TMA4245 Statistics
Exam December 9 2013

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

In the card game blackjack, one gets the highest win if the first two cards you get dealt is an ace along with either a ten, a jack, a queen or a king. This is called "getting a blackjack".

In casinos where blackjack is played for money, the cards are often drawn from 8 decks of cards that are shuffled together. (The purpose of this is to make it harder to "count cards", i.e., to calculate how many aces and face cards are left in the deck.) In this case, there are so many cards in the deck that we can regard this as a situation "with replacement", i.e., that we draw each card from a regular deck of cards and replace it before drawing the next one. A standard deck has 52 cards, with 4 suits that each have their own ten, jack, queen, king and ace.

a) What is the probability of getting a blackjack ?

What is the probability of getting a blackjack if the first card you have been dealt is an ace?

Assume that a player who counts cards has found that the probability of getting a blackjack is 0.06, that the probability of getting an ace as the first card is 0.1 and that the probability that the first card was an ace if you have gotten a blackjack is 0.4. Use these numbers to compute what the probability now is of getting a blackjack if the first card you have been dealt is an ace.

Lars is vacationing in Las Vegas and goes into the largest casino he comes across. He sits down at the blackjack table and decides to play until he wins and to double the bet for each game. He bets one dollar in the first game, two dollars in the second game, and so on until he wins.

Assume (for simplicity) that he always gets back twice what he bet if he wins, that his probability of winning is 0.3 in every game, and that Lars stops playing after winning once.

b) Let X be the number of times Lars player before he quits. What is the probability distribution of X?

Assume in the rest of this problem that Lars runs out of money and thus stops playing if he has not won after playing five games. Let W be the number of dollars Lars wins at the casino (regardless of how much he has bet). What is the expected value of W?

Let Y be the winnings Lars is left with after his bets are deducted. What is the expected value of Y?

Oppgave 2

Agent John Bang goes to regular shooting practice. Experience tells him the probability of a hit is $p = 0.8$. During a practice session he has 20 trials. Assume that each shot is either a hit or a miss, and that the trials are independent.

a) What is the expected number of hits?

What is the probability that the number of hits is larger than the expected number of hits?

What is the conditional probability that he has 20 hits when we know he hits more than expected?

The boss decides that John should have a new gun. They hope this new one results in a better hitting probability. They want to check if this may hold, and John does a usual practice session consisting of 20 trials with the new gun.

b) Formulate the problem as a hypothesis test.

Use the common normal approximation to perform the test at significance level $\alpha = 0.1$ when the observed number of hits is 18.

c) Describe how the test can be done exactly using the binomial distribution.

What is the P-value of the test when he hits 18?

Assume significance level $\alpha = 0.1$ for the test, and compute the power of the exact test when the true probability of a hit is $p = 0.9$.

Oppgave 3

The median of a data set, $\tilde{X}$, is the middle value. If we have random variables $X_1, X_2, ..., X_n$ and sort them by size such that $X_{(1)} < X_{(2)} < ... < X_{(n)}$, the median is defined as

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})} & \text{hvis } n \text{ er et oddetall,} \\ \frac{1}{2}\left(X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)}\right) & \text{hvis } n \text{ er et partall.} \end{cases}$$

When the random variables are independent and normally distributed with expected value μ and variance σ^2, i.e., $X_i \sim N(\mu, \sigma^2)$, and we have that the number of variables n is large, we can assume that the variance of the median is

