UNIVERSITY OF OSLO

Faculty of Mathematics and Natural Sciences

Trial exam in GEO4300/9300 – Geophysical Data Science

This paper consists of 4 pages including this page.

1 Random variables

(a) Explain or define the mean, median and mode of a random variable.

(b) Explain or define a measure of dispersion of a random variable.

(c) For bivariate random variables, explain how the Pearson correlation coefficient and the Spearman rank correlation coefficients are calculated and explain the differences between them. You may draw a simple sketch to illustrate the differences.

2 Hypothesis testing

Based on two samples with respectively 1000 and 100 observations, the following estimates for mean and standard deviations have been obtained:
Sample 1 ($n_1 = 1000$): $\bar{x}_1 = 44.1$ $\sigma_1 = 11.3$
Sample 2 ($n_2 = 100$): $\bar{x}_2 = 48.0$ $\sigma_2 = 8.2$

(a) Test if the estimates for mean and standard deviation are significantly different. Use a significance level of 5%.

(b) We make two types of errors in hypothesis testing. Explain these type I and type II errors, and illustrate graphically the relation between them.

3 Goodness-of-fit testing

The table below shows July precipitation data (P in inches) for $N = 30$ years of observations between 1951 and 1980 in Ithaca, New York.

TABLE 4.8 July precipitation at Ithaca, New York, 1951–1980 (inches).

1951	4.17	1961	4.24	1971	4.25
1952	5.61	1962	1.18	1972	3.66
1953	3.88	1963	3.17	1973	2.12
1954	1.55	1964	4.72	1974	1.24
1955	2.30	1965	2.17	1975	3.64
1956	5.58	1966	2.17	1976	8.44
1957	5.58	1967	3.94	1977	5.20
1958	5.14	1968	0.95	1978	2.33
1959	4.52	1969	1.48	1979	2.18
1960	1.53	1970	5.68	1980	3.43

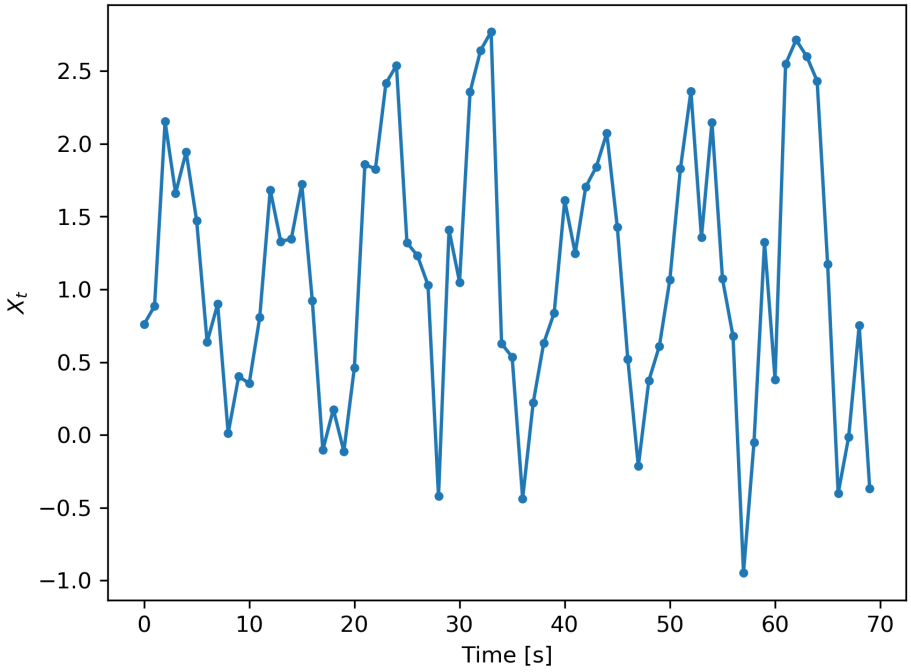
The mean value of the data is $\bar{P} = 3.54$ inches and the standard deviation $s = 1.77$ inches.

(a) Bin the data in suitable classes and sketch the cumulative histogram of the dataset.

(b) Use the Chi-square method to test the hypothesis that the data fit the normal distribution. Use a significance level of 5%.

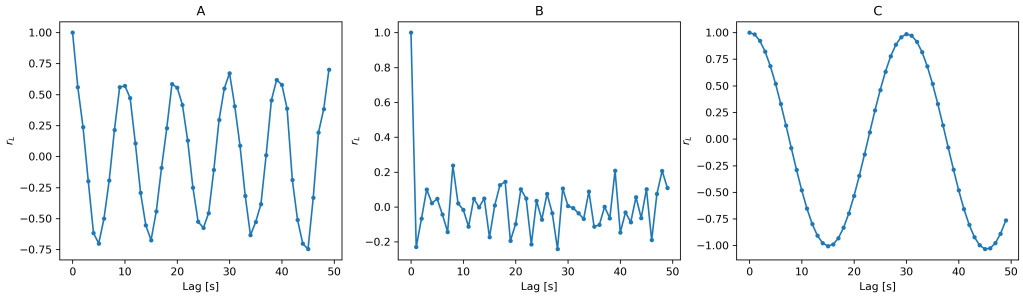
4 Time series analysis

Consider the following time series X_t sampled at $n = 70$ time steps:



(a) The following three graphs show autocorrelation functions for lags $L \leq 50$, defined as:

$$r_L = \frac{1}{n-L}\sum_{t=1}^{n-L}(X_t - \bar{X})(X_{t+L} - \bar{X}) / \frac{1}{n}\sum_{t=1}^{n}(X_t - \bar{X})^2$$



Which one of them (A, B or C) shows the autocorrelation of X_t? Explain your answer.

5 Machine learning

(a) What is the difference between supervised and unsupervised machine learning? Give an example of one model/learning algorithm and argue why you can categorize it as supervised (or unsupervised).

(b) Explain the difference between the two main types of supervised machine learning: regression and classification. Give an example of each type.