TMA4245 Statistics
Exam December 2016

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

A company produces electrical components. The components can have two types of faults. We randomly choose one component from the production process and define two events: A=the component has a fault of type A, and B=the component has a fault of type B. Let A' and B' be the associated complementary events.

It is known that $P(B) = 0.09$, $P(A \mid B) = 0.5$ and $P(A \mid B') = 0.01$.

a) We study a randomly chosen component from the production process.

What is the probability that the component has both a fault of type A and a fault of type B, that is, $P(A \cap B)$?

What is the probability that the component has a fault of type A, that is, $P(A)$.

Given that the component has a fault of type A, what is the probability that the component has an a fault of type B, that is, $P(B \mid A)$?

We are now only interested in whether a component is fault-free or not. The management of the company has over many years monitored the production process, and is confident that the probability that a component is fault-free is 0.9. We randomly choose 20 components from the production prosess, and investigate if the components are fault-free. Let X be a random variable denoting the number of fault-free components.

b) What is the distribution of X? Justify your answer.

What is the probability that exactly 19 components are fault-free?

What is the probability that more than 15 components are fault-free?

The management of the company has implemented changes in the production process and hopes that the changes have lead to an increased proportion of fault-free components. We denote this unknown proportion of fault-free components p. We draw a random sample of size n components from the new production process and denote by X the number of fault-free components.

An intuitive estimator for p is the proportion of fault-free components in the random sample, that is, $\hat{P} = \frac{X}{n}$. When we have observed $X = x$ fault-free components we may calculate the estimate $\hat{p} = \frac{x}{n}$ for p. The random sample of size n is large enough for us to assume that $\frac{X-np}{\sqrt{n\hat{p}(1-\hat{p})}}$ approximately follows a standard normal distribution.

c) Derive a 90% confidence interval for p.

Calculate numerical values for the confidence interval when $n = 500$ and $x = 470$.

Explain briefly how to interpret the interval.

Oppgave 2

In this problem we consider the calculation of the expected value and the variance of an average when the observations making up the average are dependent.

Let X_1 and X_2 be random variables with $\mathrm{E}(X_1) = \mathrm{E}(X_2) = 2$, $\mathrm{Var}(X_1) = \mathrm{Var}(X_2) = 1$ and $\mathrm{Cov}(X_1, X_2) = \frac{1}{2}$.

Calculate $\mathrm{E}(\frac{1}{2}X_1 + \frac{1}{2}X_2)$ and $\mathrm{Var}(\frac{1}{2}X_1 + \frac{1}{2}X_2)$.

Further, let $X_1, X_2, \ldots, X_{10}$ be random variables with $\mathrm{E}(X_i) = 2$ and $\mathrm{Var}(X_i) = 1$ for $i = 1, 2, \ldots, 10$ and $\mathrm{Cov}(X_i, X_j) = \frac{1}{2}$ for all $i = 1, 2, \ldots, 10$ and $j = 1, 2, \ldots, 10$ where $i \neq j$. Let $\bar{X} = \frac{1}{10}\sum_{i=1}^{10} X_i$.

Calculate $\mathrm{E}(\bar{X})$ and $\mathrm{Var}(\bar{X})$.

Hint: you may use the following formula for the variance of a sum (the formula is also given in *Tabeller and formler i statistikk*)

$$\mathrm{Var}(\sum_{i=1}^{n} a_i X_i + b) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2\sum_{i=2}^{n}\sum_{j=1}^{i-1} a_i a_j \mathrm{Cov}(X_i, X_j).$$

Oppgave 3

We study a population of male students, and assume that the height of a randomly chosen male from the population is normally distributed with mean μ and variance σ^2.

a) Assume (only in this subproblem) that $\mu = 181\,\mathrm{cm}$ and $\sigma = 6\,\mathrm{cm}$. We randomly select two male students from the population and let X_1 denote the height of the first student and X_2 the height of the second student. We assume that X_1 and X_2 are independent random variables.

Calculate the following probabilites:
$\mathrm{P}(X_1 > 190)$

$\mathrm{P}(X_1 > 190 | X_1 > 185)$

$\mathrm{P}(X_1 > 190 | X_2 > 185)$

Two research groups have independently of each other estimated the mean height of male students, μ. Research group 1 used a random sample of size n and observed the heights $x_1, x_2, \ldots, x_n$, and research group 2 used a random sample of size m and observed the heights $y_1, y_2, \ldots, y_m$. The two random samples were drawn independently of each other from the given population.