$$\text{var}(\tilde{X}) = \frac{1}{4n\big(f(\mu)\big)^2},$$

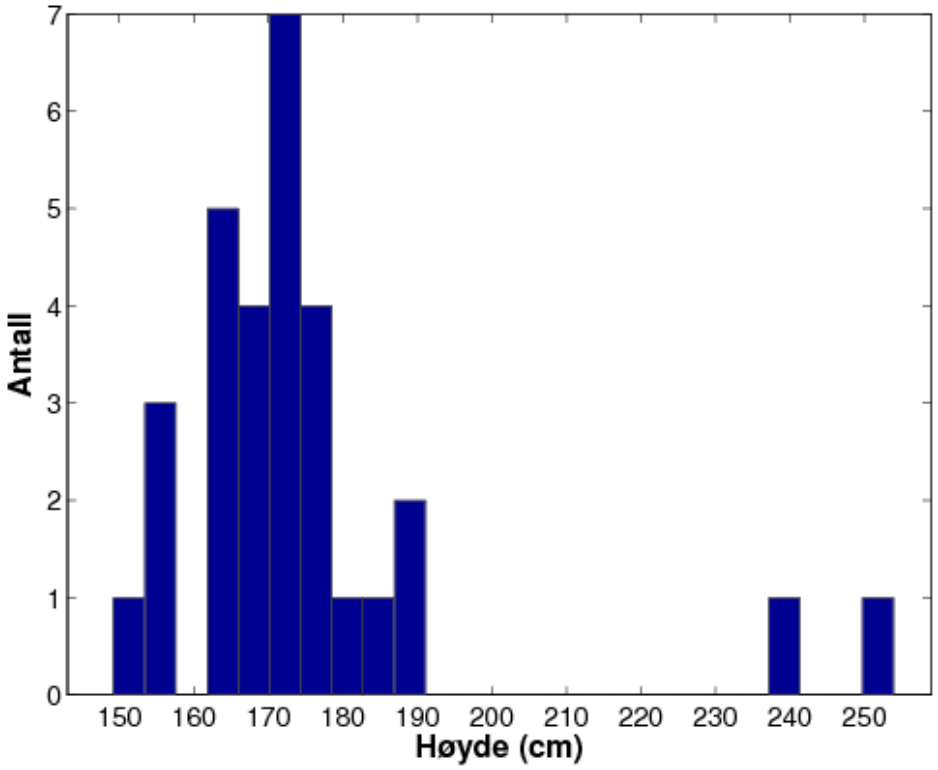
where $f(x)$ is the probability density of the normal distribution.

a) For this case, show that

$$\text{var}(\tilde{X}) = \frac{\pi}{2}\text{var}(\bar{X}),$$

where the average $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$\tilde{X}$ is an unbiased estimator for the expected value μ. Why we do we typically prefer $\bar{X}$ instead of $\tilde{X}$ as estimator for μ?



Figur 1: HÃ¸ydene til 30 rekrutter, kanskje fra 1814.

Statistics Norway has data on the heights of male Norwegian recruits to the army for every year back to 1878. In this problem, you can assume that you know that the heights of recruits in any year are normally distributed.

On Terningmoen camp Lieutenant Munthe has found a form with the heights of 30 recruits that he believes to be from 1814. The paper has yellowed and the ink is faded, but the lieutenant gets one of his current recruits to enter the data into a spreadsheet to the best of his ability. Figure 1 shows a histogram of this data.

b) For this dataset, will the median $\tilde{X}$ be greater than, less than or about equal to the average $\bar{X}$?

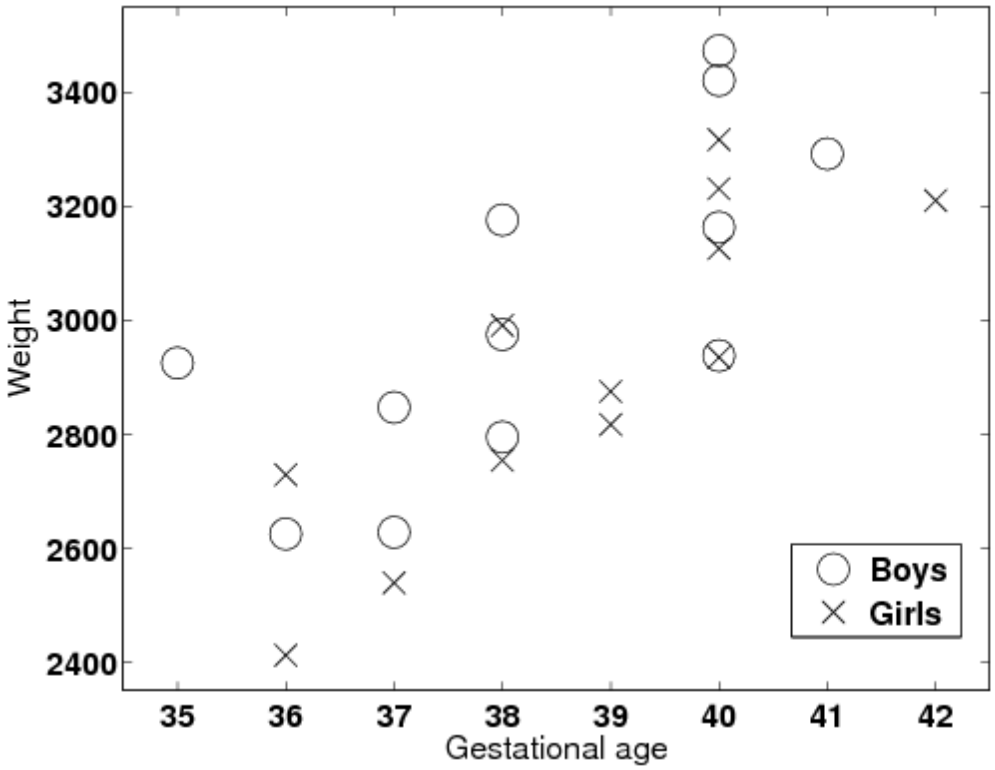
Would you have used the median or average to estimate the expected value μ here? Explain your answer.

