In exercises 1–3 the sample is the artificial dataset 1, 2, 3, 4, 5, 6, 7.

1. Find the sample quantiles \( q(0.2), q(0.4), q(0.6) \) and \( q(0.8) \) both a) with the method presented in the book WMMY and b) with the presented better method (pages 3–4 in the lecture notes).

2. A statistical program has to produce an unambiguous sample quantile. The statistical program JMP computes the quantiles in its own way, as follows. In order to calculate \( q(f) \), first formulate the corresponding ordinal number \( j = (n+1)f \). Then proceed as follows:
   
   1. If \( j \) is an integer, then choose \( q(f) = x(j) \) (the \( j \):th smallest element of the sample).
   2. If \( j \) isn’t an integer, find its integer part \( k \) and choose
      
      \[ q(f) = (1 - j + k)x(k) + (j - k)x(k+1). \]

   Use this method to find a) the sample median and b) the sample quantiles \( q(0.2) \) and \( q(0.4) \) of the sample.

3. The (default) method of the R program to compute quantiles is rather special. Now the ordinal is \( j = (n-1)f \) and the quantile \( q(f) \) is obtained in the following manner:
   
   1. If \( j \) is an integer, then choose \( q(f) = x(j+1) \).
   2. If \( j \) isn’t an integer, find its integer part \( k \) and choose
      
      \[ q(f) = (k - j)x(k) + (1 + j - k)x(k+1). \]

   Use this method to find a) the sample median and b) the sample quantiles \( q(0.2) \) and \( q(0.4) \) of the sample.

   (In R there are no less than nine different ways to compute the sample quantiles; one of them is the above mentioned method that JMP uses.)

4. A dice is thrown 100 times. Using the Central limit theorem, approximate the probability that the sum of the spots is at least 330.

5. In the table below there are two (fictitious) samples of 55 students in two (fictitious) universities. The students are grouped according to whether they have studied the short or long syllabus of mathematics in upper secondary school. Looking the groups’ corresponding grade point averages (GPA) of all subjects in upper secondary school, it appears that students who study the long mathematics syllabus have higher GPA. But what happens when the 110 students are pooled?

   (The phenomenon is called Simpson’s paradox and is well known: Google finds about 69 000 results. The more “philosophical” aspect of the matter is considered for example on the web-page http://plato.stanford.edu/entries/paradox-simpson/.)
1. Some BOD₅ values (related to the rate at which microorganisms in water take up oxygen) were measured. In October the sample mean of four measurements was $\overline{x}_{\text{Oct}} = 11 \text{ mg/l}$ and in November the sample mean of four measurements was $\overline{x}_{\text{Nov}} = 9 \text{ mg/l}$. The population distribution of the BOD values is normal and its standard deviation is always $\sigma = 2.95 \text{ mg/l}$. Is there a difference between October and November?

2. Consider a random variable $V$ that is $\chi^2$-distributed with $v$ degrees of freedom.
   a) Find an $x$ such that $P(V \geq x) = 0.99$, when $v = 4$.
   b) Find an $x$ such that $P(V \leq x) = 0.025$, when $v = 19$.
   c) Find an $x$ such that $P(37 \leq V \leq x) = 0.045$, when $v = 25$.

   Use a table, web-based calculators or statistical programs.

3. A packing machine packs beer into cans. The amount of beer per can is normally distributed. In order for the machine to meet the quality control requirements, the standard deviation of the amount of beer should be at most 10 g. In a random sample of $n = 40$ elements the obtained sample variance was $s^2 = 151.09 \text{ g}^2$. What do you conclude?

4. Consider a random variable $T$ that is $t$-distributed with $v$ degrees of freedom.
   a) Find an $x$ such that $P(T \geq x) = 0.99$, when $v = 4$.
   b) Find an $x$ such that $P(T \leq x) = 0.025$, when $v = 19$.
   c) Find an $x$ such that $P(|T| \geq x) = 0.001$, when $v = 25$.

   Use a table, web-based calculators or statistical programs.

