Statistiek (WISB263)

Final Exam (Sketch of Solutions)

June 27, 2022

Schrijf uw naam op elk in te leveren vel. Schrijf ook uw studentnummer op blad 1.
(The exam is an *open--book* exam: notes and book are allowed. The use of a laptop is allowed as well, under the restriction that the invigilator can look at the screen at all times, and that students are not allowed to type on the computer and wifi is off. The scientific calculator is also allowed).

The maximum number of points is 110 (10 extra BONUS points!!).
Grade= $min(100, \text{points})$.
Points distribution: 25–20–20-25-10 (+10 extra BONUS points!!)

1. Suppose that n integers are drawn uniformly at random with replacement from the set $\mathcal{I} \equiv \{1, 2, \ldots, N\}$.

   (a) [5pt] Find the method of moments estimator $\hat{N}_1$ of N.
   Solution: If $X \sim \text{Unif}[\{1, \ldots, N\}]$, then:

   $$\mathbb{E}(X) = \frac{1}{N}\sum_{i=1}^{N} i = \frac{N+1}{2}$$

   Given the sample $\mathbf{X} = \{X_1, \ldots, X_n\}$, with $X_i \overset{i.i.d}{\sim} \text{Unif}[\{1, \ldots, N\}]$, by the definition of the *method of moments* we have:

   $$\frac{\hat{N}_1 + 1}{2} = \overline{\mathbf{X}}_n \implies \hat{N}_1 = 2\overline{\mathbf{X}}_n - 1$$

   (b) [5pt] Calculate $\mathbb{E}(\hat{N}_1)$ and $\text{Var}(\hat{N}_1)$.
   Solution:

   $$\mathbb{E}(\hat{N}_1) = 2\mathbb{E}(\overline{\mathbf{X}}_n) - 1 = 2\mathbb{E}(X_i) - 1 = N + 1 - 1 = N$$

   so that $\hat{N}_1$ is an unbiased estimator of N.

   $$\begin{aligned}\text{Var}(\hat{N}_1) &= 4\text{Var}(\overline{\mathbf{X}}_n) = \frac{4}{n}\text{Var}(X_i) = \frac{4}{n}\left(\frac{1}{N}\sum_{i=1}^{N} i^2 - \frac{(N+1)^2}{4}\right) = \frac{4}{n}\left(\frac{(N+1)(2N+1)}{6} - \frac{(N+1)^2}{4}\right) \\ &= \frac{N+1}{3n}(4N + 2 - 3N - 3) = \frac{N^2 - 1}{3n}\end{aligned}$$

   (c) [7pt] Find the Maximum Likelihood Estimator (MLE) $\hat{N}_2$ of N.
   Solution: The likelihood $L(N; \mathbf{X})$, can be written as:

   $$L(N; \mathbf{X}) = \frac{1}{N^n}\prod_{i=1}^{n} \mathbf{1}_{\{X_i \in \mathcal{I}\}} = \frac{1}{N^n}\mathbf{1}_{\{1 \le \max_i X_i \le N\}}$$

   Since $L(N; \mathbf{X})$ is decreasing in N, it follows that $\hat{N}_2 = \max_i X_i$.

   (d) [4pt] Show that $\mathbb{E}(\hat{N}_2)$ is approximately $\frac{n}{n+1}N$ (use a continuous approximation of the discrete distribution of the MLE). With the same approximation show that $\text{Var}(\hat{N}_2)$ is approximately $N^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)$.
   Solution: If $Y := \max_i X_i$, then the CDF of the maximum of i.i.d. uniform RVs, by using the continuous approximation:

   $$F_Y(y) = (F_{X_i}(y))^n \approx \frac{y^n}{N^n}$$

   so that the pdf $f_Y(y)$ is:

   $$f_Y(y) = \frac{n y^{n-1}}{N^n}$$

Therefore

$$\mathbb{E}(Y) \approx \frac{n}{N^n}\int_0^N y^n dy = \frac{n}{n+1}N$$

$$\mathrm{Var}(Y) \approx \frac{n}{N^n}\int_0^N y^{n+1}dy - \left(\frac{n}{n+1}N\right)^2 = \frac{n}{n+2}N^2 - \left(\frac{n}{n+1}N\right)^2 = N^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right)$$

