Resit – Mathematical Statistics 2025

Course: 202500380; Module: M5 Statistics and Analysis
Date : November 6, 2025; Time : 08:45 – 11:45 (3hrs)

Instructions

- This test consists of 9 exercises. Please check if your exam paper is complete before you start.
- Please write your name on each sheet of exam paper that you submit.
- Round all numerical results to three digits.
- An ordinary calculator is allowed, not a programmable one (GR).
- The formula sheet is provided separately.
- Please write legibly. We cannot evaluate what we do not understand.
- Always justify your answers. An answer without justification will not be given full points.

Part 1: Basic concepts

Exercise 1. ***1+1+2+2+2+1+1+1+2=13 Points***

a) What is the difference between a categorical and a discrete random variable?

b) Let $\{P_\theta : \theta \in \Theta\}$, θ unknown, be a statistical model for observations $X_1, \dots, X_n$.

(i) Define the bias of an estimator $\hat{\theta}$ for the unknown parameter θ.

(ii) Assuming $X_1, \dots, X_n$ are independent, identically distributed and $\theta = \mathrm{E}(X_1) < \infty$. Provide an example for an unbiased estimator of θ and an example for a biased estimator of θ.

c) Let $\alpha > 0$. Compute the quantile function $Q(p)$ for $p \in (0,1)$ of the distribution with probability density function $f(x) = 0$ for $x < 1$ and $f(x) = \alpha x^{-\alpha-1}$ for $x \geq 1$.

d) For a random variable X with mean μ and variance σ^2, the skewness is defined by $\sigma^{-3} E\left[(X-\mu)^3\right]$. If $a > 0$ and $b \in (-\infty, \infty)$ are non-random, show that the random variables X and $aX + b$ have the same skewness.

e) Describe the difference between uncorrelated and independent random variables.

f) Given an observation X with continuous density f_θ, where $\theta \in \Theta := \{\theta_0, \theta_1\}$ is unknown and $\mathrm{supp}(f_{\theta_0}) = f_{\theta_1}$. Consider the test

$$\Phi_1(X) = \begin{cases} 1, & \text{if } f_1(X) > c_\alpha f_0(X), \\ 0, & \text{if } f_1(X) \leq c_\alpha f_0(X), \end{cases}$$

with c_α chosen such that the test has level α. Assume you want to consider the same test, but for a smaller level α', i.e. you consider

$$\Phi_2(X) = \begin{cases} 1, & \text{if } f_1(X) > c_{\alpha'} f_0(X), \\ 0, & \text{if } f_1(X) \leq c_{\alpha'} f_0(X), \end{cases}$$

where $\alpha' < \alpha$. How does the power of the second test compare to the power of the first test?

g) Let $X_1, \dots, X_n$ be observations from a statistical model $\{P_\theta : \theta \in \Theta\}$, and let $T = T(X_1, \dots, X_n)$ be a statistic. When is T called complete for the unknown parameter θ?

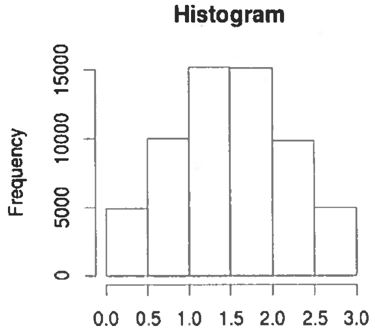
h) Let $X_1, \dots, X_n$ be observations from a statistical model $\{P_\theta : \theta \in \Theta\}$, let $T = T(X_1, \dots, X_n)$ be a statistic and let $\hat{\theta}$ be an unbiased estimator for θ. Under which conditions on T can we generate a minimum-variance unbiased (MVU) estimator from T and $\hat{\theta}$? Provide a formula for the corresponding MVU estimator.

Part 2: Visual interpretation of data

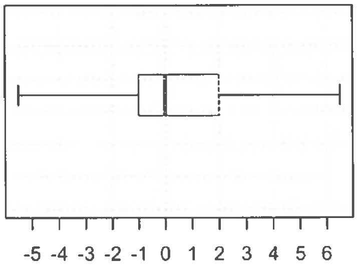
Exercise 2. *2 Points*
Draw a scatter plot for a sample of two negatively correlated random variables. In a second step add one outlier to the dataset, such that the regression line has positive slope.

Exercise 3. *2+2=4 Points*
A distribution can be visualized using histograms or boxplots.

(a) Given the histogram below, draw the corresponding boxplot (ignoring potential outliers).



(b) Given the boxplot below, draw a histogram representing the same data [Histograms are more informative than boxplots. Therefore, given a boxplot there could be many different histograms. Your task is to draw one possible histogram with at least four bins.]



Exercise 4. *2 Points*
The random variable X is generated by the following procedure: Toss a fair coin, that is, heads and tails occur with probability 1/2. If heads appears, draw X from a $\mathcal{N}(2,1)$ distribution and if tails appears draw X from a $\mathcal{N}(-2,1)$ distribution. Make a plot of the probability density function of X (for the solution it is enough to get the shape right) and provide an explanation for the density function's shape.

