Statistiek (WISB261)

Final exam: sketch of solutions

June 28, 2023

Schrijf uw naam op elk in te leveren vel. Schrijf ook uw studentnummer op blad 1.
`(The exam is a CLOSED-book exam: students can bring only two A4-sheets with personal notes. The use of the statistical tables is allowed. The scientific calculator is also allowed).`

The maximum number of points is 100.
Points distribution: 20–24–26–30

1. (a) [8pt] Let Y_1 and Y_2 be two i.i.d. random variables such that $Y_i \sim \text{Poi}(\lambda)$ for $i \in \{1,2\}$ and $\lambda \in \mathbb{R}_+$, i.e.,

$$\mathbb{P}(Y_i = k) = e^{-\lambda}\frac{\lambda^k}{k!} \text{ with } k \in \mathbb{N}_0.$$

Show that $Y_1 + Y_2 \sim \text{Poi}(2\lambda)$.
Solution:
Since $M_{Y_i}(t) = e^{\lambda(e^t-1)}$ for $i \in \{1,2\}$, and being $Y_1 \perp Y_2$, we have:

$$M_{Y_1+Y_2}(t) = M_{Y_1}(t)M_{Y_2}(t) = e^{2\lambda(e^t-1)}$$

which is the MGF of a $\text{Poi}(2\lambda)$ distributed random variable.

(b) [8pt] For the random variable $Y \sim \text{Poi}(100)$ find an approximated value of the probability $\mathbb{P}(Y > 120)$.
Solution:
From point (a) the random variable $Y \sim \text{Poi}(100)$ can be written as:

$$Y \overset{d}{=} \sum_{i=1}^{100} Y_i,$$

with $Y_i \overset{i.i.d.}{\sim} \text{Poi}(1)$. Hence, by the classical CLT:

$$\mathbb{P}(Y > 120) = 1 - \mathbb{P}(Y \leq 120) = 1 - \mathbb{P}\left(\frac{Y-100}{10} \leq 2\right) \overset{CLT}{\approx} 1 - \Phi(2) \approx 0.028.$$

(c) [4pt] Show that:

$$\lim_{n\to\infty} e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!} = \frac{1}{2}$$

Solution:
We pose $I_n := e^{-n}\sum_{k=0}^{n}\frac{n^k}{k!}$ and we notice that $I_n = \mathbb{P}(Y_n \leq n)$, with $Y_n \sim \text{Poi}(n)$. Similarly to point (b), by the CLT we know that:

$$\frac{Y_n - n}{\sqrt{n}} \overset{d}{\to} Z \sim N(0,1).$$

Therefore

$$\lim_{n\to\infty} I_n = \lim_{n\to\infty}\mathbb{P}(Y_n \leq n) = \lim_{n\to\infty}\mathbb{P}\left(\frac{Y_n - n}{\sqrt{n}} \leq 0\right) = \Phi(0) = \frac{1}{2}.$$

2. (a) [8pt] Assume that the outcome of an experiment is the realization of a single random variable Y, whose distribution depends on an unknown parameter $\theta \in \mathbb{R}$. A 80% confidence interval for θ has the form $[Y-1, Y+2]$. Determine a decision rule and the rejection region for testing:

$$\begin{cases} H_0: & \theta = 5, \\ H_1: & \theta \neq 5. \end{cases}$$

at $\alpha = 0.2$ level of significance.

Solution:

By the CI-hypothesis test duality, we do not reject H_0 iff $5 \in [Y-1, Y+2]$, so that we reject H_0 if $5 \leq Y-1$ or $5 \geq Y+2$. Hence the required rejection region $\mathcal{B}(5)$ is then:

$$\mathcal{B}(5) = \{y \in \mathbb{R} : y \leq 3 \cup y \geq 6\}$$

(b) Let Y_1 and Y_2 be two independent random variables, such that $Y_i \sim \text{Uniform}[0, \theta]$, for $i \in \{1, 2\}$. We want to test:

$$\begin{cases} H_0: & \theta = 1, \\ H_1: & \theta > 1. \end{cases}$$

and we reject H_0 when $\max(Y_1, Y_2) > c$.

(i) [8pt] Find c so that the test has significance level 19/100.

Solution:

$$\frac{19}{100} = \mathbb{P}_{\theta=1}(\max\{Y_1, Y_2\} > c)$$

Being $Y_i \overset{i.i.d.}{\sim} \text{Unif}[0, \theta]$, we have:

$$\mathbb{P}_{\theta=1}(\max\{Y_1, Y_2\} \leq c) = (\mathbb{P}_{\theta=1}(Y_1 \leq c))^2 = c^2$$

so that $c = 9/10$.

(ii) [8pt]Which is the power function of the test (as a function of θ_1 of H_1)?

