TMA4240 Statistics
Exam December 2015

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

The distribution function (probability density function) for the continuous random variable X is given as

$$f(x) = \begin{cases} \theta \cdot x^{-(\theta+1)} & \text{for } x > 1 \\ 0 & \text{else,} \end{cases} \tag{1.1}$$

where $\theta > 0$. This distribution is a popular model when X is a normalised measure of wealth in a population.

a) Find the cumulative distribution function for X, $F(x) = P(X \leq x)$.

Let $\theta = 1.16$ and find $P(X \leq 2)$, $P(X > 4)$, and $P(X > 4 \mid X > 2)$.

Let $X_1, X_2, \ldots, X_n$ be a random sample (independent and identically distributed random variables) from a population described by $f(x)$ given in Equation (1.1).

b) Derive the maximum likelihood estimator for θ.

Oppgave 2

a) Let X and Y be two independent and normally distributed random variables, with $\mathrm{E}(X) = 1$, $\mathrm{E}(Y) = 2$, $\mathrm{Var}(X) = 4$ and $\mathrm{Var}(Y) = 1$.

Find $P(X \leq 0)$.

What is the distribution of $X + Y$?

Find $P(X + Y > 4)$ and $P(X - Y \leq -2)$.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a normally distributed population with expected value μ_X and standard deviation σ. Further, let $Y_1, Y_2, \ldots, Y_m$ be a random sample from a normally distributed population with expected value μ_Y and standard deviation σ. Assume that the two samples are independent.

Define $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{Y} = \frac{1}{m}\sum_{j=1}^{m} Y_j$, $S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ and $S_Y^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_j - \bar{Y})^2$.

b) We wish to estimate σ^2, and consider the following two estimators:

$$S^2_{\text{pooled}} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$
$$S^2_{\text{mean}} = \frac{1}{2}S_X^2 + \frac{1}{2}S_Y^2$$

Are the estimators unbiased?

Find the variance for each of the two estimators.

Let $n = 10$ and $m = 20$. Which of the two estimator would you prefer?

You may use that $V_X = \frac{(n-1)S_X^2}{\sigma^2}$ has a chi-square distribution with parameter $n-1$, and $V_Y = \frac{(m-1)S_Y^2}{\sigma^2}$ has a chi-square distribution with parameter $m-1$.

c) We would like to investigate if there is reason to believe that μ_X is greater than μ_Y.

Write down the null hypothesis and the alternative hypothesis.

Choose a test statistic, and give the distribution of the test statistics when the null hypothesis is true.

Find a rejection rule when we choose significance level 0.05.

What is the result of the test if we observe: $n = 129$, $\bar{x} = 75.2$, $s_X^2 = 174.6$, $m = 141$, $\bar{y} = 61.0$ and $s_Y^2 = 292.1$?

These numbers were taken from the Internet survey in TMA4240, autumn 2015. The X and Y-value give the answer to the question: "If you attend a lecture, or watch the lecture on video, how many percent of the lecture would you say that you on average understand". The X-sample consists of the students who answered "Agree" or "Strongly agree" to the the phrase "At this point in time I find that Statistics is a fun course" and the Y-sample are the students who answered "Neutral", "Disagree", or "Strongly disagree".

How would you translate your answer from the hypothesis test (reject or not reject the null hypothesis) to this situation?

Oppgave 3

Two robots communicate via binary signals. The communication of a signal is not perfect: There is a chance of 0 being send and 1 received, and if 1 is sent the other robot can receive 0.

a) Let X be the discrete random variable representing the sent signal, and denote by Y the discrete random variable for the received signal. The possible outcomes for X and Y are 0 and 1. Set $P(X = 1) = 0.2$. The communication is such that $P(Y = 1 \mid X = 1) = P(Y = 0 \mid X = 0) = 0.9$.

Use the law of total probability to compute $P(Y = 1)$.

Use Bayes' rule to compute the probability $P(X = 1 \mid Y = 1)$.