(e) [4pt] Suppose that you have collected the sample $\mathbf{x} = \{28, 6, 22, 15\}$. Among $\hat{N}_1, \hat{N}_2$, which estimator would you prefer for estimating N?
Solution: In our sample $n = 4$, so that if we compute the mean squared errors (MSE):

$$\mathrm{MSE}(\hat{N}_1) = \frac{N^2-1}{3n} = \frac{N^2-1}{12}$$

$$\mathrm{MSE}(\hat{N}_2) = \left(N - \frac{4}{5}N\right)^2 + N^2\left(\frac{4}{6} - \frac{16}{25}\right) = \frac{1}{25}N^2 + \frac{2}{75}N^2 = \frac{1}{15}N^2$$

Since in our case $N \geq 28$, the MLE estimator is preferable.

2. Let $\mathbf{X} = \{X_1, \dots X_n\}$ a random sample of i.i.d. random variables with probability density function (pdf):

$$f_X(x) = \frac{1-\beta}{x^\beta},$$

with $x \in (0, 1)$ and $0 < \beta < 1$.

(a) [6pt] Find a sufficient statistic for β.
Solution: We write the likelihood $L(\beta; \mathbf{X})$ as:

$$L(\beta; \mathbf{X}) = (1-\beta)^n \frac{1}{\prod_{i=1}^n X_i^\beta}\mathbf{1}_{\{0<X_{(1)}\}}\mathbf{1}_{\{X_{(n)}<1\}},$$

so that from the factorization theorem $T := \prod_{i=1}^n X_i$ is a sufficient statistic (i.e., $h(\mathbf{X}) = 1$).

(b) [8pt] Determine the MLEs for β and for $n/(\beta-1)$. Can you write these MLEs in terms of the sufficient statistic? Explain clearly the reason.
Solution: The log-likelihood can be written as:

$$\ell(\beta; \mathbf{X}) = n\log(1-\beta) - \beta\sum_{i=1}^n \log X_i$$

and its derivatives:

$$\ell'(\beta; \mathbf{X}) = \frac{-n}{1-\beta} - \sum_{i=1}^n \log X_i$$

$$\ell''(\beta; \mathbf{X}) = \frac{-n}{(1-\beta)^2} \leq 0$$

so that the MLE can be written from the score equations:

$$\hat{\beta} = \frac{n}{\sum_{i=1}^n \log X_i} + 1$$

By the *invariance theorem*, the MLE of $n/(\beta-1)$ is $n/(\hat{\beta}-1)$, so that:

$$\frac{n}{\hat{\beta}-1} = \sum_{i=1}^n \log X_i = \log\prod_{i=1}^n X_i = \log T$$

By the *factorization theorem*, the MLE has to be indeed a function of the sufficient statistic T, since g_θ is the only factor depending on the parameter.

(c) [6pt] Which is the asymptotic variance of the MLE?
Solution: By the asymptotic theory of MLE the asymptotic variance of β is $\frac{1}{nI(\beta)}$, with

$$I(\beta) = -\mathbb{E}\left(\frac{\partial^2}{\partial\beta^2}\log f_X(X)\right) = \frac{1}{(1-\beta)^2}$$

3. The time Y (in years) served in prison for a certain crime was believed to follow a distribution with the following probability density function (pdf):

$$f_Y^{(1)}(y) = \frac{1}{9}y^2, \ \ 0 \le y \le 3.$$

In order to check this law, a study was conducted in one prison on 100 people. It was reported that 16 convicted served less than one year in jail, 32 served between one and two years, and 52 served between two and three years.

(a) [8pt] Are these data consistent with the pdf $f_Y^{(1)}(y)$? Perform an appropriate hypothesis test at 0.05 level of significance.
Solution: The probability of the multinomial model are:

$$\pi_1^{(1)} = \int_0^1 f_Y^{(1)}(y)dy = \frac{1}{27}; \ \ \pi_2^{(1)} = \int_1^2 f_Y^{(1)}(y)dy = \frac{7}{27}, \ \ \pi_3^{(1)} = 1 - \pi_1^{(1)} - \pi_2^{(1)} = \frac{19}{27},$$

We perform a goodness of fit test with the Pearson's χ^2 statistic (asymptotically equivalent to the GLRT for the multinomial distribution, with the sample $\mathbf{x} = (16, 32, 52)$ and $n = 100$.

$$X^2 = \sum_{i=1}^{3} \frac{(x_i - n\pi_i^{(1)})^2}{n\pi_i^{(1)}} = 45.81$$

We reject H_0 at 0.05 level of significance since $X^2 > \chi_2^2(0.05) = 5.99$.

