Statistics II - Test I

Coursecode: 202300026
Time: 21 January 2026
Time: 13:45 - 16:45
Teacher: Frank Röttger

Student's name:____________________
Student ID:____________________

General information:

- A regular scientific calculator is allowed, a programmable calculator ("GR") is not.
- Other than a pen, no means are needed (or allowed) for answering the exam questions.
- All electronic devices (e.g., phones, smartwatches, earbuds, tablets, laptops) must be switched off and stored away; they may not be on your person or at your desk during the exam.
- Please write your name and student number on every exam paper you hand in.
- Please write legibly. I cannot evaluate what I do not understand.

Achieved points:

Part A:

Task	1	2	3	4	5	6	7	8	9	10	11	Sum
Maximum	1	1	2	1	2	2	1	1	1	1	1	14
Achieved												

Part B:

Task	1	2	3	4	5	6	7	Sum	Total A+B
Maximum	5	6	4	5	5	5	4	34	48
Achieved									

Part A: Basic concepts

1. **(1P)** True or False? If $\mathbf{X}$ follows a d-variate Gaussian distribution, then $A\mathbf{X} + b$ follows an n-variate Gaussian distribution for any $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$.

2. **(1P)** True or False? In logistic regression, the response variable is a categorical variable.

3. Assume n independent categorical variables $X_1, \ldots, X_n$, where for each $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, k\}$ we have

$$\mathbb{P}(X_i = j) = p_j$$

with $0 < p_j < 1$ for all j and $\sum_{j=1}^{k} p_j = 1$. Let $N_j = \sum_{i=1}^{n} \mathbf{1}\{X_i = j\}$.

   (i) **(1P)** Which distribution does $\mathbf{N} = (N_1, \ldots, N_k)$ follow?

   (ii) **(1P)** True or false? The sum $\sum_{j=1}^{k} N_j$ is equal to $n - 1$.

4. **(1P)** True or false? The skewness of a normal random variable is zero.

5. Let $Z_1, \ldots, Z_n$ be i.i.d. standard normal, and let $m_i = \mathbf{E}(Z_{(i)})$ be the expectation of the i-th order statistic for all $i \in \{1, \ldots, n\}$. Let $x_1, \ldots x_n$ be a realization of an i.i.d. random sample $X_1, \ldots, X_n$.

   (i) **(1P)** If X_i follows a normal distribution, which values would you expect for the sample correlation between $x_{(1)}, \ldots, x_{(n)}$ and $m_1, \ldots, m_n$?

   (ii) **(1P)** Can you name a nonparametric test that is based on the sample correlation from (i)?

6. Consider independent samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ for a Wald–Wolfowitz runs test.

   (i) **(1P)** True or false? Under the null hypothesis, the number of runs R in the joint sample follows a binomial distribution.

   (ii) **(1P)** True or false? Under the null hypothesis, the number of runs can be standardized by its mean and standard deviation such that it converges to the standard normal distribution.

7. **(1P)** Explain the data-generating mechanism in the nonparametric Bootstrap world.

8. **(1P)** What is the main difference between Bayesian statistics and frequentist statistics?

(c) **(2P)** Estimate the parameters in your models from (a) for X, Y, and (X, Y).

(d) **(1P)** Compute the value of the test statistic from (b) for the given data.

3. Let X be a Uniform$(0, a)$ random variable with density $f_X(x) = \frac{1}{a}\mathbf{1}\{0 \le x \le a\}$, and let Y be a Exp(λ) random variable with density $f_Y(y) = \lambda e^{-\lambda y}\mathbf{1}\{y \ge 0\}$.

   Recall that X is said to be *stochastically smaller* than Y (written $X \le_{st} Y$) if

$$\mathbb{P}(X > t) \le \mathbb{P}(Y > t) \qquad \text{for all } t \in \mathbb{R}.$$

   (a) **(2P)** Compute the survival functions $\mathbb{P}(X > t)$ and $\mathbb{P}(Y > t)$.

   (b) **(2P)** Determine all parameter values $a > 0$ and $\lambda > 0$ for which the stochastic ordering $X \le_{st} Y$ holds.

4. A researcher wants to investigate whether a new teaching method improves student performance. Two independent groups of students were tested after a 4-week instructional period:

   - **Group A** (standard method): 72, 65, 78, 70, 69, 73
   - **Group B** (new method): 80, 74, 85, 79, 83

   The researcher decides to use the **Wilcoxon rank–sum test** to compare the groups.

   (a) **(1P)** State the null hypothesis for the Wilcoxon rank–sum test.

   (b) **(1P)** Combine the observations from both groups and assign ranks to all values.

   (c) **(1P)** Compute the rank sum for group A.

