**WISB263 Mathematical Statistics**
**Resit - 14th July 2020**
**Total amount of points: 100**
**Grade exam: number of points/10 rounded to 0.5 or integers**

Lecturer: Dr. Wioletta Ruszel
Teaching assistants: Leandro Chiarini, Ben Balkenende, Esther Steenkamer

**Exercise 1** (25 p.) We want to investigate the percentage of products on a production line which are defective. Components are inspected independently until the first defective is encountered. We observe that 5 distinct components are found defective after the first inspection, 11 distinct components are found defective at their second inspection (the same distinct 11 components were checked twice and at the second inspection they were broken), 12 at their third inspection, 8 at their forth inspection, 6 at their fifth inspection and 8 at their sixth inspection.

We have checked in total $n = 50$ components (the components which were checked and not defective are not considered here) and performed $\sum_{i=1}^{50} x_i = 173$ inspections. $x_i$ denotes the number of inspections of component $i$ until the first time it is defective.

1. (10 p.) Perform a hypothesis test to decide whether the number of inspections until the first defective one is encountered is geometrically distributed at $1 - \alpha = 0.99$ significance. What is the p-value?

2. (10 p.) Under the assumption that the data comes from a geometric distribution with parameter $p$, perform a hypothesis test for $(H_0) : p = 0.3$ against $(H_1) : p \neq 0.3$ using a generalized likelihood ratio statistic $R(\mathbf{X})$ at $\alpha = 0.01$. You can use the appropriate asymptotic distribution of $R(\mathbf{X})$. Compute the p-value. (Hint: You can use that the MLE is equal to $\hat{p}_n = \frac{1}{\overline{X}_n}$.)

3. (5 p.) Compare both tests, which one is more appropriate? Comment also on the assumption that the underlying distribution is geometric in the second test. Was it a reasonable assumption?

**Solution:**

1. We perform a goodness-of-fit test, Theorem 10.2. (1 point) If the data comes from a geometric distribution then $\mathbb{P}(X_i = k) = p(1-p)^{k-1}$. The MLE of $p$ is equal to $\frac{1}{\overline{X}_n}$. (1 point) From the data we have the estimate for $p$

$$\hat{p}_{50} = \left(\frac{1 \cdot 5 + 2 \cdot 11 + ... + 6 \cdot 8}{50}\right)^{-1} = 0.29. \text{(1 point)}$$

| Nr of inspections until first defective | Observations | $\hat{p}_{i_j}$ | $n\hat{p}_{i_j}$ |
|---|---|---|---|
| 1 | 5 | 0.29 | 14.5 |
| 2 | 11 | 0.20 | 10 |
| 3 | 12 | 0.15 | 7.5 |
| 4 | 8 | 0.10 | 5 |
| 5 | 6 | 0.07 | 3.5 |
| 6 | 8 | 0.05 | 2.5 |

(2 points)

| Nr of inspections until first defective | Observations | $\hat{p}_{i_j}$ | $n\hat{p}_{i_j}$ |
|---|---|---|---|
| 1 | 5 | 0.29 | 14.5 |
| 2 | 11 | 0.20 | 10 |
| 3 | 12 | 0.15 | 7.5 |
| 4 | 8 | 0.10 | 5 |
| 5+ | 14 | 0.12 | 6.5 |

(2 points)

The test statistic is equal to

$$d_5^2 = \frac{(5-14.5)^2}{14.5} + ... + \frac{(14-6.5)^2}{6.5} = 19.48, \text{(1 point)}$$

we have to compare it to $\chi^2_{5-1-1}(0.99) = 11.34$, (1 point) since $d_5^2 \geq \chi_3^2(0.99)$ we reject $H_0$. (1 point) The p-value is 0.00022. (1 point)

2. Argue that we can use Theorem 9.1. (3 points) The rejection region is given by $\{2\ln(R(\mathbf{X})) \geq \chi_1^2(0.99)\}$ where $\chi_1^2(0.99) = 6.63$. (2 points) We have that

$$2\ln(R(\mathbf{x})) = 2\ln\left(\frac{\overline{x}_{50}^{-50}(1-\overline{x}_{50}^{-1})^{\sum_{i=1}^{50} x_i - 50}}{0.3^{50}(0.7)^{\sum_{i=1}^{50} x_i - 50}}\right) = 0.098 \text{(3 points)}$$

so we would not reject $H_0$. (1 point) The p-value is equal to 0.75. (1 point)

3. The first test since it does not assume anything about the distribution and has lower p-value. (2 points) The assumption on the distribution seems a natural one since these type of questions are typically modelled with geometric distributions. (3 points)

**Exercise 2** (25 p.) Let $X_1, ..., X_N$ be independent Bernoulli random variables $B(p)$ modelling the outcome whether a person $i$ has a sickness or not. We consider the estimator $\hat{p}_n$ for the proportion of people with the sickness in a sample of size $n$ taken out of a population of size $N$ using simple random sample scheme.

1. (6 p.) Determine for which $p \in (0,1)$ the standard error of the estimator $\hat{p}_n$ is maximal.

