Statistics II - Test 1

Coursecode: 202300026
Quartile: 1B 2023/24
Time: 26 January 2024
Time: 8:45 - 11:45
Teacher: Sophie Langer

Student's name:_____
Student ID:_____

General information:

- A regular scientific calculator is allowed, a programmable calculator ("GR") is not.
- All tables of probability distribution needed are attached.
- Other than a pen, no means are needed (or allowed) for answering the exam questions.
- Please write your name on every exam paper you hand in.
- Please write legibly. I cannot evaluate what I do not understand.

Distribution of the points:

Part A

Task	(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)	Sum
	1	1	1	1	2	2	1	1	1	11

Part B

Task	1	2	3	4	5	6	7	8	9	Sum
	2	2+2+1+2 Bonus	1+1+2+2+2	7	+3 Bonus	8	+4 Bonus	2+4+4	4+4	48

Achieved points:

Task	Part A	1	2	3	4	5	6	7	Sum

Part A: Basic concepts

(a) [1 point] If X and Y are normally distributed, then (X, Y) is also multivariate normally distributed. True or false?
Solution: False

(b) [1 point] Give the formula for the multiple linear regression model?
Solution: Either $Y = X\beta + \varepsilon$ or $Y_i = \beta_0 + x_{i,1}\beta_1 + \ldots x_{i,p-1}\beta_{p-1} + \varepsilon_i$, for $i = 1, \ldots, n$

(c) [1 point] When do interaction effects occur in a multiple linear regression model?
Solution: When an independent variable has a different effect on the outcome depending on the value of another independent variable.

(d) [1 point] What is a dummy variable? Provide an example.
Solution: A categorical explanatory variable. Example is e.g. the diet variable in the dinosaur dataset.

(e) [2 points] Describe a hypothesis test that can be used to check normality of data. What is the null hypothesis in this test and why aren't the null and the alternative hypothesis interchanged.
Solution: Shapiro-Wilk's test on normality. The null hypothesis is that the data are normally distributed. In the classical hypothesis framework, we want that the 'effect' is in the alternative hypothesis. However, since the null hypothesis is used to derive the test, it would be extremely difficult to work with a null hypothesis stating that the data are generated from any distribution that are not the normal distribution.

(f) [2 point] What is the difference between parametric and non-parametric bootstrap?
Solution: In parametric bootstrap we assume that data are sampled from a parametric model P_θ with θ unknown. We estimate θ with $\hat{\theta}$ by using our data and then generate bootstrap samples from $P_{\hat{\theta}}$. In non-parametric bootstrap we sample with replacement from the given data.

(g) [1 point] Bayesian statistics: We estimate a parameter θ with a point estimator $\hat{\theta}$. If the loss function $L(\theta, \hat{\theta}) = (\hat{\theta} - \theta)^2$ then the posterior median is the best estimator. True or false?
Solution: False

(h) [1 point] Describe a random walk time series.
Solution: $S_t = \sum_{i=1}^{t} X_i$ with $\{X_i\}$ iid noise.

(i) [1 point] For Gaussian processes, weakly stationary is not equivalent to strictly stationary. True or false?
Solution: False.

Part B: Theory

1. [2 points] For $\mathbf{X} = (X_1, X_2, X_3)^T \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, find the distribution of

$$\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix}.$$

Solution: Note that

$$\begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix} = A \cdot \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \text{ with } A = \begin{pmatrix} 1 & -1 & 1 \\ 0 & 1 & -1 \end{pmatrix}$$

As $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ then $AX = \begin{pmatrix} X_1 - X_2 \\ X_2 - X_3 \end{pmatrix} \sim \mathcal{N}(A\boldsymbol{\mu}, A\Sigma A^T)$.

