leqslant p - q \leqslant \overline{X} - \overline{Y} + z_{1-\alpha/2} S_n\right) = 1 - \alpha.$$

We can notice that

$$S_n = \sqrt{\frac{\overline{X}(1-\overline{X})}{n} + \frac{\overline{Y}(1-\overline{Y})}{n}} \leqslant$$

$$\leqslant \left\{\overline{X}(1-\overline{X}) \leqslant \frac{1}{4}\right\} \leqslant \frac{1}{\sqrt{2n}}.$$

Then

$$\mathbb{P}\left(\overline{X} - \overline{Y} - z_{1-\alpha/2}\frac{1}{\sqrt{2n}} \leqslant p - q \leqslant \overline{X} - \overline{Y} + z_{1-\alpha/2}\frac{1}{\sqrt{2n}}\right) = 1 - \alpha.$$

Finally, n_{min} is a solution of the next equation

$$\frac{z_{1-\alpha/2}}{\sqrt{2n}} = 0.05,$$

$$\boxed{n_{min} = \frac{1}{2}\left(\frac{z_{1-\alpha/2}}{0.05}\right)^2, \ \alpha = 0.05.}$$

T3

Consider a null hypothesis $H_0 : p - q = 0$ and an alternative $H_1 : p - q = 0.03$. Let's calculate the Test Statistics

$$T = \frac{\overline{X} - \overline{Y} - (p-q)}{\sqrt{\frac{p(1-p)}{n} + \frac{q(1-q)}{n}}} = \frac{\overline{X} - \overline{Y} - (p-q)}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n} + \frac{\overline{Y}(1-\overline{Y})}{n}}} = \sqrt{n}\frac{0.05 - (p-q)}{\sqrt{0.2023}}.$$

Reject the null hypothesis if

$$T > z_{1-\alpha}$$

where $z_{1-\alpha}$ is a $1-\alpha$-quantile of $\mathcal{N}(0,1)$.
Let's calculate type 1 error as following

$$\mathbb{P}(\text{error 1 type}) = \mathbb{P}(\text{reject } H_0 \,|\, H_0 \text{ is true}) = \mathbb{P}(T > z_{1-\alpha} | p = q).$$

Let's calculate type 2 error as following

$$\mathbb{P}(\text{error 2 type}) = \mathbb{P}(\text{don't reject } H_0 \,|\, H_0 \text{ is false}) = \mathbb{P}(T \leqslant z_{1-\alpha} | p = q + 0.03).$$

Since we want that

$$\mathbb{P}(\text{error 1 type}) = 0.05,$$
$$\mathbb{P}(\text{error 2 type}) = 0.04,$$

we have to solve the next system of inequalities

$$\begin{cases} \sqrt{n}\dfrac{0.05}{\sqrt{0.2023}} > z_{0.95} = 1.65, \\ \sqrt{n}\dfrac{0.02}{\sqrt{0.2023}} \leqslant z_{0.96} = 1.76, \end{cases} \Leftrightarrow \begin{cases} n > 250.65, \\ n < 1566.61. \end{cases}$$

Finally, we get

$$\boxed{n_{min} = 251.}$$

Task 3

T4

Consider a null hypothesis $H_0 : \theta = \theta_0$, an alternative $H_1 : \theta > \theta_0$ and

LR (Likelihood Ratio) test

$$\lambda(\vec{x}) = \frac{\mathcal{L}(\vec{x}, \theta_0)}{\sup\limits_{\theta \geqslant \theta_0} \mathcal{L}(\vec{x}, \theta)}.$$

Let $\hat{\theta}$ be the solution for the next problems $\sup\limits_{\theta \geqslant \theta_0} \mathcal{L}(\vec{x}, \theta)$. Now the LR can be written as following

$$\lambda(\vec{x}) = \left(\frac{\theta_0}{\hat{\theta}}\right)^n \prod_{i=1}^{n} (1 - x_i)^{\theta_0 - \hat{\theta}}.$$

Thus, we accept H_0 if

$$\lambda(\vec{x}) \geqslant c \Leftrightarrow n \log\left(\frac{\theta_0}{\hat{\theta}}\right) + (\theta_0 - \hat{\theta}) \sum_{i=1}^{n} \log(1 - x_i) \geqslant \log c.$$