2. (4 p.) Show that the maximal standard error of the estimator of the variance of $\hat{p}_n$ is equal to $\frac{1}{2}\sqrt{\frac{N-n}{N(n-1)}}$.

3. (4 p.) Choose now $N = 1000$. How many samples $n$ do you need to ensure that the standard error is at most 0.05?

4. (8 p.) Use the result from (2) to construct an asymptotic confidence interval for $p$ at confidence of at least 0.99.

5. (3 p.) Let $N = 1000$. If in the simple random sample 40 out of 100 tested people have the sickness, determine the 99% (asymptotic) confidence interval for $p$.

**Solution:**

1. The standard error of the estimator is equal to

$$\sigma_{\hat{p}_n} = \sqrt{\frac{p(1-p)}{n}\frac{N-n}{N-1}}. \text{ (2 points)}$$

It is maximal for $p$ if and only if $\sigma^2_{\hat{p}_n}$ is maximal for $p$. Consider

$$\frac{d}{dp}\sigma^2_{\hat{p}_n} = \frac{1-2p}{n}\frac{N-n}{N-1} \text{ (2 points)}$$

which is equal to 0 for $p = \frac{1}{2}$, since $\frac{d^2}{dp^2}\sigma^2_{\hat{p}_n} = -\frac{2}{n}\frac{N-n}{N-1} < 0$ we have that it is indeed a maximum. (2 points)

2. Now we want to maximize the square root of

$$s^2_{\hat{p}_n} = \frac{\hat{p}_n(1-\hat{p}_n)}{n-1}\frac{N-n}{N}, \text{ (2 points)}$$

using (1) we get that

$$\sup_{\hat{p}_n \in (0,1)} s_{\hat{p}_n} = \frac{1}{2}\sqrt{\frac{N-n}{N(n-1)}}. \text{ (2 points)}$$

3. We look for $n$ such that

$$\frac{1}{2}\sqrt{\frac{N-n}{N(n-1)}} \leq 0.05 \text{ (2 points)}$$

which is true for $n \geq \lfloor 91.81 \rfloor + 1 = 92$. (2 points)

4. Use Theorem 2.3 from the lecture to deduce that $\frac{\hat{p}_n - p}{s_{\hat{p}_n}}$ converges in distribution to a standard normal random variable. (1 point) Hence asymptotically

$$\mathbb{P}\left(-z\left(\frac{\alpha}{2}\right) \leq \frac{\hat{p}_n - p}{s_{\hat{p}_n}} \leq z\left(\frac{\alpha}{2}\right)\right) \approx 1-\alpha. \text{ (1 point)}$$

which is equivalent to

$$\mathbb{P}\left(\left[\hat{p}_n - z\left(\frac{\alpha}{2}\right)s_{\hat{p}_n}; \hat{p}_n + z\left(\frac{\alpha}{2}\right)s_{\hat{p}_n}\right] \ni p\right) \approx 1-\alpha. \text{ (2 points)}$$

We have that $z(0.005) = 2.58$ (2 points) and we take the maximal standard error since increasing the CI is decreasing $\alpha$. We obtain

$$\mathbb{P}\left(\left[\hat{p}_n - 1.29\sqrt{\frac{N-n}{N(n-1)}}; \hat{p}_n + 1.29\sqrt{\frac{N-n}{N(n-1)}}\right] \ni p\right) \geq 0.99. \text{ (2 points)}$$

5. The CI is equal to $[0.28, 0.52]$. (3 points)

**Exercise 3** (25 p.) We want to study different estimators for the parameter $\lambda$ in an exponential distribution with parameter $\frac{1}{\lambda}$ for $\lambda > 0$.

1. (5 p.) Determine the MoM estimator $\hat{\lambda}_1$ of $\lambda$ using an i.i.d. sample $X_1, ..., X_n$ of exponential random variables with parameter $\frac{1}{\lambda}$. Compute the mean and variance of $\hat{\lambda}_1$.

2. (5 p.) Prove that the rescaled estimator $\sqrt{n}(\hat{\lambda}_1 - \lambda)$ is asymptotically normal as $n \to \infty$ and determine the limiting variance.

3. (5 p.) Determine the MoM estimator $\hat{\lambda}_2$ of $\lambda$ using an i.i.d. sample $X_1, ..., X_n$ of exponential random variables with parameter $\frac{1}{\sqrt{\lambda}}$. Compute the mean and variance of $\hat{\lambda}_2$. (Hint: You can use that $\int_0^\infty x^4 e^{-ax} dx = \frac{24}{a^5}$.)

4. (5 p.) Prove that the rescaled estimator $\sqrt{n}(\hat{\lambda}_2 - \lambda)$ is asymptotically normal as $n \to \infty$ and determine the limiting variance.

5. (5 p.) Is one of the estimators $\hat{\lambda}_1$ or $\hat{\lambda}_2$ most efficient in their respective class $K_b$? Which one is asymptotically more efficient?

