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	8	4	2+4+4	4+4	55

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?

(b) [1 point] Give the formula for the multiple linear regression model?

(c) [1 point] When do interaction effects occur in a multiple linear regression model?

(d) [1 point] What is a dummy variable? Provide an example.

(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.

(f) [2 point] What is the difference between parametric and non-parametric bootstrap?

(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?

(h) [1 point] Describe a random walk time series.

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

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

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

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

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

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

   (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?

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

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.

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.

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

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

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

6. [8 points] Suppose that $X_1, X_2, \ldots, X_n \sim \mathrm{Ber}(\theta)$ and that θ has a $\mathrm{Beta}(a, b)$ 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 $Z \sim \mathrm{Beta}(a, b)$, $\mathbf{E}(Z) = \frac{a}{a+b}$.

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 $\theta \in \mathbb{R}$.

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*?

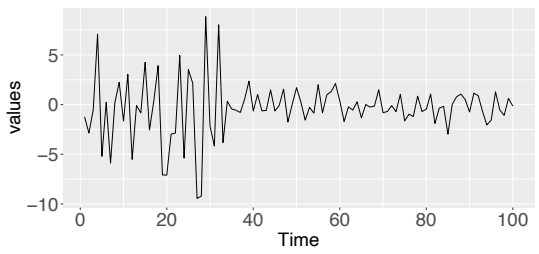
(b) [4 points] What are the mean function and the autocovariance function of this time series? Is this time series stationary? Justify your answer.

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

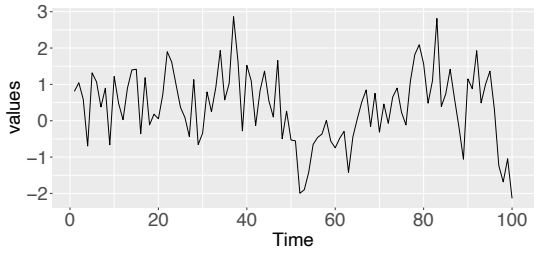
$$Y_t = \nabla X_t = X_t - X_{t-1}.$$

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

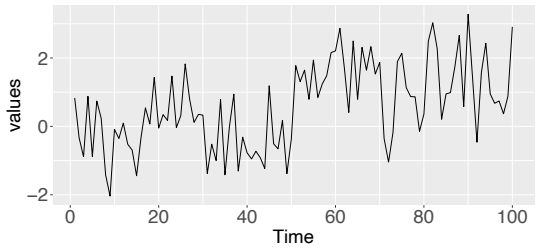
9. (a) [4 points] Consider the time series plot (i) - (iv).



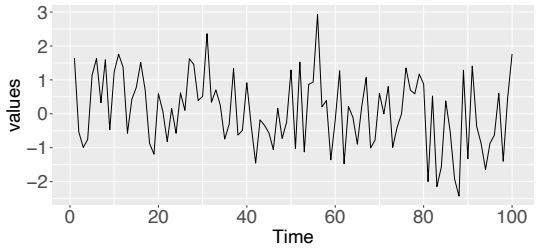
(i)



(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.