2. Consider the Gaussian multivariate regression model $Y = X\beta + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, assuming that $X^\top X$ is invertible. Recall that $\widehat{\beta}^{\mathrm{MLE}} = (X^\top X)^{-1} X^\top Y$ is the MLE. The residuals are given by $Y - X\widehat{\beta}^{\mathrm{MLE}}$.

   (a) [2 points] Show that the residuals are given by the formula

$$\big(I - X(X^\top X)^{-1} X^\top\big)\varepsilon.$$

   Solution: We have $Y - X\widehat{\beta}^{\mathrm{MLE}} = (I - X(X^\top X)^{-1}X^\top)Y = (I - X(X^\top X)^{-1}X^\top)(X\beta + \varepsilon = (I - X(X^\top X)^{-1}X^\top)\varepsilon$.

   (b) [2 points] Show that the distribution of the residuals is given by

$$\mathcal{N}\big(0, I - X(X^\top X)^{-1} X^\top\big).$$

   Solution: We have $\varepsilon \sim \mathcal{N}(0, I)$. From the lecture, we know that $P\varepsilon \sim \mathcal{N}(0, PP^\top)$. We apply this formula to $P = I - X(X^\top X)^{-1}X^\top$. Since in this case $P^2 = P$ and P symmetric, we have $P\varepsilon \sim \mathcal{N}(0, P)$. The distribution of the residuals is thus

$$\mathcal{N}\big(0, I - X(X^\top X)^{-1} X^\top\big).$$

(c) [1 point] Under the assumptions on linear regression, the residuals have consequently a normal distribution. In practice, the assumptions on linear regression are often not met and the residuals have a different distribution. To test whether the assumptions on the regression model are met, one could be tempted to apply Shapiro-Wilk's test on normality. Which assumption of Shapiro-Wilk's test on normality is violated?

Solution: Under the null hypothesis, the residuals are normal but not necessarily independent. This is also the problem for all the other tests that we discussed. Another reason is that the variances are also not all the same. Since the variances are known, this can, however, be easily fixed dividing the data by the corresponding standard deviations.

(d) [+2 Bonus] What could be done to make Shapiro-Wilk's test on normality applicable?

Solution: Because $P = I - X(X^\top X)^{-1}X^\top$ is symmetric, we can find an orthogonal matrix D and a diagonal matrix Λ such that $P = D\Lambda D^\top$. Since P is a projection, that is, $P^2 = P$, all eigenvalues of Λ are either zero or one. We can now multiply $D^\top$ and the residual vector. The distribution is then $\mathcal{N}(0, D^\top P D) = \mathcal{N}(0, \Lambda)$. Since Λ is diagonal and all eigenvalue are either zero or one, one can extract the components of the vector corresponding to eigenvalue one. Under the null hypothesis, these random variables are thus independent and follow a $\mathcal{N}(0, 1)$ distribution.

3. A group of researcher accompanies a group of penguins on a 10-years period and measures their survival time. One group of the penguins is tagged with a metal tag, the other group with an electronic tag. We are interested in testing whether the type of tag has an effect on the penguin survival rate, using a χ^2 test. In the study 33 of the 167 metal-tagged penguins survived while 68 of the 189 electronic tagged penguins survived.

(a) [1 point] Create a two-way table from the information given.

(b) [1 point] State the null and alternative hypothesis.

(c) [2 points] Give a table with the expected counts for each of the four categories.

(d) [2 points] Calculate the chi-square statistic.

(e) [2 points] Determine the p-value and state the conclusion.

Solution.

(a) The two-way table of penguin survival time vs. type of tag is

	Metal	Electronic	Total
Survived	33	68	101
Died	134	121	255
Total	167	189	356

(b) The hypothesis are

H_0 : Type of tag is not related to the survival time

H_1 : Type of tag is related to the survival time.

(c) The table below shows the expected counts obtained for each cell by multiplying the row total by the column total and dividing by $n = 356$.