(b) [7pt] Another theory was proposed for the distribution of Y. According to this new theory, we have a linear pdf:

$$f_Y^{(2)}(y) = \frac{2}{9}y, \ \ 0 \le y \le 3.$$

Perform the same analysis as in point (a).
Solution: The probability of the multinomial model are:

$$\pi_1^{(2)} = \int_0^1 f_Y^{(2)}(y)dy = \frac{1}{9}; \ \ \pi_2^{(2)} = \int_1^2 f_Y^{(2)}(y)dy = \frac{3}{9} =, \frac{1}{3} \ \pi_3^{(2)} = 1 - \pi_1^{(2)} - \pi_2^{(2)} = \frac{5}{9},$$

We perform a goodness of fit test with the Pearson's χ^2 statistic (asymptotically equivalent to the GLRT for the multinomial distribution, with the sample $\mathbf{x} = (16, 32, 52)$ and $n = 100$.

$$X^2 = \sum_{i=1}^{3} \frac{(x_i - n\pi_i^{(2)})^2}{n\pi_i^{(2)}} = 2.43$$

We do not reject H_0 at 0.05 level of significance since $X^2 < \chi_2^2(0.05)$.

(c) [5pt] Imagine now we have only one realization y of the random variable Y. Perform the most powerful test for testing:

$$\begin{cases} H_0: & f_Y = f_Y^{(1)}, \\ H_1: & f_Y = f_Y^{(2)}. \end{cases}$$

If $\alpha = 0.05$, find the rejection region of the test and its power. In case $y = 0.3$ do we reject H_0 at the 0.05 level of significance?
Solution: By Neyman-Pearson Lemma, the most powerful test for testing H_0 vs. H_1 is a Likelihood Ratio test, whose test statistic is:

$$\Lambda(Y) = \frac{f_Y^{(1)}(Y)}{f_Y^{(2)}(Y)} = \frac{Y}{2}$$

The rejection region is of the type $\{y : \Lambda(y) < K\}$, with K such that:

$$0.05 = \mathbb{P}(\Lambda(Y) < K | H_0) = \frac{1}{9}\int_0^K y^2 dy = \frac{K^3}{27}$$

Therefore $K = 1.11$. As regards the power:

$$\pi = \mathbb{P}(\Lambda(Y) < K | H_1) = \frac{2}{9}\int_0^K y dy = \frac{2}{9}\int_0^{1.11} y dy = \frac{(1.11)^2}{9} = 0.14.$$

4. Consider the two samples $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, with $X_i \overset{i.i.d}{\sim} N(\theta_1, \theta_2^2)$ and $\mathbf{Y} = \{Y_1, Y_2, \ldots, Y_m\}$, with $Y_i \overset{i.i.d}{\sim} N(\theta_3, \theta_4^2)$. The two samples are independent, i.e., $X_i \perp\!\!\!\perp Y_j, \forall i, j$.

   (a) [4pt] Find the Maximum Likelihood Estimators (MLE) of $\theta_1, \theta_2, \theta_3$ and θ_4.
   Solution: Given $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, we can write the Likelihood as:

$$L(\theta_1, \theta_2, \theta_3, \theta_4; \mathbf{X}, \mathbf{Y}) = \left(\frac{1}{2\pi\theta_2^2}\right)^{n/2} e^{-\frac{1}{2\theta_2^2}\sum_{i=1}^n (X_i - \theta_1)^2} \left(\frac{1}{2\pi\theta_4^2}\right)^{m/2} e^{-\frac{1}{2\theta_4^2}\sum_{i=1}^m (Y_i - \theta_3)^2}$$