ν	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.25}$	$\chi^2_{.50}$	$\chi^2_{.75}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$	$\chi^2_{.999}$
1	.0000	.0002	.0010	.0039	.0158	.102	.455	1.32	2.71	3.84	5.02	6.63	7.88	10.8
2	.0100	.0201	.0506	.103	.211	.575	1.39	2.77	4.61	5.99	7.38	9.21	10.6	13.8
3	.0717	.115	.216	.352	.584	1.21	2.37	4.11	6.25	7.81	9.35	11.3	12.8	16.3
4	.207	.297	.484	.711	1.06	1.92	3.36	5.39	7.78	9.49	11.1	13.3	14.9	18.5
5	.412	.554	.831	1.15	1.61	2.67	4.35	6.63	9.24	11.1	12.8	15.1	16.7	20.5
6	.676	.872	1.24	1.64	2.20	3.45	5.35	7.84	10.6	12.6	14.4	16.8	18.5	22.5
7	.989	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.0	14.1	16.0	18.5	20.3	24.3
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.2	13.4	15.5	17.5	20.1	22.0	26.1
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.4	14.7	16.9	19.0	21.7	23.6	27.9
10	2.16	2.56	3.25	3.94	4.87	6.74	9.34	12.5	16.0	18.3	20.5	23.2	25.2	29.6
11	2.60	3.05	3.82	4.57	5.58	7.58	10.3	13.7	17.3	19.7	21.9	24.7	26.8	31.3
12	3.07	3.57	4.40	5.23	6.30	8.44	11.3	14.8	18.5	21.0	23.3	26.2	28.3	32.9
13	3.57	4.11	5.01	5.89	7.04	9.30	12.3	16.0	19.8	22.4	24.7	27.7	29.8	34.5
14	4.07	4.66	5.63	6.57	7.79	10.2	13.3	17.1	21.1	23.7	26.1	29.1	31.3	36.1
15	4.60	5.23	6.26	7.26	8.55	11.0	14.3	18.2	22.3	25.0	27.5	30.6	32.8	37.7
16	5.14	5.81	6.91	7.96	9.31	11.9	15.3	19.4	23.5	26.3	28.8	32.0	34.3	39.3
17	5.70	6.41	7.56	8.67	10.1	12.8	16.3	20.5	24.8	27.6	30.2	33.4	35.7	40.8
18	6.26	7.01	8.23	9.39	10.9	13.7	17.3	21.6	26.0	28.9	31.5	34.8	37.2	42.3
19	6.84	7.63	8.91	10.1	11.7	14.6	18.3	22.7	27.2	30.1	32.9	36.2	38.6	43.8
20	7.43	8.26	9.59	10.9	12.4	15.5	19.3	23.8	28.4	31.4	34.2	37.6	40.0	45.3
21	8.03	8.90	10.3	11.6	13.2	16.3	20.3	24.9	29.6	32.7	35.5	38.9	41.4	46.8
22	8.64	9.54	11.0	12.3	14.0	17.2	21.3	26.0	30.8	33.9	36.8	40.3	42.8	48.3
23	9.26	10.2	11.7	13.1	14.8	18.1	22.3	27.1	32.0	35.2	38.1	41.6	44.2	49.7
24	9.89	10.9	12.4	13.8	15.7	19.0	23.3	28.2	33.2	36.4	39.4	43.0	45.6	51.2
25	10.5	11.5	13.1	14.6	16.5	19.9	24.3	29.3	34.4	37.7	40.6	44.3	46.9	52.6
26	11.2	12.2	13.8	15.4	17.3	20.8	25.3	30.4	35.6	38.9	41.9	45.6	48.3	54.1
27	11.8	12.9	14.6	16.2	18.1	21.7	26.3	31.5	36.7	40.1	43.2	47.0	49.6	55.5
28	12.5	13.6	15.3	16.9	18.9	22.7	27.3	32.6	37.9	41.3	44.5	48.3	51.0	56.9
29	13.1	14.3	16.0	17.7	19.8	23.6	28.3	33.7	39.1	42.6	45.7	49.6	52.3	58.3
30	13.8	15.0	16.8	18.5	20.6	24.5	29.3	34.8	40.3	43.8	47.0	50.9	53.7	59.7
40	20.7	22.2	24.4	26.5	29.1	33.7	39.3	45.6	51.8	55.8	59.3	63.7	66.8	73.4
50	28.0	29.7	32.4	34.8	37.7	42.9	49.3	56.3	63.2	67.5	71.4	76.2	79.5	86.7
60	35.5	37.5	40.5	43.2	46.5	52.3	59.3	67.0	74.4	79.1	83.3	88.4	92.0	99.6
70	43.3	45.4	48.8	51.7	55.3	61.7	69.3	77.6	85.5	90.5	95.0	100	104	112
80	51.2	53.5	57.2	60.4	64.3	71.1	79.3	88.1	96.6	102	107	112	116	125
90	59.2	61.8	65.6	69.1	73.3	80.6	89.3	98.6	108	113	118	124	128	137
100	67.3	70.1	74.2	77.9	82.4	90.1	99.3	109	118	124	130	136	140	149

Source: E. S. Pearson and H. O. Hartley, *Biometrika Tables for Statisticians*, Vol. 1 (1966), Table 8, pages 137 and 138, by permission.