TMA4245 Statistics
Exam May 19 2014

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1 Collectable cards

Agnes collects cards of the set *Animals of the World*. A complete set consists of θ different types of cards. On each card, there is a picture of an animal species and information about the species. In addition, one of the numbers 1, 2, ..., θ is printed on the card – this number identifies the type of the card.

Let X be the type of a card that is bought in the shop. We assume that $P(X = x) = 1/\theta$ for $x = 1, 2, \ldots, \theta$ and $P(X = x) = 0$ for all other x. That is, the probability is the same of getting each type of card. We also assume that, when we buy several cards, the card types are independent.

a) Assume (only here) that there are 50 types of cards, that is $\theta = 50$.

Agnes buys 2 cards. What is the probability that the cards are of different type?

What is the probability that all cards are of different type if Agnes buys 8 cards?

The card manufacturer advertises that there are 200 types of cards in the set. Agnes has bought 20 cards, but she has never got any higher type than 170. Assume that $X_1, X_2, \ldots, X_n$ are independent types, and let $\max X_i$ be the largest of these types.

b) Find the cumulative distribution function $P(X_i \leq x)$ for $x = 1, 2, \ldots, \theta$.

Show that $P(\max X_i \leq x) = (x/\theta)^n$ for $x = 1, 2, \ldots, \theta$.

What is $P(\max X_i \leq 170)$ if $n = 20$ and there are $\theta = 200$ different types in a complete set?

Assume that θ is unknown. The types of Agnes' cards are 7, 8, 25, 32, 55, 72, 74, 74, 89, 100, 102, 114, 121, 124, 126, 129, 131, 151, 165 and 170.

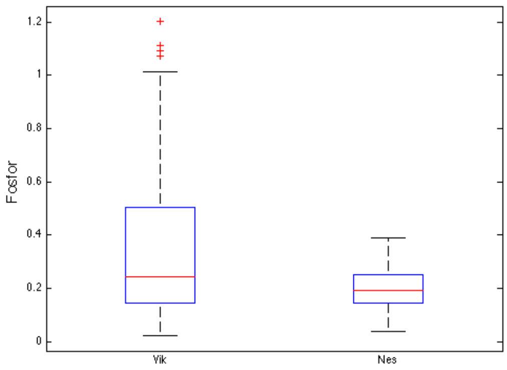
Agnes wishes to test the null hypothesis $\theta = 200$ against the alternative $\theta < 200$, and uses $\max X_i$, that is, the highest card type, as test statistic. She finds a critical region given by $\max x_i \leq 172$.

c) What is the conclusion of the hypothesis test with Agnes' data?

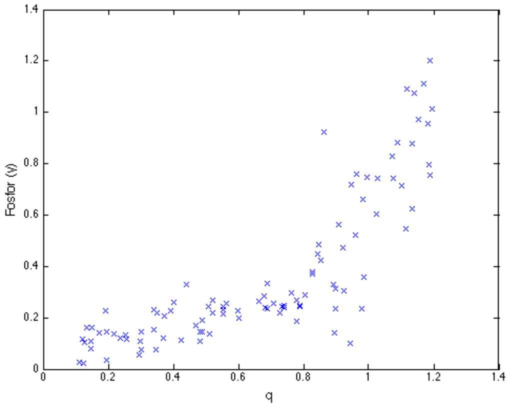
Find the significance level of the test.

Find the test power in $\theta = 180$ and in $\theta = 160$.

d) Show that the maximum likelihood estimate of θ is 170 with Agnes' data.



Figur 1: Box plots from two plants



Figur 2: 100 observations of phosphorous content and flow rate

Is the maximum likelihood estimator unbiased? Argue for your conclusion (you are not required to calculate the estimator's expected value).

Oppgave 2 Phosphorous at cleaning plant

We are interested in phosphorous content (in grams per cubic metre) in cleaned water at a cleaning plant.

a) Figure 1 shows box plots of measurements of phosphorous content from two plants, Vik and Nes. On the basis of the box plots, discuss whether the phosphorous content might come from normal distributions, and whether phosphorous content from the two plants have equal medians, equal expected values and equal variances. Argue briefly for your conclusions.

b) We assume (only here) that the phosphorous content Y of a sample is normally distributed with expected value $\mu = 0.3$ and variance $\sigma^2 = 0.1^2$.

Find the probability that Y is less than 0.5.

Find the probability that Y is greater than 0.3.

Find the conditional probability that Y is less than 0.5 given that Y is greater than 0.3.

A reason that the phosphorous content varies might be that it depends on the flow rate at the plant. Let q_i be the flow rate (in cubic metres per second) where sample no. i was taken and Y_i the phosphorous content in sample no. i. We assume a simple linear regression model

$$Y_i = \alpha + \beta q_i + \epsilon_i,$$

where α and β are regression parameters. Further, we assume that the error terms ϵ_i are independent and normally distributed with expected value 0 and variance σ_ϵ^2.

c) We assume (only here) that the regression parameters are known: $\alpha = 0.05$, $\beta = 0.3$ and $\sigma_\epsilon^2 = 0.05^2$.

Show that the phosphorous content in a sample at flow rate 0.5 is normally distributed with expected value 0.2 and variance 0.05^2, and that the phosphorous content in a sample at flow rate 1.0 is normally distributed with expected value 0.35 and variance 0.05^2.