Both research groups used the empirical mean (average) as the estimator for μ, $\bar{X} = (X_1 +$

$X_2 + \ldots + X_n)/n$ and $\bar{Y} = (Y_1 + Y_2 + \ldots + Y_m)/m$, and research group 1 found $\bar{x} = 180\,\text{cm}$ and research group 2 found $\bar{y} = 183\,\text{cm}$.

You have studied statistics and know that you can combine the estimates from the independent studies to construct an estimate for μ that has lower uncertainty than each of the separate estimates. You have decided to use the estimator

$$\hat{\mu} = a\bar{X} + b\bar{Y},$$

where a and b are real numbers.

b) Explain which two properties characterize a good estimator.

Find expressions for a and b (as functions of n and m) so that $\hat{\mu}$ is an estimator for μ that satisfies the two properties.

What is your estimate for μ when $n = 64$ and $m = 192$?

After closer consideration you find the difference between the two estimates from the two research groups to be unreasonably large taking the sample sizes $n = 64$ and $m = 192$ into consideration. Your claim is that the two research group have collected the random samples from different populations.

Assume that research group 1 drew a random sample from a normally distributed population with mean μ_1 and standard deviation σ_1, and that research group 2 drew a random sample from a normally distributed population with mean μ_2 and standard deviation σ_2. You have earlier been informed that $\bar{x} = 180\,\text{cm}$ and $\bar{y} = 183\,\text{cm}$. You contact the research groups and they send you the empirical standard deviations for their observations, $s_1 = 6.0$ for research group 1 and $s_2 = 5.5$ for research group 2.

c) Use your claim (given earlier in the text) to formulate a null- and an alternative hypothesis.

You may consider it known that the formula for the number of degrees of freedom in a test for the difference in means when σ_1 can be different from σ_2, is

$$\nu = \frac{(s_1^2/n + s_2^2/m)^2}{(s_1^2/n)^2/(n-1) + (s_2^2/m)^2/(m-1)} = 100.6,$$

where the value is calculated with the numerical values for s_1, s_2, n and m as given earlier in the text.

Argue why this test can be used and find the rejection region of the test when the significance level is chosen to be $\alpha = 0.05$.

What is the conclusion of the hypothesis test when you use the data given in the text?

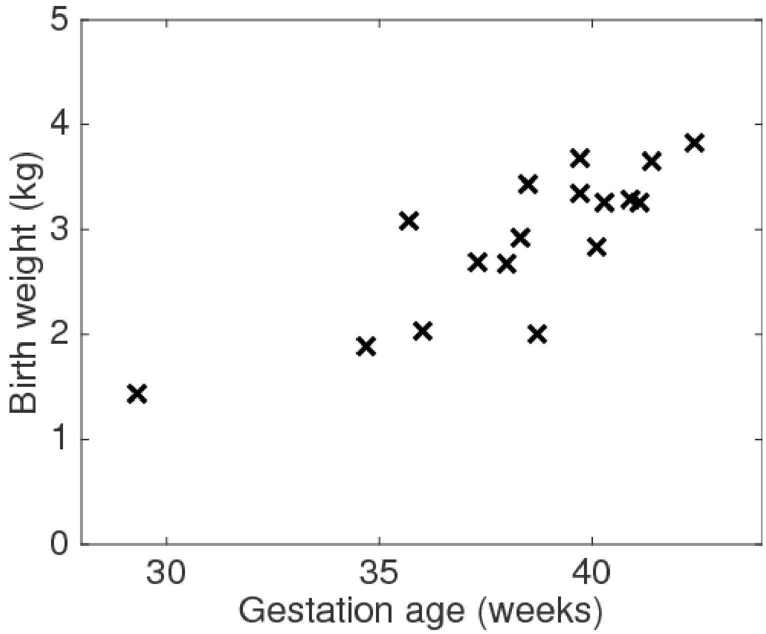
Oppgave 4

In Figure 1 you find a scatter plot of birth weight (measured in kg) and gestational age (time from the first day of the last menstrual cycle of the mother, measured in weeks) for $n = 17$ births.