Part 3: Theory

Exercise 5. ***2+3=5 Points***
Suppose we observe n independent random variables $X_i \sim \text{Pois}(\lambda)$, $i = 1,\dots,n$, where $\text{Pois}(\lambda)$ denotes the Poisson distribution with intensity parameter $\lambda \in (0,\infty)$.

a) Compute the maximum likelihood estimator $\widehat{\lambda}$ for λ.

b) Consider the class of estimators $a\widehat{\lambda}$ for λ with a a real number. For which value of a is the mean squared error (MSE) minimized? Let a^* be the value minimizing the MSE. Is $a^*\widehat{\lambda}$ an estimator?

Exercise 6. ***2+2+2+2=8 Points***
Study participants are sometimes reluctant to share sensitive information as they are concerned about their privacy. Suppose we conduct a study asking study participants whether they feel addicted to alcohol. For each individual, the possible outcome is either *yes* or *no*. Suppose that the outcomes are independent, identically distributed and that *yes* occurs with probability p and *no* with probability $1-p$ with p the unknown proportion parameter of alcohol addicts in the population. To protect the participants' privacy, we propose the following mechanism: roll a die once. If you get "1" or "2", report the wrong answer (if you feel addicted to alcohol choose *no* and if you do not feel addicted to alcohol choose *yes*). If you get a number larger than "2", report the right answer. We are now given a sample $X_1,\dots,X_n$ generated from this mechanism.

a) Show that $P(X_i = \text{yes}) = (1+p)/3$ and $P(X_i = \text{no}) = (2-p)/3$.

b) Propose an unbiased estimator for the true proportion p of alcohol addiction based on the sample $X_1,\dots,X_n$.

c) Compute the variance and the mean-squared error of the estimator you proposed in b).

d) Suppose every participant would have given the correct answer. Then we would have simply taken the relative frequency of participants answering *yes* as estimator for p. Compare the mean-squared error obtained in c) with the mean-squared error of the relative frequency estimator.

Exercise 7. ***2+2+2=6 Points***
Let $X_1,\dots,X_n$ be independent $N(0,\sigma^2)$-distributed random variables with $\sigma \in (0,\infty)$ unknown.

a) Determine the Neyman-Pearson test for the testing problem

$$H_0 : \sigma^2 = 1 \text{ vs. } H_1 : \sigma^2 = 0.5,$$

i.e. the test

$$\phi(X) = \begin{cases} 1, & \text{if} \quad L_1(X_1,\dots,X_n;\sigma^2) \geq cL_0(X_1,\dots,X_n;\sigma^2), \\ 0, & \text{if} \quad L_1(X_1,\dots,X_n;\sigma^2) < cL_0(X_1,\dots,X_n;\sigma^2), \end{cases}$$

for some $c \in \mathbb{R}$ and with $L_1(X_1,\dots,X_n;\sigma^2)$ and $L_0(X_1,\dots,X_n;\sigma^2)$ denoting the likelihood functions under the alternative hypothesis and null hypothesis, respectively. Show that this test can be written as

$$\phi(X) = \begin{cases} 1, & \text{if} \quad \sum_{i=1}^n X_i^2 \leq k, \\ 0, & \text{if} \quad \sum_{i=1}^n X_i^2 > k \end{cases}$$

for some $k \in \mathbb{R}$.

b) Determine k such that the test has exact level 5% and report the power of the level-5%-test. (Express both, k and the power of the test, as a function of the same quantile function/cumulative distribution function that may depend on n.)

c) For $n = 50$ and based on an approximation by the central limit theorem determine k such that the test has approximate level 5%. (You can make use of the fact that if X is $N(0,\sigma^2)$-distributed, $\text{E}(X^4) = 3\sigma^4$.) Your solution should be a numerical value.

Exercise 8. ***2+1+2=5 Points***
Let $X_1,\dots,X_n$ be independent, identically Bernoulli(θ) distributed random variables and let $\tau(\theta) = (1-\theta)^2$.

a) Show that $T := \sum_{i=1}^n X_i$ is a sufficient statistic for $\tau(\theta)$.

b) Show that $\hat{\tau} = 1_{\{X_1+X_2=0\}}$ is an unbiased estimator for $\tau(\theta)$.

c) Use a) and the Rao-Blackwell Theorem to improve the estimator $\hat{\tau}$ in b).

Part 4: Application

Exercise 9. *2+1+1+2=6 Points*

To explore the relationship between the number of weekly training hours and the running performance (5K time in minutes) for a group of runners, the following data has been collected:

training hours per week (x)	5K time (minutes) (Y)
2	29
4	26
6	26
8	23
10	21

Based on the above table, an analyst decides to fit a simple linear regression model.

a) Specify the simple linear regression model (assuming normally distributed measurement errors), state which parameters in the model are unknown, and provide the estimated values for these parameters based on the given information.

b) Use the fitted regression model to predict the 5K time for a runner who trains for $x = 7$ hours per week.

c) Compute the correlation coefficient between training hours and 5K time and interpret the result.

d) Test at 5% level of significance whether the explanatory variable (training hours per week (x)) has an effect on the response variable (5K time (Y)). For this, you may approximate the t-distribution (with any degrees of freedom) by the standard normal distribution.

$$\text{Grade} = 1 + \frac{\#\text{ points}}{51} \times 9$$

Rounded to 1 decimal

question	1	2	3	4	5	6	7	8	9	Total
points possible	1+1+2+2+2+1+1+1+2	2	2+2	2	2+3	2+2+2+2	2+2+2	2+1+2	2+1+1+2	51