	Metal	Electronic
Survived	47.4	53.6
Died	119.6	135.4

(d) We calculate the χ^2 test statistic

$$\begin{aligned}\chi^2 &= \frac{(33-47.4)^2}{47.4} + \frac{(68-53.6)^2}{53.6} + \frac{(134-119.6)^2}{119.6} + \frac{(121-135.4)^2}{135.4} \\ &= 4.37 + 3.87 + 1.73 + 1.53 \\ &= 11.5\end{aligned}$$

(d) We compare the test statistic to a chi square with 1 degree of freedom to get a p-value of 0.0007. There is a strong evidence that the type of tag and the survival rate are related.

4. [7 points] What do you test by running a permutation test? What assumption is needed? Describe the different steps of the test. Name *two* advantages of a permutation test.

Solution. In general, a permutation test checks whether two samples follow the same distribution. It underlies the assumption that the test statistic is **invariant under permutations**, i.e., under H_0 any permutation of the data has the same distribution. To run a permutation test, one needs to proceed in four steps: 1 Pt.

(a) Compute the observed values of the test statistic 1 Pt.

$$t_{obs} = T(X_1, \ldots, X_n, Y_1, \ldots, Y_n)$$

(b) Randomly permute the data and compute the statistic again using the permuted data

1 Pt.

(c) Repeat the previous step for all possible (or a large number) of permutations

(d) Approximate the p-value

1 Pt.

$$\hat{p}(Z_{obs}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(T(Z_i) \geq t_{obs})$$

(e) Evaluate with

1 Pt.

$$\phi_{perm}(Z_{obs}) = \mathbf{1}[\hat{p}(Z_{obs} < \alpha].$$

The advantages of a permutation test is that

(a) it is a distribution-free test

(b) it works for tiny samples

(c) works for categorical and numerical data

(d) probability of wrongly rejecting is limited to $\alpha = 0.05$

(e) the test statistic can be flexible chosen

1 Pt. for each advantage max. 2

5. [3 points] Wilcoxon's rank sum test is based on an analysis of the ranks. Given a sample of independent and identically distributed random variables $(Z_1, \ldots, Z_N)$, denote by $R(Z_1), \ldots, R(Z_N)$ the ranks of $Z_1, \ldots, Z_N$. Show that

$$\text{Cov}\left(R(Z_1), R(Z_2)\right) = -\frac{(N+1)}{12}.$$

Hint: You may use that $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$.

Solution: Write $R_i = R(Z_i)$. We have $E(R_i) = (N+1)/2$. Using that $P(R_1 = k, R_2 = \ell) = P(R_1 = k|R_2 = \ell)P(R_2 = \ell) = 1/[(N-1)N]$ if $k \neq \ell$ and $P(R_1 = k, R_2 = \ell) = 0$ for $k = \ell$, we find that

$$\begin{aligned} E\left(R_1 R_2\right) &= \sum_{k=1}^{N}\sum_{\ell=1}^{N} k\ell P(R_1 = k, R_2 = \ell) \\ &= \sum_{k=1}^{N}\sum_{\substack{k=1 \\ k\neq\ell}}^{N} k\ell \frac{1}{N(N-1)} \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{N(N-1)}\left(\sum_{k=1}^{N} k \sum_{\ell=1}^{N} \ell - \sum_{k=1}^{N} k^2\right) \\
&= \frac{1}{N(N-1)}\left(\frac{N(N+1)}{2}\frac{N(N+1)}{2} - \frac{N(N+1)(2N+1)}{6}\right) \\
&= \frac{N(N+1)}{N(N-1)}\left(\frac{N(N+1)}{4} - \frac{2N+1}{6}\right) \\
&= \frac{N+1}{N-1}\left(\frac{3N^2+3N-4N-2}{12}\right) \\
&= \frac{N+1}{N-1}\frac{(3N+2)(N-1)}{12} \\
&= \frac{(N+1)(3N+2)}{12}
\end{aligned}
$$

and

$$
\begin{aligned}
\operatorname{Cov}(R_1, R_2) = &\ E(R_1 R_2) - E(R_1)E(R_2) = \frac{1}{12}(N+1)(3N+2) - \left(\frac{N+1}{2}\right)^2 \\
&= -\frac{N+1}{12}.
\end{aligned}
$$