**Solution:**

1. Let $X_1, ..., X_n$ be an i.i.d. sample with $X_i \sim Exp(\lambda^{-1})$ (2 points)then the MoM is equal to $\hat{\lambda}_1 = \overline{X}_n$. (1 point) The mean is $\lambda$ and the variance equal to $\frac{\lambda^2}{n}$. (2 points)

2. We have that $q(x) = x$ so $q \in C^1$ and $q'(x) = 1 \neq 0$. (1 point) The variance of $X_i$ is also finite so we can use Theorem 4.4 (1 point) to deduce asymptotic normality of $\sqrt{n}(\hat{\lambda}_1 - \lambda)$. (2 points)The limiting variance is equal to $Var(X_1) = \lambda^2$. (1 point)

3. Let $X_1, ..., X_n$ be an i.i.d. sample with $X_i \sim Exp(\lambda^{-1/2})$ (2 points) then the MoM is equal to $\hat{\lambda}_2^2 = \frac{1}{2n} \sum_{i=1}^n X_i^2$. (2 points) The mean is $\lambda$ and variance $\frac{5\lambda^2}{n}$. (1 point)

4. $q \in C'$ and $q'(x^2) = 2$, (1 point) $Var(X_1^2) < \infty$ so by Theorem 4.4. (1 point) the estimator is asymptotically normal. (1 point) The limiting variance is equal to $5\lambda^2$. (2 points)

5. The first estimator is asymptotically more efficient since its variance is smaller. (1 point) Both estimators are unbiased hence we check for efficiency in $K_0$. We want to use Theorem 6.2 All conditions are trivially satisfied. (1 point)We check whether the Cramer-Rao lower bound is attained. We compute

$$\frac{d^2}{d\lambda^2} \ln(f_\lambda(x_1)) = \frac{1}{\lambda^2} - \frac{2x_1}{\lambda^3} \text{ (1 point)}$$

and hence

$$I(\lambda) = -\mathbb{E}_\lambda \left( \frac{1}{\lambda^2} - \frac{2X_1}{\lambda^3} \right) = \frac{1}{\lambda^2} \text{ (1 point)}$$

so the lower bound is equal to $\frac{1}{nI(\lambda)} = \frac{\lambda^2}{n} = Var_\lambda(\hat{\lambda}_1)$. The first estimator is the unique efficient estimator in $K_0$. (1 point)

**Exercise 4** (25 p.) Consider independent random variables $X_1, ..., X_n$ such that $X_i \sim N(\mu, \sigma^2 x_i)$ and $x_i \neq 0$ for all $i = 1, ..., n$.

1. (5 p.) Write the random variables $X_1, ..., X_n$ as a linear regression model $Y_1, ..., Y_n$ with respect to errors $\epsilon_i$ which have mean 0 and common variance $\sigma^2$ for all $i = 1, ..., n$.

2. (7 p.) Compute directly the unbiased MLE's $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ for $\mu$ resp. $\sigma^2$.

3. (5 p.) Determine the distribution of $\frac{n-1}{\sigma^2}\hat{\sigma}_n^2$.

4. (5 p.) Discuss the hypothesis test $(H_0) : \sigma = 1$ against $(H_1) : \sigma > 1$. Which test statistic could you choose? Determine the rejection region for $\alpha = 0.05$ and $n = 10$. For which $\hat{\sigma}^2_{10}$ do we accept $H_0$?

5. (3 p.) Interpret now the vector $(Y_1, ..., Y_n)$ in (1) as a linear regression model. Which design matrix $\mathbf{X}$ and vector of parameters $B$ correspond to $(Y_1, ..., Y_n)$? Discuss the connection of the MLE $\hat{\mu}_n$ from (2) with the least-square estimator in a simple linear regression model.

**Solution:**

1. We can write first of all
$$X_i = \mu + \epsilon_i' \text{ (1 point)}$$
and $\epsilon_i' \sim N(0, \sigma^2 x_i)$ (1 point) and then the linear model as
$$Y_i = \frac{\mu}{x_i} + \epsilon_i \text{ (1 point)}$$
setting $Y_i = \frac{X_i}{x_i}$ and where $\epsilon_i \sim N(0, \sigma^2)$. (2 points)

2. We get that
$$\hat{\mu}_n = \frac{\sum_{i=1}^n Y_i/x_i}{\sum_{i=1}^n 1/x_i^2} \text{ (5 points)}$$
and $\hat{\sigma}^2_n = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \hat{\mu}_n)^2$. (2 points)

3. From the proof of Theorem 7.1 we know that $\frac{n-1}{\sigma^2}\hat{\sigma}^2_n \sim \chi^2_{n-1}$.

4. We can take as test statistic $(n-1)\hat{\sigma}^2_n$. (1 point) The test rejects for large values of $\sigma$. (1 point) The rejection region is equal to
$$\{(10-1)\hat{\sigma}^2_{10} \geq \chi^2_{10-1}(0.95)\} = \{\hat{\sigma}^2_{10} \geq 1.88\}. \text{ (2 points)}$$
We accept for all $\hat{\sigma}_{10} < 1.88$. (1 point)

5. The MLE $\hat{\mu}_n$ is equal to the least-square estimator of a SLR model. (1 point) The design matrix is equal to the vector $(1/x_1, ..., 1/x_n)^T$ and $B = \mu$. (2 points)