   By standard calculations, we find:

$$\hat{\theta}_1 = \overline{\mathbf{X}}_n;\ \hat{\theta}_2^2 = S_X^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{\mathbf{X}}_n)^2;\ \hat{\theta}_3 = \overline{\mathbf{Y}}_m;\ \hat{\theta}_4^2 = S_Y^2 = \frac{1}{m}\sum_{i=1}^n (Y_i - \overline{\mathbf{Y}}_m)^2$$

   (b) [7pt] if we want to test:

$$\begin{cases} H_0 : & \theta_1 = \theta_3 \\ H_1 : & \theta_1 \neq \theta_3. \end{cases}$$

   with θ_2 and θ_4 unknown, find the null parameter space Θ_0 and the Generalized Likelihood Ratio Test (GLRT) statistics. Which is its asymptotic distribution?
   Solution: Given $\theta = (\theta_1, \theta_2, \theta_3, \theta_4)$, we have that the parameter space $\Theta = \{\theta \in \mathbb{R}^4 : \theta_2 > 0, \theta_4 > 0\}$, $\dim(\Theta) = 4$, and the null parameter space:

$$\Theta_0 = \{\theta \in \Theta : \theta_1 = \theta_3\},\ \dim(\Theta_0) = 3$$

   Let us denote the common value of θ_2 and θ_4 with θ_0. Then:

$$\Lambda(\mathbf{X}, \mathbf{Y}) = \frac{\sup_{\theta \in \Theta_0} L(\theta_0, \theta_2, \theta_0, \theta_4; \mathbf{X}, \mathbf{Y})}{\sup_{\theta \in \Theta} L(\theta_1, \theta_2, \theta_3, \theta_4; \mathbf{X}, \mathbf{Y})} = \frac{L(\hat{\theta}_0^{(0)}, \hat{\theta}_2^{(0)}, \hat{\theta}_0^{(0)}, \hat{\theta}_4^{(0)}; \mathbf{X}, \mathbf{Y})}{L(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4; \mathbf{X}, \mathbf{Y})}$$

   with

$$\hat{\theta}_0^{(0)} = \frac{n\overline{\mathbf{X}}_n + m\overline{\mathbf{Y}}_m}{n+m};\ \ (\hat{\theta}_2^{(0)})^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \hat{\theta}_0^{(0)})^2;\ (\hat{\theta}_4^{(0)})^2 = \frac{1}{m}\sum_{i=1}^m (Y_i - \hat{\theta}_0^{(0)})^2$$

   and $\hat{\theta}_i$ as calculated in point (a). By Wilks's theorem, we have that $-2\log(\Lambda(\mathbf{X}, \mathbf{Y})) \overset{d}{\to} X$, as $n, m \to \infty$ and where X is χ^2 with $d = \dim(\Theta) - \dim(\Theta_0) = 1$ degree of freedom.

   (c) [7pt] If we assume now that $\theta_2 = \theta_4$ (still unknown), find also in this case the GLRT statistics and its asymptotic distribution for:

$$\begin{cases} H_0 : & \theta_1 = \theta_3 \\ H_1 : & \theta_1 \neq \theta_3. \end{cases}$$

   Solution: In this case the parameter space is $\Theta = \{\theta \in \mathbb{R}^4 : \theta_2 = \theta_4 > 0\}$, $\dim(\Theta) = 3$, while the null parameter space has the form:

$$\Theta_0 = \{\theta \in \Theta : \theta_1 = \theta_3\},\ \dim(\Theta_0) = 2$$

Let us pose $\theta_2 = \theta_4 = \sigma_0$ and $\theta_1 = \theta_3 = \mu$. Then:

$$\Lambda(\mathbf{X}, \mathbf{Y}) = \frac{\sup_{\theta \in \Theta_0} L(\mu, \sigma_0, \mu, \sigma_0; \mathbf{X}, \mathbf{Y})}{\sup_{\theta \in \Theta} L(\theta_1, \sigma, \theta_3, \sigma; \mathbf{X}, \mathbf{Y})} = \frac{L(\hat{\mu}, \hat{\sigma}_0, \hat{\mu}, \hat{\sigma}_0; \mathbf{X}, \mathbf{Y})}{L(\hat{\theta}_1, \hat{\sigma}, \hat{\theta}_3, \hat{\sigma}; \mathbf{X}, \mathbf{Y})}$$