6. [8 points] Suppose that $X_1, X_2, \ldots, X_n \sim \text{Ber}(\theta)$ and that θ has a Beta(a, b) prior. Find the posterior mean of θ. Show that it is a weighted average of the sample mean and the prior mean. You may use that for $Z \sim \text{Beta}(a, b)$, $\mathbf{E}(Z) = \frac{a}{a+b}$.
Solution: We know that

$$
\pi(\theta|X_1, \ldots, X_n) \propto \pi(\theta) \cdot \prod_{i=1}^{n} p(x_i|\theta).
$$

1 Pt.

The likelihood function for a sequence of independent Bernoulli distributed random variables is given by

$$
L(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} p(x_i|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}.
$$

2 Pt.

By definition of the Beta distribution

$$
\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}(0 \leq \theta \leq 1).
$$

1 Pt.

This, in turn, means that

$$
\begin{aligned}
\pi(\theta|X_1, \ldots, X_n) &\propto \theta^{a-1}(1-\theta)^{b-1}\mathbf{1}(0 \leq \theta \leq 1) \cdot \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \\
&\propto \theta^{\sum_{i=1}^n x_i + a - 1} \cdot (1-\theta)^{n+b-1-\sum_{i=1}^n x_i}
\end{aligned}
$$

1 Pt.

$$\sim \text{Beta}\left(\sum_{i=1}^{n} x_i + a, n + b - \sum_{i=1}^{n} x_i\right).$$ -1 Pt.

By definition the posterior mean is

$$\begin{aligned}
\hat{\theta} &= \int_{\Theta} \theta \cdot \pi(\theta|X) d\theta = E(\theta|X_1, \ldots, X_n) \\
&= \frac{\sum_{i=1}^{n} x_i + a}{\sum_{i=1}^{n} x_i + a + n + b - \sum_{i=1}^{n} x_i} \\
&= \frac{\sum_{i=1}^{n} x_i + a}{a + n + b} \qquad \text{1 Pt.} \\
&= \frac{n}{a+n+b} \cdot \frac{1}{n} \sum_{i=1}^{n} x_i + \frac{a+b}{a+n+b} \cdot \frac{a}{a+b}, \qquad \text{1 Pt.}
\end{aligned}$$

showing that the posterior mean is the weighted average of the sample mean and the prior mean.

7. [4 points] Determine the Bayes estimator $\hat{\theta}$ for scaled squared error loss, $L(\theta, \hat{\theta}) = c(\theta - \hat{\theta})^2$ for some $c > 0$ and $\theta \in \mathbb{R}$.
Solution: We want to find the minimizer for $\mathbf{E}(L(\theta, \hat{\theta}))$. For that we take the gradient with respect to $\hat{\theta}$:

$$\begin{aligned}
\mathbf{E}(L(\theta, \hat{\theta})|x) &= \mathbf{E}(c\theta^2 - 2c\hat{\theta}\theta + c^2\hat{\theta}^2|x) \\
&= c\mathbf{E}(\theta^2|x) - 2c\hat{\theta}\mathbf{E}(\theta|x) + c^2\hat{\theta}^2 \\
\frac{d}{d\hat{\theta}}\mathbf{E}(L(\theta, \hat{\theta})|x) &= -2c\mathbf{E}(\theta|x) + 2c\hat{\theta} \\
\frac{d^2}{d\hat{\theta}}\mathbf{E}(L(\theta, \hat{\theta})|x) &= -2c.
\end{aligned}$$

Setting the first derivative to 0 results in $\hat{\theta} = \mathbf{E}(\theta|x)$ and the second derivative being positives means this estimate minimizes the scaled squared error loss.

