TMA4245 Statistics
Exam December 20 2012

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

In this problem, we consider the situation where a coin is tossed several times. The coin is fair, i.e., the probability of the outcome “heads” is 0.5, as is the probability of the outcome “tails”. We assume that the outcomes of the coin tosses are independent of one another.

a) Suppose we toss the coin five times.

What is the probability of getting 5 “heads”?

What is the probability of getting 3 “heads”?

What is the probability of getting at least 4 “heads” in a row, i.e., a streak of “heads” outcomes that is of length at least 4?

b) We toss the coin 30 times. The distribution of the longest streak of “heads” is hard to calculate. Instead, we can make a computer generate 30 random and independent coin tosses. We record the longest streak of “heads”. This procedure is repeated B times, and the results are representative of the distribution of the longest streak of “heads”.

The figure shows a histogram of such streaks. These are the result of $B = 10\,000$ simulations of 30 coin tosses. The height of the bars indicate in how many of the 10 000 trials the longest streak of “heads” was of a particular length.

Estimate the probability of getting a longest streak of 5 or 6.

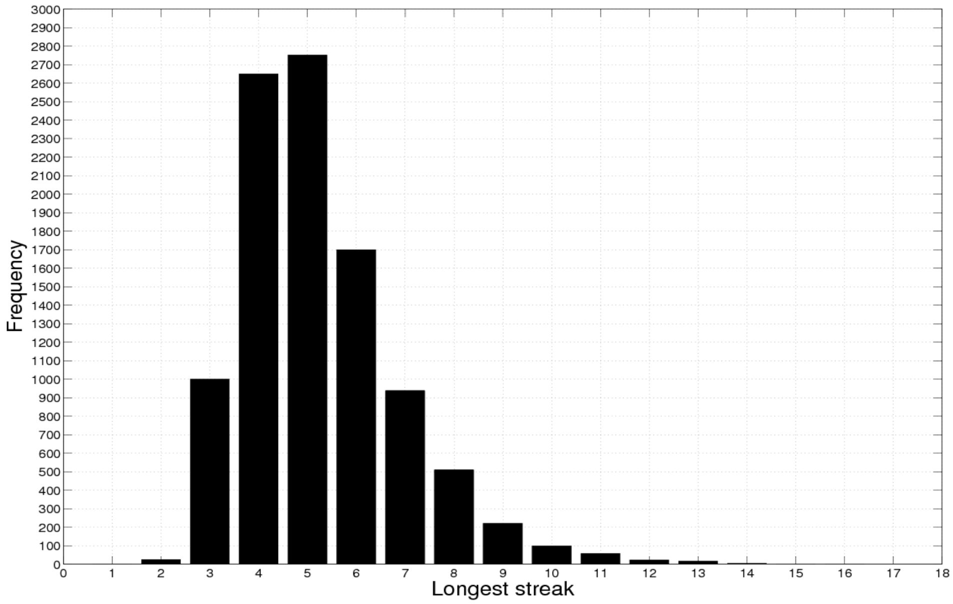
Miriam has been given a homework assignment of tossing a coin 30 times and recording the results. The results she handed in are as follows, where 1 means “heads” and 0 means “tails”:
$(0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 1, 0, 1, 0, 1, 0, 1, 0, 0, 1)$.

The teacher suspects that Miriam has cheated and just made up the numbers instead of actually tossing a coin, and the teacher wants to investigate this. Formulate this as a hypothesis test about the longest streak of “heads”. Use the histogram in the figure to answer.

Oppgave 2

Kjell Inge often goes fishing for an hour or so near where he lives.

a) He believes that the weight of a fish is normally distributed with mean value 800 grams and standard deviation 100 grams. Assume Kjell Inge is right about this.



What is the probability of a fish weighing more than 1000 grams?

What is the probability of a fish weighing between 500 grams and 1000 grams?

Kjell Inge thinks that the number of fish is more important than the weight. He wonders about the probability distribution of the number of fish per trip.

b) Let X be the number of fish caught on one trip. All trips take 1 hour. We assume that X is Poisson distributed with parameter (mean) $\mu = 3$.

What is the probability that Kjell Inge gets no fish on a trip?

Given that he gets fish, what is the probability that he gets more than 3 fish?

An alternative probability model is as follows: With probability θ he surely gets 0 fish. With probability $1 - \theta$ the number of fish that he gets is Poisson distributed with parameter μ. Then the point probability of the number of fish caught is:

$$P(X = x) = \theta I(X = 0) + (1 - \theta)\frac{\mu^x}{x!}\mathrm{e}^{-\mu}, \qquad x = 0, 1, 2, \ldots$$

where $I(A) = 1$ if the event A occurs, and $I(A) = 0$ otherwise.

c) Assume that $\theta = 0.5$ and $\mu = 4$.

What is the probability that Kjell Inge gets one or more fish on a trip?

Use the probability model above to calculate the expected number of fish caught per trip.

We now assume that we have data from $n = 20$ independent fishing trips. Of these twenty trips, $r = 8$ ended with no fish. In the remaining twelve trips, the total number of fish caught was 40. We use this data to estimate the model parameters μ and θ.

d) State the likelihood function for μ and θ.