with

$$\hat{\mu} = \frac{n\overline{\mathbf{X}}_n + m\overline{\mathbf{Y}}_m}{n+m};\ \ \hat{\sigma}_0^2 = S_p^2 = \frac{1}{n+m}\left(n\sum_{i=1}^{n}(X_i - \hat{\mu})^2 + m\sum_{i=1}^{m}(Y_i - \hat{\mu})^2\right);$$

$$\hat{\sigma}^2 = \frac{1}{n+m}\left(n\sum_{i=1}^{n}(X_i - \overline{\mathbf{X}}_n)^2 + m\sum_{i=1}^{m}(Y_i - \overline{\mathbf{Y}}_m)^2\right),\ \hat{\theta}_1 = \overline{\mathbf{X}}_n,\ \hat{\theta}_3 = \overline{\mathbf{Y}}_m$$

By Wilks's theorem, we have that $-2\log(\Lambda(\mathbf{X}, \mathbf{Y})) \xrightarrow{d} X$, as $n, m \to \infty$ and where X is χ^2 with $d = \dim(\Theta) - \dim(\Theta_0) = 1$ degree of freedom.

(d) [7pt] We collected the samples $\mathbf{x} = \{9.00, 10.87, 11.32\}$, $\mathbf{y} = \{9.54, 7.82, 7.74, 7.14\}$. Perform the test of point (c) at 0.05 level of significance by using a GLRT.
Solution: By equivalence of GLRT and the two-sample t-test for normal samples with equal variances, we just perform a t-test, with a realization of the test statistic:

$$t = 2.74$$

Since $t > t_7(0.025) = 2.57$, we reject H_0 at the 0.05 level of significance.

5. [10pt] We have performed an experiment and we collected 48 measurements. However, each of these 48 real numbers is rounded to the nearest integer. The sum of the original 48 numbers is approximated by the sum of these integers. If we assume that the errors made by rounding off are *i.i.d.* and have a uniform distribution over the interval $(-1/2, 1/2)$, compute *approximately* the probability that the sum of the integers is within two units of the *true* sum.
Solution: Given the measurements $\mathbf{X} = (X_1, \ldots, X_{48})$, and the rounded off integers $\mathbf{I} = (I_1, \ldots, I_{48})$, we have the model:

$$X_i = I_i + \epsilon_i$$

with $\epsilon_i \overset{i.i.d.}{\sim} \text{Unif}[-0.5, 0.5]$. Since $\mathbb{E}(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \mathbb{E}(\epsilon_i^2) = \int_{-1/2}^{1/2} x^2 dx = \frac{1}{12}$, we have that:

$$\Delta_i := X_i - I_i \overset{i.i.d.}{\sim} \text{Unif}[-0.5, 0.5].$$

We want to calculate

$$p_n := \mathbb{P}(|\sum_{i=1}^{n} \Delta_i| < 2) = \mathbb{P}\left(\frac{|\sum_{i=1}^{n} \Delta_i|}{\sqrt{n\text{Var}(\Delta_i)}} < \frac{2}{\sqrt{48/12}}\right) = \mathbb{P}\left(\frac{|\sum_{i=1}^{n} \Delta_i|}{\sqrt{n\text{Var}(\Delta_i)}} < 1\right)$$

by classical CLT, we know that $\frac{\sum_{i=1}^{n} \Delta_i}{\sqrt{n\text{Var}(\Delta_i)}} \xrightarrow{d} Z \sim N(0,1)$ as $n \to \infty$, so that:

$$p_n \approx \mathbb{P}(|Z| < 1) = \Phi(1) - \Phi(-1) \approx 0.68$$

BONUS [10pt] Imagine that you are a fraudulent scientist and that you have decided to repeat the experiment if its p-value is smaller than a priori fixed value β, with $0 < \beta < 1$, and then to report only the largest one. Show that under H_0, the cumulative distribution function of the p-values of the fraudulent experiment is:

$$F(x; \beta) = \begin{cases} x^2, & \text{if } 0 \le x \le \beta, \\ (1+\beta)x - \beta, & \text{if } \beta < x \le 1. \end{cases}$$