8. Consider the time series

$$X_t = -2t + W_t + 0.5W_{t-1},$$

where $W_t \sim \mathcal{N}(0, \sigma^2)$.

   (a) [2 points] What does it mean for a time series to be *stationary*?

   (b) [4 points] What are the mean function and the autocovariance function of this time series? Is this time series stationary? Justify your answer.

(c) [4 points] Consider the first order differences of the time series above, that is, consider

$$Y_t = \nabla X_t = X_t - X_{t-1}.$$

What is the mean function and autocovariance function of this time series? Is this time series stationary? Justify your answer.

Solution.

(a) The mean of the series is constant and the autocovariance between Y_t and Y_s is a function only of $t - s$.

(b) The mean of this time series is $\mathbf{E}(X_t) = -2t$, which varies with t. Thus the series is not stationary. The autocovariance function is

1 Pt.
1 Pt.

$$\begin{aligned}\gamma(\tau) &= \mathrm{Cov}(-2t + W_t + 0.5W_{t-1}, -2(t+\tau) + W_{t+\tau} + 0.5W_{t+\tau-1}) \\ &= \mathbf{E}((2t + W_t + 0.5W_{t-1} + 2t) \cdot (-2(t+\tau) + W_{t+\tau} + 0.5W_{t+\tau-1} + 2(t+\tau)) \\ &= \mathbf{E}((W_t + 0.5W_{t-1})(W_{t+\tau} + 0.5W_{t+\tau-1}))\end{aligned}$$

For $\tau = 0$, it then follows

$$\gamma(0) = \mathbf{E}(W_t^2 + 0.25W_{t-1}^2) = 1.25\sigma^2$$

For $|\tau| = 1$, we have

$$\begin{aligned}\gamma(1) &= \mathbf{E}((W_t + 0.5W_{t-1}) \cdot (W_{t+1} + 0.5W_t)) \\ &= \mathbf{E}(0.5W_t^2) \\ &= 0.5\sigma^2.\end{aligned}$$

2 Pt.

For all other τ, $\gamma(\tau) = 0$.

(c) The differenced time series $\{Y_t\}$ is given by

$$Y_t = -2t + W_t + 0.5W_{t-1} - (-2(t-1) + W_{t-1} + 0.5W_{t-2}) = W_t - 0.5W_{t-1} - 0.5W_{t-2} - 2.$$

This has mean function $\mathbf{E}Y_t = 2$, which is constant. The autocovariance function is

1 Pt.

$$\begin{aligned}\gamma(\tau) &= \mathrm{Cov}(Y_t, Y_{t+\tau}) \\ &= \mathrm{Cov}(W_t - 0.5W_{t-1} - 0.5W_{t-2} - 2, W_{t+\tau} - 0.5W_{t+\tau-1} - 0.5W_{t+\tau-2} - 2)\end{aligned}$$

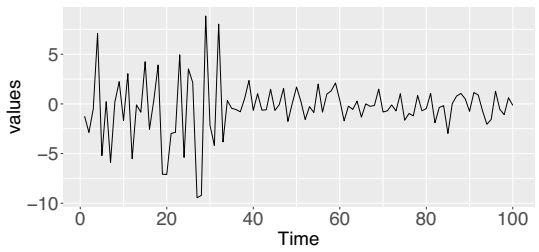
$$
\begin{aligned}
&= \operatorname{Cov}(W_t, W_{t+\tau}) - 0.5\operatorname{Cov}(W_t, W_{t+\tau-1}) - 0.5\operatorname{Cov}(W_t, W_{t+\tau-2}) \\
&\quad - 0.5\operatorname{Cov}(W_{t-1}, W_{t+\tau}) + 0.25\operatorname{Cov}(W_{t-1}, W_{t+\tau-1}) + 0.25\operatorname{Cov}(W_{t-2}, W_{t+\tau-2}) \\
&\quad - 0.5\operatorname{Cov}(W_{t-2}, W_{t+\tau}) + 0.25\operatorname{Cov}(W_{t-2}, W_{t+\tau-1}) + 0.25\operatorname{Cov}(W_{t-2}, W_{t+\tau-2}) \\
&= \begin{cases} 1.25\sigma^2 & \text{if } \tau = 0, \\ -0.25\sigma^2 & \text{if } |\tau| = 1 \\ -0.5\sigma^2 & \text{if } |\tau| = 2, \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

2Pt.