5. When studying the typical level of use of a certain model of consumer electronic device, the number of days of use per year were collected on 10 randomly chosen devices. The mean was $\overline{x} = 340.6 \text{ d/y}$ and the standard deviation $s = 10.1 \text{ d/y}$. Could we say that the devices are being used 95% of the time (that is about 346.75 days per year)? The number of days of use can be supposed to be normally distributed.

6. The population distribution is a normal distribution $N(\mu, \sigma^2)$.

   a) Find the variance $S^2$ of the sample variance $\text{var}(S^2)$ by using the variance of the $\chi^2$-distribution and show that it decreases as the sample size $n$ increases.

   b) Two independent samples are taken from the population, the population variances of which are $S^2_1$ and $S^2_2$. The sample sizes are $n_1$ and $n_2$ respectively. From these sample variances the pooled sample variance $S^2_p$ (to which we’ll return later in section 2.5) is calculated with the formula

   $$S^2_p = \frac{n_1 - 1}{n_1 + n_2 - 2} S^2_1 + \frac{n_2 - 1}{n_1 + n_2 - 2} S^2_2.$$

   Show that the variance of $S^2_p$ is smaller than the variances of $S^2_1$ and $S^2_2$.

   (The moral of the story is: When estimating the variance there has to be as much data as possible and all the data has to be used!)
1. Consider a random variable $F$ that is F-distributed with $v_1$ and $v_2$ degrees of freedom.

a) Find an $x$ such that $P(F \geq x) = 0.07$, when $v_1 = 4$ and $v_2 = 10$.

b) Find an $x$ such that $P(F \leq x) = 0.025$, when $v_1 = 9$ and $v_2 = 12$.

c) Find an $x$ such that $P(F \leq x) = 0.5$ (median), when $v_1 = 20$ and $v_2 = 15$.

Use a table, web-based calculators or statistical programs.

2. Holes drilled by a human operator (M) and by a robot (R) were studied. The numbers of holes were $n_M = 7$ and $n_R = 8$. The obtained sample standard deviations of the hole depths were $s_M = 0.0435$ cm and $s_R = 0.0131$ cm. The drill hole depths are assumed to be normally distributed in both cases. Can it be claimed that the standard deviation is smaller in holes drilled by the robot?

3. Coffee beans are packed in 100 kg packages for shipping. The exact weight of a package is however a random variable that can be assumed to be normally distributed with variance $\sigma^2 = 0.85$ kg$^2$.

a) Find the 95 % confidence interval for the (mean) weight of a package, when the sample mean of $n = 10$ studied packages was $\mu = 99.62$ kg.

b) The manufacturer wants the 95 % confidence interval for the mean weight to be at most 1 kg. How large a sample should be taken?

c) Find also the 95 % lower confidence limit for the mean weight in part a.

4. In order to study the heights of adult men in a certain region, $n = 50$ were randomly chosen. The obtained sample mean was $\mu = 178.15$ cm and the sample standard deviation was $s = 5.92$ cm. Assume heights to be approximately normally distributed.

a) Find the 98 % confidence limit for the mean height of men in that region.

b) How large should the sample size be to ensure that the estimation error of the mean length would at the probability of 98 % be at most 1 cm? Use the sample standard deviation as the population standard deviation.

(There is considerable variation in heights of men in different regions; women’s heights vary less. According to statistics on the Internet, Dutch men are on average many centimeters taller than Finnish men!)