Denote

$$\widetilde{c} = \frac{\log c - n \log\left(\frac{\theta_0}{\hat{\theta}}\right)}{\theta_0 - \hat{\theta}}.$$

Searching for the threshold c

To find the constant $\widetilde{c}$ we are going to calculate a type 1 error

$$\mathbb{P}(\text{type 1 error}) = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is true}) = \mathbb{P}\left(\sum_{i=1}^{n} \log(1 - x_i) < \widetilde{c} \,|\, H_0 \text{ is true}\right).$$

Statement 2. *Consider $X_1, \ldots, X_n \sim p(x; \theta)$. Then*

$$\sum_{i=1}^{n} \log(1 - X_i) \sim \Gamma\left(n, \frac{1}{\theta}\right).$$

Proof. Consider exponential random variable $X \sim p(x;\theta)$. Then

$$F_{\log(1-X)}(x) = \mathbb{P}(\log(1-X) \leqslant x) = 1 - F_X(1-e^x).$$

Recall that

$$F_X(x) = \begin{cases} 0, & x < 0, \\ 1-(1-x)^\theta, & 0 \leqslant x \leqslant 1, \\ 1, & x > 1. \end{cases}$$

Thus, we have

$$F_{\log(1-X)}(x) = e^{x\theta} \Leftrightarrow f_{\log(1-X)}(x) = \theta e^{x\theta} \sim Exp(\theta), \ x \in (-\infty, 0].$$

As we know

$$\sum_{i=1}^{n} \log(1-X_i) \sim \Gamma\left(n, \frac{1}{\theta}\right).$$

$\square$

Back to the type 1 error

$$\mathbb{P}(\text{type 1 error}) = \mathbb{P}\left(\sum_{i=1}^{n} \log(1-x_i) < \widetilde{c} \,|\, H_0 \text{ is true}\right) = F_{(n,\theta_0)}(\widetilde{c}) = \alpha,$$

$$\widetilde{c} = F^{-1}_{(n,\theta_0)}(\alpha) = q_{(n,\theta_0)}(\alpha)$$

Threshold c can be found the next way

$$\boxed{c = \exp\left((\theta_0 - \hat{\theta}) \cdot q_{(n,\theta_0)}(\alpha) + n\log\left(\frac{\theta_0}{\hat{\theta}}\right)\right)}$$

where $F_{(n,\theta_0)}$ is a CDF of Gamma distribution $\Gamma\left(n, \frac{1}{\theta_0}\right)$, and $q_{(n,\theta_0)}(\alpha)$ is a α – quantile of Gamma distribution $\Gamma\left(n, \frac{1}{\theta_0}\right)$ (if we consider the left tailed test).

The power of test

According to the definition of the power of statistic test

$$W(\theta) = \mathbb{P}(\text{reject } H_0 \,|\, H_1 \text{ is true}) =$$

$$= \mathbb{P}\left(\sum_{i=1}^{n} \log(1-x_i) < \widetilde{c} \,|\, H_1 \text{ is true}\right) = F_{(n,\theta)}(q_{(n,\theta_0)}(\alpha)),$$

where $F_{(n,\theta)}$ is a CDF of Gamma distribution $\Gamma\left(n, \frac{1}{\theta}\right)$ and $\theta > \theta_0$. Finally, we have

$$W(\theta) = F_{(n,\theta)}(q_{(n,\theta_0)}(\alpha)) = \int\limits_0^{q_{(n,\theta_0)}(\alpha)} \frac{\theta^n t^{n-1} e^{-t\theta}}{\Gamma(n)} dt =$$

$$= \left\{ z = \frac{\theta}{\theta_0} t \right\} = \frac{\theta_0}{\theta} \int\limits_0^{\frac{\theta}{\theta_0} \cdot q_{(n,\theta_0)}(\alpha)} \frac{\theta^n \theta_0^{n-1}}{\theta^{n-1}} \frac{z^{n-1} e^{-t\theta_0}}{\Gamma(n)} dz = F_{(n,\theta_0)}\left(\frac{\theta}{\theta_0} \cdot q_{(n,\theta_0)}\right).$$

$$\boxed{W(\theta) = F_{(n,\theta_0)}\left(\frac{\theta}{\theta_0} \cdot q_{(n,\theta_0)}\right).}$$