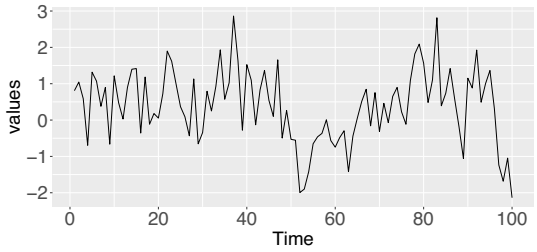
Since this function is constant and the autocovariance function depends on τ, the series is stationary.

1Pt.

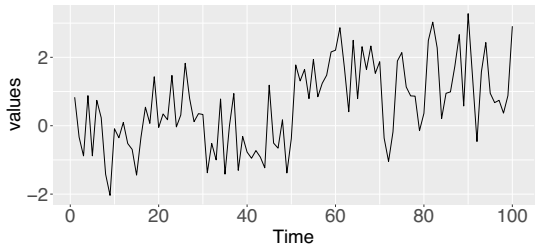
9. (a) [4 points] Consider the time series plot (i) - (iv).



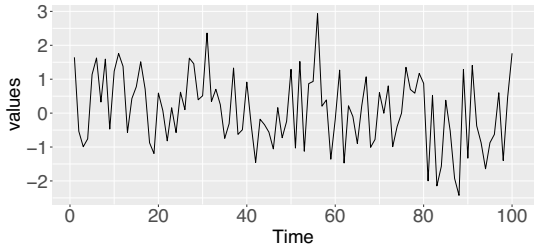
(i)



(ii)



(iii)



(iv)

Each of the time series is a realisation of one of the following stochastic processes:

(1) A Gaussian white noise process.

(2) An AR(1) process, i.e., $X_t = \phi X_{t-1} + \epsilon_t$, with parameter $\phi = 0.7$ and white noise process ϵ_t, $t = 1, \ldots, N$.

(3) A stochastic process X_t, $t = 1, \ldots, N$, (for some $N \in \mathbb{N}$) defined by

$$
X_t = \begin{cases} \epsilon_t, & t = 1, \ldots, \lfloor \frac{N}{2} \rfloor \\ \epsilon_t + \mu, & t = \lfloor \frac{N}{2} \rfloor + 1, \ldots, N \end{cases}
$$

for some $\mu > 0$ and a white noise process ϵ_t, $t = 1, \ldots, N$. (For a real number, x, $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x.)

(4) A stochastic process, X_t, $t = 1, \ldots, N$, (for some $N \in \mathbb{N}$) defined by

$$X_t = \begin{cases} \sigma\epsilon_t, & t = 1, \ldots, \lfloor \frac{N}{3} \rfloor \\ \epsilon_t, & t = \lfloor \frac{N}{3} \rfloor + 1, \ldots, N \end{cases}$$

for some $\sigma > 1$ and a white noise process ϵ_t, $t = 1, \ldots, N$.

(b) [4 points] Which of the time series (i) - (iv) can be considered as stationary? For each plot, justify your answer.

Solution. (a) (1)=(iv), (2)=(ii), (3)=(iii), (4)=(i) 1Pt. each

(b) (iii) is non-stationary, one possible justification: expectation is not constant over time, (i) is non-stationary, one possible justification: variance is not constant over time, (ii) and (iv) are stationary. } 1Pt. each