5. Consider a population distribution having a single unknown parameter $\theta$; any other parameters are assumed to be known. The density function of the distribution is $f(x; \theta)$. (The notation shows that the density depends also on the parameter $\theta$.) The Fisher information of the distribution with respect to $\theta$ is the expectation

$$I(\theta) = E\left(\left(\frac{\partial}{\partial \theta} \ln f(X; \theta)\right)^2\right),$$

where $X$ has a population distribution with the parameter $\theta$. The Fisher information measures how much information a sample from the population gives about $\theta$.

a) Calculate the Fisher information of a normal distribution $N(\mu, \sigma^2)$ with respect to $\mu$. (Consider the variance $\sigma^2$ known.)

b) According to the Cramér–Rao inequality, if $\hat{\theta}$ is an unbiased estimator of the parameter $\theta$ in a sample of size $n$, then $\text{var}(\hat{\theta}) \geq 1/(nI(\theta))$. If the inequality is in fact an equality, the estimator is said to be efficient. Show that the estimate $\bar{X}$ (sample mean) is an efficient estimator for the expectation $\mu$ of the normal distribution.

(The Cramér–Rao inequality is related to the Heisenberg uncertainty principle in quantum physics.)
1. The critical thickness of a metal part, which is desired to be exactly 1 cm, was measured. The thickness can be supposed to be normally distributed. As a result the sample mean $\overline{x} = 1.0008$ cm and sample standard deviation $s = 0.0028$ cm were obtained. Find a) the 99% prediction interval for the thickness and b) the tolerance interval in which at least 95% of the population distribution lies with probability 0.90.

2. The times of use of cutting tools in a metal lathe were studied at two feeding speeds. The number of cutting tools was 40 in both tests. The sample means $\overline{x}_1 = 41.525$ min (speed 1) and $\overline{x}_2 = 40.439$ min (speed 2) were obtained. The population distributions of times of use are supposed to be normal $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ with equal known variance $\sigma^2 = 20$ min$^2$ (based on the obtained sample variances).

   a) Find the 95% confidence interval for the difference of the expectations $\mu_1 - \mu_2$.

   b) Find also the 95% lower confidence limit for the difference.

3. $n_1 = 8$ equal-size ore samples were taken from location 1 in a mine and their copper content was measured (grams). The obtained sample mean was $\overline{x}_1 = 2.61$ g and the sample variance $s^2_1 = 0.040$ g$^2$. From location 2 in the same mine $n_2 = 10$ samples were taken, from which were obtained the sample mean $\overline{x}_2 = 2.32$ g and the sample variance $s^2_2 = 0.022$ g$^2$. The population distributions of copper contents are supposed to be normal and their population variances to be equal. Find the 98% confidence interval for the difference of the expectations of copper contents in locations 1 and 2.

4. A certain blood pressure medicine was studied with two groups of 9 subjects. The medicine was given to group 1 and a placebo was given to group 2. The tests subjects weren’t told which group they belong to. The subjects’ diastolic blood pressure was measured. For group 1 it was obtained $\overline{x}_1 = 89.1$ mmHg and $s_1 = 6.6$ mmHg and for group 2 $\overline{x}_2 = 93.3$ mmHg ja $s_2 = 11.9$ mmHg. The population distributions are supposed to be normal. The difference of the sample standard deviations is however so large that the population variances cannot be supposed to be equal. Find the 96% confidence interval for the difference of expectations of diastolic blood pressure $\mu_1 - \mu_2$.

5. If the variance $\sigma^2$ of the population distribution is known, the tolerance intervals are easier to construct. (The population distribution is still supposed to be a normal distribution $N(\mu, \sigma^2)$.) Especially easy is the case where only a one-sided tolerance interval is needed.

   For example the upper tolerance limit $\overline{x} + k\sigma$ is obtained from an $n$-sized sample by solving for $k$ in the equation

   $$P\left(\frac{\overline{X} + k\sigma - \mu}{\sigma} \geq z_\alpha\right) = 1 - \gamma$$

   by using the information that the random variable $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ has the standard normal distribution (see the footnote on page 23 of the course notes) What is $k$?

   (The above mentioned upper limit is interpreted to mean that at least $(1 - \alpha)100\%$ of the population distribution is included in the interval $(-\infty, \overline{x} + k\sigma)$ at the probability $1 - \gamma$. Easier is also the case where the expectation $\mu$ of the population distribution is known but the variance isn’t. One-sided tolerance limits are then obtained quite directly.)
1. Wind velocity at Pirkkala airport was measured at noon on \( n = 12 \) days using data from a ground station and using satellite data. The sample mean of differences of the obtained measurements was \( d = -0.412 \) m/s and the sample standard deviation \( s = 1.140 \) m/s. The difference can be assumed to be normally distributed. Find the 95 \% confidence interval for the difference of the expectations.