We next consider signals of length 5. Let X_i be a discrete random variable for signal element i, and Y_i a discrete random variable for received signal element i. Here $i = 1, 2, \ldots, 5$, and the possible outcomes for X_i and Y_i are 0 and 1. Assume that every element ($i = 1, \ldots, 5$) is communicated correctly with probability $P(Y_i = 1 \mid X_i = 1) = P(Y_i = 0 \mid X_i = 0) = 0.9$ for all $i = 1, 2, \ldots, 5$. The random variable Y_i only depends on X_i and not the other Y_j or X_j, for $j \neq i$.

b) What is the probability that the sent signal has been received correctly for at least 4 of the 5 elements?

For a particular situation one has knowledge saying that signal $(0, 0, 0, 0, 0)$ and $(1, 1, 1, 1, 1)$ are the most likely to be sent; each with probabilities 0.35. While the remaining 30 signals all have probabilities 0.01. The received signal is $(0, 0, 0, 0, 0)$. What is the probability that the signal $(0, 0, 0, 0, 0)$ was sent?

Oppgave 4

Last winter, on Sundays, Alexander sold cups of hot chocolate by the ski-tracks near his house. This winter he plans to have a similar business.

a) Denote by Y the number of cups Alexander sells on a given day. Assume this random variable Y is Poisson distributed with expected value $\lambda = 18$.

What is the probability that he sells 18 cups or less?

What is the probability that he sells more than 10 cups, given that he sells 18 or less?

We know that the Poisson distribution is closely approximated by a normal distribution when λ is large. In the following we assume that the number of cups of hot chocolate that Alexander sells on a given day is normal distributed.

Alexander experienced that the sales changed dramatically with the weather and skiing conditions. He made a condition index, x, where $x = 1$ means "bad conditions", $x = 2$ means "good conditions", $x = 3$ means "very good conditions" and $x = 4$ means "excellent conditions".

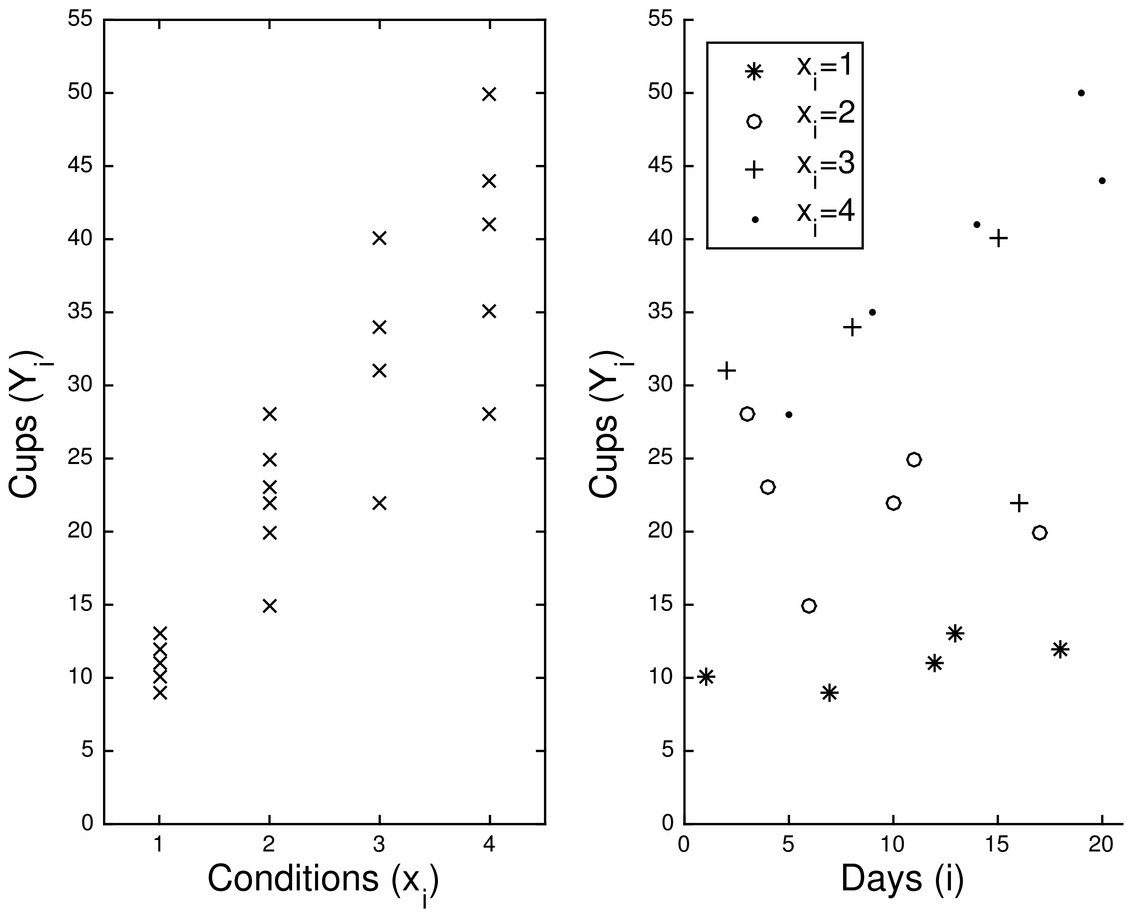
For 20 Sundays, $i = 1, \ldots, 20$, he registered both the number of cups sold, denoted y_i and the associated condition index, x_i.