   (d) **(2P)** In class, we learned that the rank sum test statistic follows approximately a normal distribution with mean $\frac{n_1(n_1+n_2+1)}{2}$ and variance $\frac{n_1 n_2(n_1+n_2+1)}{12}$, where n_1 is the sample size of group A and n_2 is the sample size of group B. Derive the approximate p-value in terms of the distribution function Φ of the standard normal distribution.

5. Let $X_1, \ldots, X_n$ be i.i.d. Geo(p) random variables with parameter $p \in (0, 1)$, that is

$$\mathbb{P}(X_i = k) = (1 - p)^{k-1}p$$

   for $k = 1, 2, 3, \ldots$

9. **(1P)** Explain the difference between an informative and non-informative prior.

10. **(1P)** What is the best Bayesian point estimator for the quadratic loss function?

11. **(1P)** True or false? An AR(1) process with parameter $\beta_1 = \frac{1}{2}$, that is a time series $\{X_t\}$ with

$$X_t = \frac{1}{2}X_{t-1} + W_t$$

with $W_t \sim \mathrm{WN}(0, \sigma^2)$, is weakly stationary.

Part B: Theory

1. Consider the linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \gamma \mathbf{1}\{z_i = 1\} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $\theta = (\beta_0, \beta_1, \gamma)$ is a vector of real parameters, x_i are the realization of a continuous covariate, and z_i are the realizations of a categorical covariate with two categories $z_i \in \{0, 1\}$. Let $\varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. standard normal random variables, that is $\varepsilon_i \sim \mathcal{N}(0, 1)$.

   (a) **(1P)** Identify the design matrix X.

   (b) **(1P)** Show that $\mathbf{Y} \sim \mathcal{N}(X\theta, I_n)$.

   (c) **(3P)** Derive the maximum likelihood estimator $\hat{\theta}$.

2. A company surveyed customers to investigate whether **customer type** (New vs. Returning) is associated with the **support channel** they used (Email, Chat, Phone). The observed counts are shown in the contingency table below:

	Email	Chat	Phone	Row Total
New customers	52	38	10	100
Returning customers	30	50	20	100
Column Total	82	88	30	200

   (a) **(1P)** Describe parametric models for the support channel $X \in \{\text{Email}, \text{Chat}, \text{Phone}\}$, customer type $Y \in \{\text{New}, \text{Returning}\}$ and the random vector (X, Y).

   (b) **(2P)** Formulate the null and alternative hypotheses for testing whether customer type and support channel are independent. Give the test statistic for the approximate χ^2-test for contingency tables.

(a) **(2P)** For a uniform prior, that is $\pi(p) = \mathbf{1}\{p \in (0,1)\}$, derive the posterior density $\pi(p|x_1, \ldots, x_n)$. **Hint:** You do not need to calculate the normalizing constant.

(b) **(3P)** Derive the maximum-a-posteriori estimator

$$\hat{p} = \arg\max_{p \in (0,1)} \pi(p|x_1, \ldots, x_n).$$

6. Let $\{Z_t\}_{t \in \mathbb{Z}}$ be an i.i.d. sequence of standard normal random variables.

Define the time series $\{X_t\}$ with

$$X_t = Z_t + 2Z_{t+1} + Z_{t+2}, \qquad t \in \mathbb{Z}.$$

(a) **(1P)** Argue that this is a Gaussian process.

(b) **(2P)** Calculate the covariance $\mathrm{Cov}(X_t, X_s)$ for arbitrary integers t and s.

(c) **(1P)** Derive the autocovariance function $\gamma_X(\tau)$ in terms of the lag $\tau = t - s$ and conclude that $\{X_t\}$ is weakly stationary.

(d) **(1P)** Can you conclude that $\{X_t\}$ is also strictly stationary? Carefully justify your answer.

7. Let $x_1, \ldots, x_n \in \mathbb{R}$ be observations of a (weakly) stationary time series, and let $\bar{x} := \frac{1}{n}\sum_{t=1}^{n} x_t$. For an integer lag τ with $-n < \tau < n$, define the sample autocovariance

$$\hat{\gamma}(\tau) := \frac{1}{n}\sum_{t=1}^{n-|\tau|} \left(x_{t+|\tau|} - \bar{x}\right)\left(x_t - \bar{x}\right).$$

(a) **(1P)** Show that $\hat{\gamma}(\tau) = \hat{\gamma}(-\tau)$ for all admissible τ.

(b) **(3P)** Consider the $n \times n$ matrix

$$\hat{\Gamma} := \left[\hat{\gamma}(i-j)\right]_{i,j=1}^{n}.$$

Show that $\hat{\Gamma}$ is positive semidefinite, i.e. for every $a \in \mathbb{R}^n$,

$$a^\top \hat{\Gamma}\, a \geq 0.$$