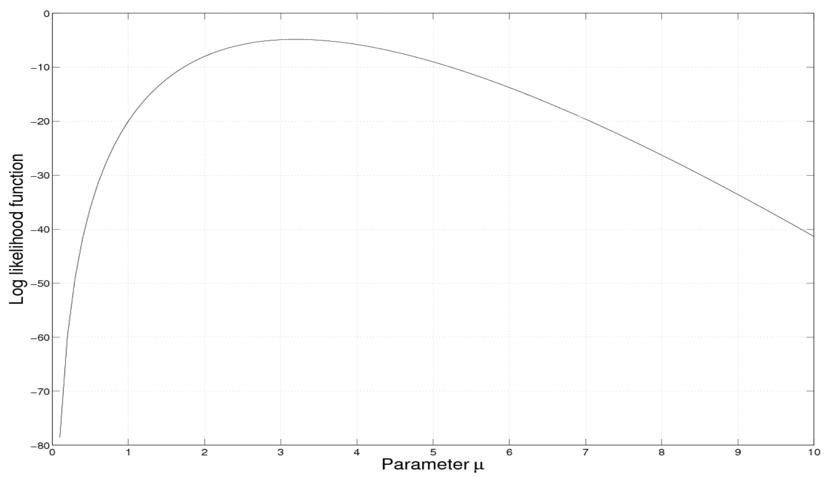
Assume that μ is known. Show that the maximum likelihood estimator for θ is given by

$$\hat{\theta} = \hat{\theta}(\mu) = \frac{r - n\mathrm{e}^{-\mu}}{n(1 - \mathrm{e}^{-\mu})}.$$

By inserting $\hat{\theta}(\mu)$ into the likelihood function, we can study the likelihood function only as a function of μ. The plot in the figure below shows this likelihood function.

Use the plot to find the maximum likelihood estimate for μ.

Use the result to calculate the maximum likelihood estimate of θ.



Oppgave 3

The figure shows the winning times in the men's 800-meter run in all the Olympic Games.

There is a total of $n = 28$ winning times. We let Y_i be the winning time in Olympics number i, and x_i be the year of Olympics number i. We assume the following regression model for the winning times:

$$Y_i = \alpha + \beta x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$

In addition, the noise terms $\epsilon_1, \ldots, \epsilon_n$ are assumed to be independent.

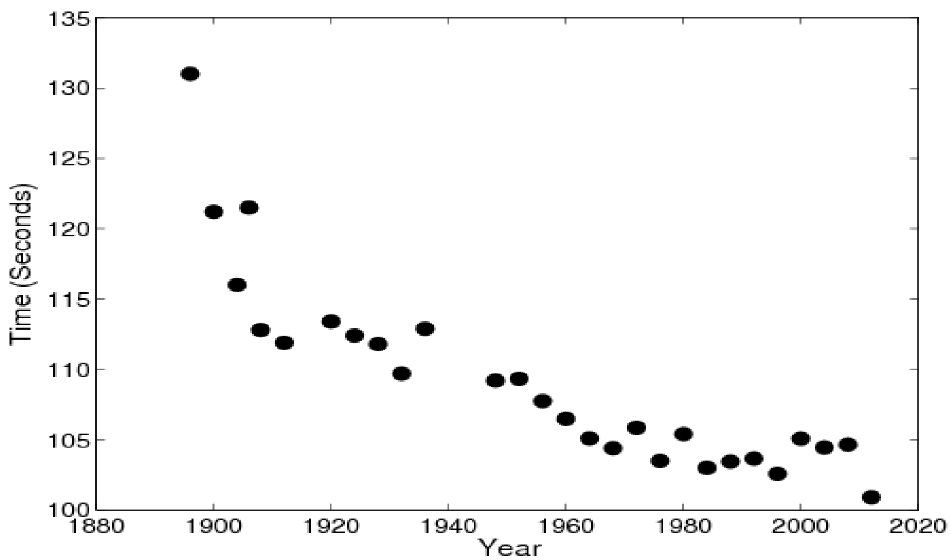
a) Give a brief explanation of the method of least squares for line fitting.

Show that this method gives the following estimators:

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}, \qquad \hat{\beta} = \frac{\sum_{i=1}^n x_i Y_i - n\bar{x}\bar{Y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

where the sample means (averages) are $\bar{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

An alternative expression for the estimator of the slope is $\hat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$.



It is given that $\bar{Y} = 109.26$, $\bar{x} = 1954.5$, $\sum_{i=1}^{n}(x_i - \bar{x})Y_i = -5942$ and $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 36517$. An estimate of the variance of the noise terms is $s^2 = 3.40^2$.

b) It can be shown that

$$T = \frac{(\hat{\beta} - \beta)}{s/\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \sim t_{n-2}.$$

Use this result to derive a 95 percent confidence interval for β.

Calculate the confidence interval using the numbers provided above.

We want to predict the winning time at the next Summer Olympics: 2016 in Brazil.

c) Calculate the predicted winning time in 2016.

Find a 95 percent prediction interval for the winning time in 2016.

d) Use the model to estimate the year when the 90-second limit will be broken, i.e., the first year in which the winning time will be below 90 seconds.

Discuss the model assumptions that have been made. What methods can be used to investigate the correctness of these assumptions?

Fasit

1. **a**) 0.031,0.313,0.09 **b**) 0.44, reject H_0

2. **a**) 0.023,0.976 **b**) 0.05,0.37 **c**) 0.49,2 **d**) 0.37

3. **b**) $[-0.199, -0.126]$ **c**) 99.3,$[91.8, 106.7]$ **d**) 2073