We will phrase the sales as a regression model taking condition as an explanatory variable:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, 20,$$

where $\epsilon_1, \ldots, \epsilon_{20}$ are independent normal distributed variables with expected value 0 and variance σ^2, and β_0 and β_1 are fixed but unknown regression parameters.

Based on the data $\bar{x} = \frac{1}{20}\sum_{i=1}^{20} x_i = 2.45$, $\bar{y} = \frac{1}{20}\sum_{i=1}^{20} y_i = 25.65$, $\sum_{i=1}^{20}(x_i - \bar{x})^2 = 24.95$ and $\sum_{i=1}^{20}(x_i - \bar{x})y_i = 237.15$. Further, $s^2 = \frac{1}{18}\sum_{i=1}^{20}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 5.65^2$ is an estimate of the variance σ^2, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the least squares estimates of β_0 and β_1, respectively.



Figur 1: Left: The number of cups sold (y-axis) for different conditions (x-axis). Right: The number of cups sold (y-axis) for all days (x-axis). The conditions index is indicated.

b) The dataset is plotted in Figure 1. Do the modeling assumptions look reasonable from this plot?

Assume that the conditions are excellent one Sunday. Alexander wants to predict the number of cups he will sell. Compute a 90 percent prediction interval for the number of cups sold that day.

In the following you can use these result: If Y is normally distributed with mean μ and variance σ^2, i.e. the probability density is $f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(y-\mu)^2}{2\sigma^2})$, then $\int_{-\infty}^{a} y f(y) dy = \mu\Phi((a-\mu)/\sigma) - \sigma\phi((a-\mu)/\sigma)$, where $\phi(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2})$ is the probability density of the standard normal and $\Phi(z)$ is the cumulative distribution function of this standard normal.

c) Alexander has sometimes run out of hot chocolate. Suppose the predicted number of cups on a given day is normal distributed with mean 20 and variance 5^2, but the maximum number he can sell is the prepared $n = 25$ cups of hot chocolate. Compute the expected number of cups sold that day.

Alexander buys and prepares one cup of hot chocolate at the cost of 5 kroner. He sells a cup for 20 kroner. If Alexander is not able to sell all n cups of chocolate, he just empties the rest (without income). The goal is to maximize the expected profit, i.e. income with cost subtracted. Find the number n of cups he should prepare to optimize his profits. Hint: Use trial and error to find the optimal n.

Fasit

1. **a**) $F(x) = 1 - \frac{1}{x^\theta}$ når $x > 1$, 0.55, 0.20, 0.44

2. **a**) 0.3085, 0.3264, 0.3264 **b**) Both are unbiased, $\text{Var}\left(S^2_{pooled}\right) = \frac{2\sigma^4}{n+m-2}$, $\text{Var}\left(S^2_{mean}\right) = \frac{\sigma^4}{2}\left(\frac{1}{n-1} + \frac{1}{m-1}\right)$, prefer S^2_{pooled} **c**) $H_0 : \mu_X = \mu_Y$, $H_1 : \mu_X > \mu_Y$, test statistic: $T_0 = (\overline{X} - \overline{Y})/(S_{pooled}(1/n + 1/m)^{1/2})$, reject H_0 if $T_0 > t_{0.05,n+m-2}$, reject H_0

3. **a**) 0.26, 0.69 **b**) 0.92, 0.98

4. **a**) 0.56, 0.946 **b**) $(29.9, 50.9)$ **c**) 19.58, 23