2. In a study of the side effects of a certain treatment, 200 treated patients were interviewed and 65 of them had experienced side effects. Find a) the 95 \% confidence interval and b) the 99 \% upper confidence limit for the proportion of patients who experience side effects because of the treatment.

3. When measuring the quality of air 16 times in the district A, the sample standard deviation of the air quality index was \( s_A = 10 \). Find the 90 \% confidence interval for the population variance \( \sigma_A^2 \) of the index in that district. Find the 90 \% confidence interval also for the population standard deviation \( \sigma_A \). (The air quality index can be assumed to be normally distributed.)

4. (Continued) Correspondingly, when measuring the quality of air 14 times in another district B, the sample standard deviation of the index was obtained \( s_B = 8 \). Find the 95 \% confidence interval for the ratio \( \sigma_A^2 / \sigma_B^2 \) of population variances of the indexes.

5. The probabilities of the Poisson distribution \( \text{Poi}(\lambda) \) are obtained from the formula

\[
P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda} \quad (x = 0, 1, 2, \ldots).
\]

You may remember that \( \text{E}(X) = \lambda \). What kind of formulas do you obtain for the 100(1 - \( \alpha \)) \% confidence limits \( \hat{\lambda}_L \) and \( \hat{\lambda}_U \) of the parameter \( \lambda \) if the realized value of \( X \) is \( x \)? Write the formulas without using infinite series.

(These equations as well are difficult to solve numerically. There are however web-calculators: See for example http://statpages.org/confint.html.)
1. The administration of a certain university estimates that about 20% of the students come to the university by bicycle, and bicycle parking space is reserved accordingly. In a random sample of \( n = 90 \) students, 28 come to the university by bicycle. Test at the risk level \( \alpha = 0.01 \) is the number of cyclists larger than was originally estimated. Use a suitable test statistic.

2. The occurrences (number of affected leaves) of a certain tree mycosis were counted in the same eight trees in an “ordinary” summer and in a “warm” summer. The sample mean of the difference (“warm” − “ordinary”) was \( \overline{d} = 10.6 \) leaves and the sample standard deviation \( s = 12.0 \) leaves. The difference can be considered approximately normally distributed (despite the fact that observations are integer-valued). Test at the levels of risk \( \alpha = 0.01 \) and \( \alpha = 0.05 \) whether or not there is a difference between the summers.

3. The delivery times of two delivery companies A and B were compared. \( n_A = 7 \) times were recorded from the company A and \( n_B = 5 \) times from the company B. The times can be assumed to be normally distributed with the same (unknown) variance. The realized sample means were \( \overline{x}_A = 20.23 \) h and \( \overline{x}_B = 18.68 \) h and the sample standard deviations \( s_A = 2.74 \) h and \( s_B = 1.64 \) h. Test at the risk \( \alpha = 0.05 \) whether the average transport times of the two companies can be considered equal or not and calculate also the P value.

4. The results of intelligence quotient (IQ) tests are always scaled so that the population is normally distributed with standard deviation \( \sigma = 15 \) points. The sample standard deviation of IQs of \( n = 24 \) randomly chosen professors is only \( s = 10.1 \) points.

   Supposing that the professors’ IQ is normally distributed, test at the risk \( \alpha = 0.05 \) if the standard deviation of professors’ IQ is actually smaller than average or not.

5. The arsenic concentration of the soil was measured in two different ways. With method 1 the sample standard deviation of \( n_1 = 10 \) measurements was \( s_1 = 0.8 \) ppm, and with method 2 the sample standard deviation of \( n_2 = 12 \) measurements was \( s_2 = 1.2 \) ppm. The results of the measurements can be assumed to be normally distributed.