Oppgave 4

Medical scientists study the weight of newborns as a function of their gestational age (time since conception). The data consists of gestational age x_i (weeks) and weight y_i (gram) for $i = 1, \ldots, n$ newborn babies, and $n = 24$.

For this dataset $\sum_{i=1}^{n} x_i y_i = 2\ 752\ 667$, $\sum_{i=1}^{n} x_i^2 = 35\ 727$, $\sum_{i=1}^{n} x_i = 925$ and $\sum_{i=1}^{n} y_i = 71\ 194$.

Assume a linear regression model for the data: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, n$, where ϵ_1, ...,



Figur 2: Scatter plot of gestational age and weight for 12 boys and 12 girls.

ϵ_n are assumed to be independent and Gaussian distributed with mean 0 and variance σ^2.

a) Use the summary statistics of data above to compute the estimates of the intercept and slope of the linear regression model: $\hat{\beta}_0$ and $\hat{\beta}_1$.

We compute an estimate of σ^2 as follows: $s^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 194^2$.

Compute a 95 percent confidence interval for the slope.

b) Use the data to construct a 90 percent prediction interval for the weight of a newborn in gestational week 40.

How wide is the 90 percent prediction interval we get for gestational week 42 compared with this one for week 40?

Figure 2 shows a scatter plot of gestational age and weight. In this plot we have split the data in two groups: boys and girls. There are $n_b = 12$ boys (numbered 1 through n_b in the following and 12 girls (numbered $i = n_b + 1$ through n in the following).

We suggest the following model for the data:

$$Y_i = \beta_b + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n_b.$$

$$Y_i = \beta_g + \beta_1 x_i + \epsilon_i, \quad i = n_b + 1, \ldots, n.$$

where we still assume $\epsilon_1, \ldots, \epsilon_n$ are independent and Gaussian distributed with mean 0 and variance σ^2.

c) Use the plot to explain why this model may seem reasonable. Explain further what elements of this model may be unwanted.

Use the model to compute the least squares estimates (or maximum likelihood estimates) denoted $\hat{\beta}_b$, $\hat{\beta}_g$ and $\hat{\beta}_1$. In addition to the above sums, we have $\sum_{i=1}^{n_b} y_i = 36\ 258$, $\sum_{i=1}^{n_b} x_i = 460$, $\sum_{i=n_b+1}^{n} y_i = 34\ 936$ and $\sum_{i=n_b+1}^{n} x_i = 465$.

Fasit

1. **a**) 0.04734, 0.3077, 0.24 **b**) 6.567, -4.4

2. **a**) 16, 0.41, 0.028 **b**) do not reject H_0 **c**) 0.21, 0.39

3. **b**) the median is lower than the average

4. **a**) -1465, 115, [69,161] **b**) [2789,3481], 690, 732 **c**) -1587, -1747, 120.22