Solution:

For any $\theta_1 > 1$, we have:

$$\begin{aligned} \pi(\theta_1) &= \mathbb{P}_{\theta_1}\left(\max\{Y_1, Y_2\} > \frac{9}{10}\right) \\ &= 1 - \mathbb{P}_{\theta_1}\left(\max\{Y_1, Y_2\} \leq \frac{9}{10}\right) \\ &= 1 - \left(\mathbb{P}_{\theta_1}\left(Y_1 \leq \frac{9}{10}\right)\right)^2 = 1 - \frac{81}{100\theta_1^2} \end{aligned}$$

3. Consider a multinomial distribution with probability mass function:

$$\mathbb{P}(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3) = \frac{m!}{y_1! y_2! y_3!} p_1^{y_1} p_2^{y_2} p_3^{y_3}$$

with $\sum_{i=1}^3 p_i = 1$ and $\sum_{i=1}^3 y_i = m$.

(a) [8pt] Show that the maximum likelihood estimate $\hat{p}_i$ of p_i is y_i/m with $i \in \{1, 2, 3\}$.

Solution:

We maximize the log-likelihood subject to the normalization constrain, i.e. we introduce a Lagrange multiplier λ, so that:

$$\ell(p; \lambda) = \log(m!) - \sum_{i=1}^3 \log y_i! + \sum_{i=1}^3 y_i \log p_i + \lambda(\sum_{i=1}^3 p_i - 1)$$

We obtain the extreme points $\hat{p}_i = -\frac{y_i}{\lambda}$. Being $\sum_{i=1}^3 y_i = m$ and $\sum_{i=1}^3 \hat{p}_i = 1$, we find that $\lambda = -m$, so that $\hat{p}_i = \frac{y_i}{m}$

(b) [8pt] It is suspected that $p_1 = p_2 = p$, where $0 < p < 1$. Show that the maximum likelihood estimate $\hat{p}$ of p is then $(y_1 + y_2)/(2m)$.
Solution:
Using the same arguments of point (a), we obtain the extreme points $\hat{p} = -\frac{y_1+y_2}{2\lambda}$, $\hat{p}_3 = -\frac{y_3}{\lambda}$. Being $\sum_{i=1}^{3} y_i = m$ and $2\hat{p} + \hat{p}_3 = 1$, we find that $\lambda = -m$, so that $\hat{p} = \frac{y_1+y_2}{2m}$

(c) [10pt] Find the generalized likelihood ratio test statistic for comparing the two models of point (a) and point (b). State its asymptotic distribution and find the rejection region for a test at $\alpha = 0.05$ level of significance.
Solution:
We will test:

$$\begin{cases} H_0: & p \in \Theta_0, \\ H_1: & p \in \Theta \end{cases}$$

where: $\Theta := \{p \in [0,1]^3 : \sum_{i=1}^{3} p_i = 1\}$ and $\Theta_0 := \{p \in \Theta_0 : p_1 = p_2\}$. Therefore $dim(\Theta) = 2$ and $dim(\Theta_0) = 1$. The GLRT statistic is then:

$$\Lambda(\mathbf{Y}) = \frac{\sup_{p \in \Theta_0} L(p; \mathbf{Y})}{\sup_{p \in \Theta} L(p; \mathbf{Y})}$$

Hence,

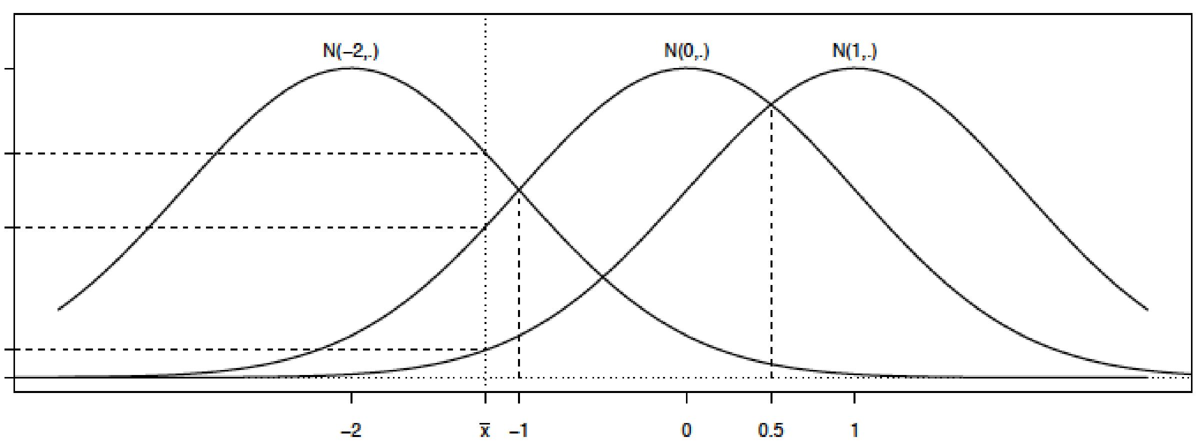
$$\Lambda(\mathbf{Y}) = \frac{\left(\frac{Y_1+Y_2}{2m}\right)^{Y_1+Y_2}}{Y_1^{Y_1} Y_2^{Y_2}}$$