What is the probability that the largest of three independent phosphorous measurements at flow rate 1.0 is greater than 0.4?

What is the probability that a measurement at flow rate 0.5 is greater than an (independent) measurement at flow rate 1.0?

Figure 2 shows $n = 100$ observations of phosphorous content and flow rate. Now we wish to estimate α and β by the method of least squares based on these data.

d) Briefly explain what the method of least squares is and illustrate by a figure.

Write down the expressions you need and explain the procedure. You are not required to derive the expressions for the estimators.

We now assume that the variance σ_ϵ^2 is known. Let $\hat{Y}_0$ be the prediction of phosphorous content given by the fitted (estimated) regression model at flow rate q_0. It is given that $\hat{Y}_0$ is normally distributed with expected value $\alpha + \beta q_0$ and variance

$$\sigma_\epsilon^2 \left(\frac{1}{n} + \frac{(q_0 - \bar{q})^2}{\sum_{i=1}^n (q_i - \bar{q})^2} \right).$$

e) Derive a 95% prediction interval for a new (independent) observation of phosphorous content when the flow rate is q_0.

Briefly explain the difference between a confidence interval and a prediction interval.

In Figure 3 data are plotted together with the fitted (estimated) regression line and the bounds of an interval. Is this a 95% prediction interval or a 95% confidence interval? Argue for your conclusion.

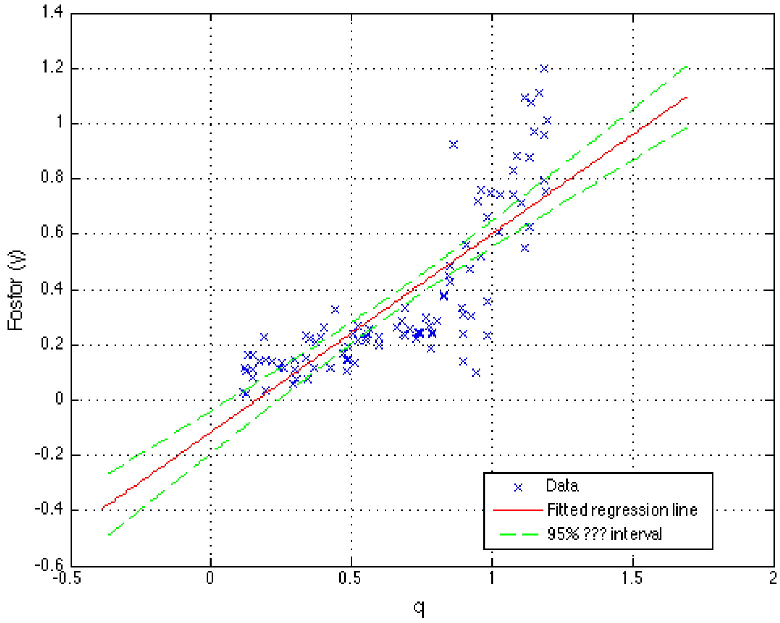
f) Specify the assumptions made in the regression model.

Discuss by means of Figures 2, 3 and 4 whether these assumptions are satisfied.

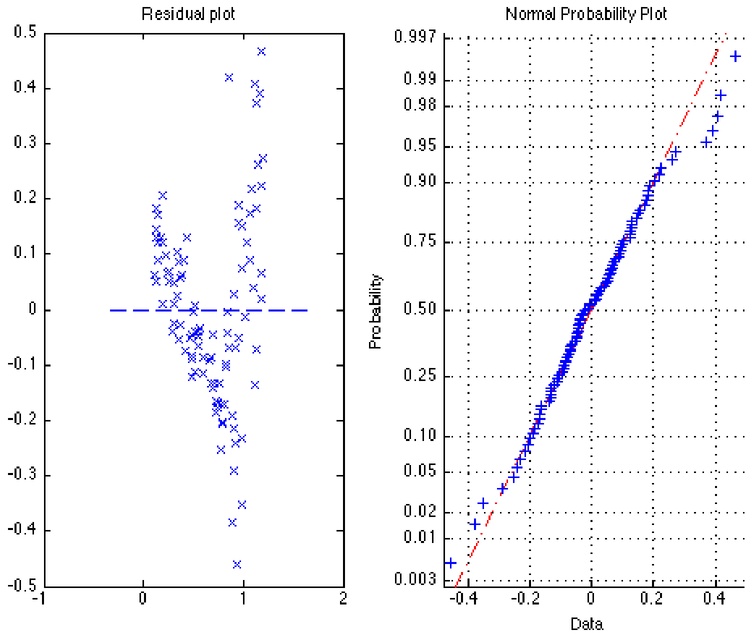
Fasit

1. **a**) 49/50,0.554 **b**) 0.0388 **c**) reject H_0,0.049,1 **d**) biased

2. **b**) 0.9772,0.5,0.9545 **c**) 0.4044,0.0169



Figur 3: Estimated regression line together with bounds of interval



Figur 4: Left: Residual plot (difference between data and estimated regression line along y axis, flow rate along x axis). Right: Normal probability plot (normal quantile–quantile plot, QQ plot) of residuals (differences between data and estimated regression line