Exam guidelines GEO4300 fall 2019

This document lists the key elements of the expected answers of the written exam (i.e. these do not represent the complete/full answers). The grade achieved in the written exam was combined with the hand-ins to produce the final grade.

Question 1

a) Mean

$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x)dx \qquad \text{for continuous variables}$$

$$E(x) = \sum_{j=1}^{n} x_j \cdot f(x_j) \qquad \text{for discrete variable}$$

Median: F(x_median)=0.5
Mode: max(f(x))

b) E.g. variance

$$V(x) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx \quad \text{for continuous variable}$$

$$V(x) = \sum_{j=1}^{n} (x_j-\mu)^2 \cdot f(x_j) \quad \text{for discrete variable}$$

c) Pearson correlation: Covariance divided by the product of the standard deviations
Spearman correlation: Calculate the rank of the data. Calculate the correlation between the ranks

Difference: Pearson: Need a linear relationship to get a correlation of 1.
Spearman: Can get correlation of 1 also for non-linear relationships

Question 2

a)

Test for the difference of mean,

Ho: mu1=mu2 Ha: not equal

Two sample, two-tail
Z=(mu1-mu2)/sqrt(var1/n1+var2/n2)=-4.36
Reject H0 if |z|>1.96

Therefore Ho is rejected, i.e. the means are different at the significance level of 5%.

Test different standard deviations. We can equally test if the variances are different.

H0: var1=var2 Ha: not equal

Fc=var1/var2=1.9

$F_{1-a,n1-1,n2-1} = F_{0.95,999,99} = 1.30 < F_c$
Therefore, Ho is rejected, i.e. there is a significant change in the standard deviations.

b)
Type I error: reject a true null hypothesis. Depends on alpha value.
Type II error: don't reject a false null hypothesis. Depends on the power of the test, i.e. the sample size, alpha, test specifics…

Question 3

a) Use around 4 to 6 classes. Cumulative histogram increases monotonically until 30 (or 1 if normalized).
b) H0: data normally distributed. Find the expected number of observation for each class assuming a normal distribution. The calculations below would differ if different bins are used:

Class	Observed	Relative frequency	Expected	(obs-ex)^2/ex
P<2	6	0.19	5.7	0.016
2<P<3	6	0.38-0.19	5.7	0.016
3<P<4	6	0.60-0.38	6.6	0.054
4<P<5	5	0.80-0.60	6.0	0.166
P>=5	7	1.00-0.80	6.0	0.166
Total	30	1	30	**0.42**

Chi2=5.99 with alpha=0.05 and ndof=k-p-1=5-2-1=2

Therefore, don't reject H0.

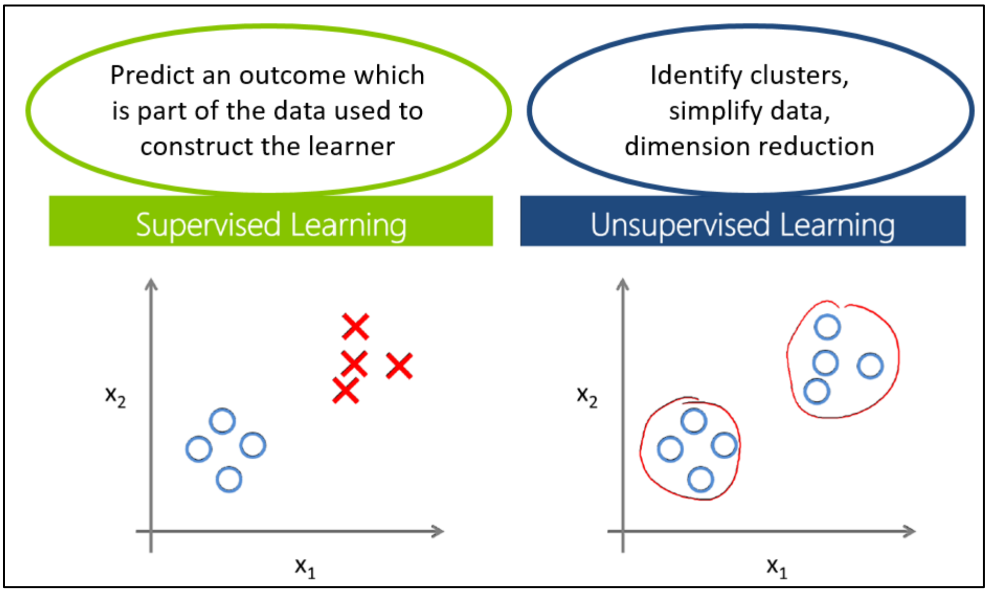
Question 4

a) A, because B doesn't have periodicity and C doesn't have any noise
b) Ordinary (least-squares) linear regression: significant slope? You'd have to check the normality of the resulting residuals before your decision.
   Mann-Kendall test: statistic based on sum of signs of differences

   Maybe also the runs test (even though it would detect the periodicity here, rather than a trend): statistic based on expected number of runs
   Also not ideal, but somehow possible would be a jump test: divide the time series in two and check if the means differ significantly

Question 5

a) See slide below. Example: linear regression is a supervised model because you predict an outcome that is part of the data used to fit the regression coefficients.



b) See slide below. Example: input is temperature, precipitation, altitude, vegetation type, latitude…. A regression could have this input and have output, Y=annual flood magnitude. A classification could have the same input, but have output, Y=whether there is permafrost or not.

- Y is called outcome, dependent variable, response, target, output…
- Two types of supervised ML:
  - Regression (if Y is quantitative, e.g. temperature)
  - Classification (if Y is categorical, e.g. vegetation type)
  - (Many ML methods can be used for both cases)