We would like to fit a simple linear regression model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n,$$

where each ϵ_i is a normally distributed random variable with expected value 0 and variance σ^2. Further, $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independent, and Y_i is birth weight and x_i is gestational age.



Figur 1: Scatter plot of birth weight, y_i, and gestational age, x_i for $i = 1, \ldots, 17$ children.

a) Is it reasonable to use a linear regression model for the observations in Figure 1? Discuss briefly.

Briefly explain how the least squares method can be used to find estimators B_0 for β_0 and B_1 for β_1, and illustrate by drawing a figure. You shall not derive the expressions for the estimators.

It is given that the estimate for β_0 is -4.02 and for β_1 is 0.18. Find the predicted birth weight for children at gestational age 40 weeks.

b) Find an expression for the variance of $\hat{Y}_0 = B_0 + B_1 x_0$, where B_0 and B_1 are the least squares estimators for β_0 and β_1. You may use that $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and B_1 are independent random variables.

Examine Figure 2 and use the expression for the variance of $\hat{Y}_0$ to explain why the estimate for the expected value $\mathrm{E}(\hat{Y}_0)$ is more uncertain at $x_0 = 29$ weeks than at $x_0 = 39$ weeks.

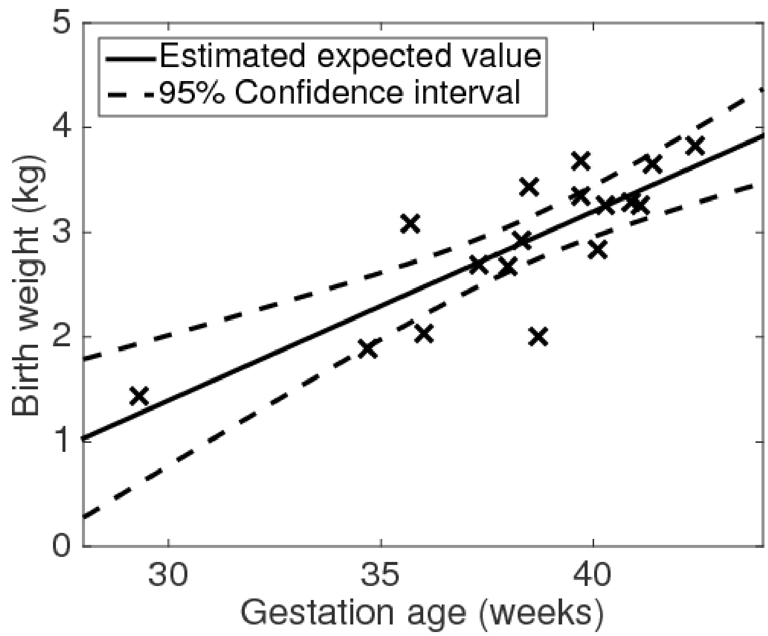
Oppgave 5

Assume that Y is uniformly distributed with probability density function

$$f_Y(y) = \begin{cases} 1, & 0 < y < 1, \\ 0, & \text{else.} \end{cases}$$

Find the cumulative distribution function $F_Y(y)$ for Y.

With the aid of a computer we often generate observations from a given distribution by first



Figur 2: Scatter plot of birth weight and gestational age for 17 children with estimated expected value of birth weight (regression line) and limits for 95% confidence intervals for expected birthweight as a function of gestational age.

drawing an observation from a uniform distribution and then transforming the observation. We will study the transformation $X = -\ln(Y)/\lambda$, where $\lambda > 0$.

Use $F_Y(y)$ to find the cumulative distribution function $F_X(x)$ for X.
Find the probability density function $f_X(x)$ of X. Which known statistical distribution is this?

Fasit

1. **a**) 0.045, 0.054, 0.83 **b**) binomial distribution, 0.270, 0.957 **c**) $[0.923, 0.957]$

2. 2, 0.75, 2, 0.55

3. **a**) 0.0668, 0.2657, 0.0668 **b**) $a = n/(n+m)$, $b = m/(n+m)$, 182.25 **c**) $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$, reject H_0

4. **a**) 3.18

5. Exponential distribution with mean value $1/\lambda$