By the Wilks theorem we know that $-2\log\Lambda(\mathbf{Y}) \xrightarrow{d} \chi_1^2$, so that we will reject H_0 at 0.05 level of significance for $-2\log\Lambda(\mathbf{Y}) > \chi_1^2(0.05) \approx 3.84$.

4. Consider the sample $\mathbf{X} = \{X_1, \ldots X_n\}$ of i.i.d. random variables such that $X_i \sim N(\theta, \sigma^2)$ with σ^2 **known** and $\theta \in \Theta$, where the parameter space is the discrete set $\Theta = \{-2, 0, 1\}$.

(a) [4pt] Show that $\bar{X} := 1/n \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ and that the likelihood $L(\theta; \mathbf{X})$ can be factorized in $L(\theta; \mathbf{X}) = h(\mathbf{X}) g_\theta(\bar{X})$, with:

$$h(\mathbf{X}) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right\}, \qquad g_\theta(\bar{X}) = \exp\left\{-\frac{n}{2\sigma^2}(\bar{X} - \theta)^2\right\}$$



In the figure above the function $g_\theta(y)$ is plotted for the three possible values of the parameter θ.
Solution:

$$\begin{aligned} L(\theta;\mathbf{X}) &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\theta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\bar{X}+\bar{X}-\theta)^2\right\} \\ &= (2\pi\sigma^2)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-\bar{X})^2\right\}\exp\left\{-\frac{n}{2\sigma^2}(\bar{X}-\theta)^2\right\} \end{aligned}$$

(b) [8pt] Find a maximum likelihood estimator $\hat{\theta}_{MLE}$ of θ.
Solution:
From the Figure follows that:

$$\hat{\theta}_{MLE} = \begin{cases} -2 & \text{if} \quad \bar{X} < -1, \\ 0 & \text{if} \quad -1 \le \bar{X} \le 0.5 \\ 1 & \text{if} \quad \bar{X} > 0.5 \end{cases}$$

(c) [8pt] Find the probability mass function of $\hat{\theta}_{MLE}$.
Solution:
If we denote with θ_0 the *true* value of the parameter θ, we have:

$$\mathbb{P}(\hat{\theta}_{MLE} = t|\theta_0) = \begin{cases} \Phi((-1-\theta_0)\sqrt{n}/\sigma) & \text{if} \quad t=-2, \\ \Phi((0.5-\theta_0)\sqrt{n}/\sigma) - \Phi((-1-\theta_0)\sqrt{n}/\sigma) & \text{if} \quad t=0, \\ 1-\Phi((0.5-\theta_0)\sqrt{n}/\sigma) & \text{if} \quad t=1, \end{cases}$$

for $\theta_0 \in \{-2,0,1\}$ and where $\Phi(\cdot)$ denote the CDF of the standard normal distribution.

(d) [4pt] Is $\hat{\theta}_{MLE}$ a biased estimator?
Solution:
$\hat{\theta}_{MLE}$ is a biased estimator. In fact, in case $\theta_0 = -2$, we have:

$$\mathbb{E}(\hat{\theta}_{MLE}|\theta_0=-2) = -2\Phi(\sqrt{n}/\sigma)+1-\Phi(2.5\sqrt{n}/\sigma) = -2\Phi(\sqrt{n}/\sigma)+\Phi(-2.5\sqrt{n}/\sigma) > -2\Phi(\sqrt{n}/\sigma) > -2$$

because $0 < \Phi(\cdot) < 1$.

(e) [6pt] Find the most powerful test for testing:

$$\begin{cases} H_0: & \theta = 0, \\ H_1: & \theta = 1, \end{cases}$$

at α level of significance. Can you say something about the rejection region of this test?
Solution:
By the Neyman-Pearson Lemma, the most powerful test is the LRT:

$$\Lambda(\bar{X}) = \frac{L(\theta=0;\bar{X})}{L(\theta=1;\bar{X})} = \frac{g_0(\bar{X})}{g_1(\bar{X})} = \exp\left(-\frac{n}{\sigma^2}\bar{X} + \frac{n}{2\sigma^2}\right)$$

and we reject for small values of this statistic. Therefore we reject for large values of $\bar{X}$, whose distribution under H_0 is $N(0,\sigma^2/n)$.