T5

Note that

$$\boxed{\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right).}$$

Let

$$E_1(c) = \mathbb{P}(\bar{X} > c | \mu = \mu_0) = 1 - \mathbb{P}(\bar{X} \leqslant c | \mu = \mu_0) = 1 - \int\limits_{-\infty}^{c} \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} e^{-\frac{n(x-\mu_0)^2}{2\sigma^2}} dx$$

be a type 1 error and

$$E_2(c) = \mathbb{P}(\bar{X} \leqslant c | \mu = \mu_1) = \int\limits_{-\infty}^{c} \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} e^{-\frac{n(x-\mu_1)^2}{2\sigma^2}} dx$$

be a type 2 error.
Using the optimality condition we get

$$\frac{d(E_1(c) + E_2(c))}{dc} = 0,$$

$$\frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} e^{-\frac{n(c-\mu_1)^2}{2\sigma^2}} = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}} e^{-\frac{n(c-\mu_0)^2}{2\sigma^2}},$$

$$\boxed{c = \frac{\mu_1 + \mu_0}{2}.}$$

Task 4

T6* (i)

Let's denote

$$\boxed{\hat{\theta} = \mathop{\mathbf{argmax}}_{\theta \geqslant \theta_0} \mathcal{L}(\vec{x}, \theta) = \mathop{\mathbf{argmax}}_{\theta \geqslant \theta_0} \prod_{i=1}^{n} p(x_i, \theta).}$$

Then

$$\Lambda(\vec{x}) = \frac{\mathcal{L}(\vec{x}, \hat{\theta})}{\mathcal{L}(\vec{x}, \theta_0)} \geqslant 0 \tag{1}$$

be a likelihood ratio. Consider the log-likelihood ratio, normalized dividing by n :

$$\Lambda_n(\vec{x}) = \frac{1}{n} \log \frac{\mathcal{L}(\vec{x}, \hat{\theta})}{\mathcal{L}(\vec{x}, \theta_0)} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i, \hat{\theta})}{p(x_i, \theta_0)}.$$

Note $X_1, \ldots, X_n$ are i.i.d. and $L_i = \log \dfrac{p(x_i, \hat{\theta})}{p(x_i, \theta_0)}$ is a random variable. In addition, we know from the strong Law of Large Numbers that

$$\frac{1}{n} \sum_{i=1}^{n} \log \frac{p(x_i, \hat{\theta})}{p(x_i, \theta_0)} \to \mathbb{E}[L_1], \ n \to +\infty.$$

By definition of expected value

$$\mathbb{E}[L_1] = \int_{-\infty}^{+\infty} q(x) \log \frac{p(x, \hat{\theta})}{p(x, \theta_0)} dx \geqslant 0$$

where we can put $q(x) = p(x, \theta_0)$ and then

$$\mathbb{E}[L_1] = \int\limits_{-\infty}^{+\infty} p(x, \theta_0) \log \frac{p(x, \hat{\theta})}{p(x, \theta_0)} dx = K(\theta_0, \hat{\theta}).$$

Finally,

$$\log \Lambda(\vec{x}) = n\,\mathbb{E}[L_1] = \begin{cases} nK(\theta_0, \hat{\theta}) & \hat{\theta} > \theta_0 \\ 0, & \hat{\theta} = \theta_0. \end{cases}$$

T6* (ii)

Using (1) we can rewrite LR test as following

$$\mathcal{L}(\vec{x}, \hat{\theta}) \geqslant c_\alpha \mathcal{L}(\vec{x}, \theta_0),$$

$$\prod_{i=1}^{n} e^{x_i \hat{\theta} - d(\hat{\theta})} \geqslant c_\alpha \prod_{i=1}^{n} e^{x_i \theta_0 - d(\theta_0)},$$

$$\sum_{i=1}^{n} (x_i \hat{\theta} - d(\hat{\theta})) \geqslant \log c_\alpha + \sum_{i=1}^{n} (x_i \theta_0 - d(\theta_0)),$$

$$\hat{\theta} \geqslant \theta_0 + \frac{\log c_\alpha + nd(\hat{\theta}) - nd(\theta_0)}{\sum\limits_{i=1}^{n} x_i}.$$

If we denote $\widetilde{c}_\alpha = \dfrac{\log c_\alpha + nd(\hat{\theta}) - nd(\theta_0)}{\sum\limits_{i=1}^{n} x_i}$ then

$$\hat{\theta} \geqslant \theta_0 + \widetilde{c}_\alpha.$$

Appendix

- Source code can be found here.