   Test at the risk \( \alpha = 0.05 \) whether method 1 is more accurate than method 2, in other words whether the variance of method 1 is smaller than the variance of method 2.

6. Like for the binomial distribution, a test statistic used for the Poisson distribution Poi(\( \lambda \)) is the realized number \( x \). (See last week’s exercise 5.)

   a) How does one calculate the P-probability for the hypothesis pair \( H_0 : \lambda = \lambda_0 \) vs. \( H_1 : \lambda \neq \lambda_0 \) concerning the parameter \( \lambda \)?

   b) How does one obtain the critical region for this hypothesis pair at the risk \( \alpha \)?

   c) The Poisson distribution is used among other things to model the number of clients arriving at a service point during a certain amount of time.

   A certain service point is designed to handle clients arriving a rate of \( \lambda_0 = 35 \) per hour. You visit the service point and observe that only \( x = 21 \) clients arrived in the hour you were there. Would there be a reason to redesign the service point? Use P-probability. The cumulative distribution function can be obtained also from web-calculators, for example:

   \[ \text{http://stattrek.com/Tables/Poisson.aspx} \]
1. Gregor Mendel concluded in his famous pea crossing experiments that the probabilities of different kinds of peas is:

- S&Y: $\frac{9}{16}$, S&G: $\frac{3}{16}$, C&Y: $\frac{3}{16}$, C&G: $\frac{1}{16}$

(S = "smooth", Y = "yellow", C = "crinkly", G = "green"). Mendel examined $n = 556$ peas, which were distributed in different types as follows:


Does this agree with Mendel’s laws? Calculate the P-probability.

2. In 2006 the 3741 students of Electrical engineering (E), Science and engineering (S) and Information technology (I) in Finland who were women (W) and men (M) is given according to the contingency table on the right.

<table>
<thead>
<tr>
<th></th>
<th>W</th>
<th>M</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>E</td>
<td>237</td>
<td>1292</td>
<td>1529</td>
</tr>
<tr>
<td>S</td>
<td>129</td>
<td>255</td>
<td>384</td>
</tr>
<tr>
<td>I</td>
<td>179</td>
<td>1649</td>
<td>1828</td>
</tr>
<tr>
<td>Σ</td>
<td>545</td>
<td>3196</td>
<td>3741</td>
</tr>
</tbody>
</table>

a) If this is considered a random sample concerning both training program and gender, then is the distribution of different training programs independent of gender?

b) What about if only Electrical engineering and Information technology are considered?

3. a) Test the independence of the contingency table on the right at the risk $\alpha = 0.05$.

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_1$</td>
<td>62</td>
<td>81</td>
<td>143</td>
</tr>
<tr>
<td>$T_2$</td>
<td>75</td>
<td>55</td>
<td>130</td>
</tr>
<tr>
<td>Σ</td>
<td>137</td>
<td>136</td>
<td>273</td>
</tr>
</tbody>
</table>

c) How are the null hypotheses of parts a) and b) logically related?

d) How can the results be explained?

4. Trucks of three different age groups $R_1$ (under 3 y), $R_2$ (3 – 5 y) and $R_3$ (over 5 y) were studied in terms of safety and the numbers on the right were obtained. Is there a difference between different age groups when concerning safety? Use P-values.

<table>
<thead>
<tr>
<th>Safety</th>
<th>$R_1$</th>
<th>$R_2$</th>
<th>$R_3$</th>
<th>Σ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>83</td>
<td>73</td>
<td>28</td>
<td>184</td>
</tr>
<tr>
<td>Satisfactory</td>
<td>25</td>
<td>35</td>
<td>39</td>
<td>99</td>
</tr>
<tr>
<td>Bad</td>
<td>7</td>
<td>12</td>
<td>23</td>
<td>42</td>
</tr>
<tr>
<td>Σ</td>
<td>115</td>
<td>120</td>
<td>90</td>
<td>325</td>
</tr>
</tbody>
</table>

5. A certain population distribution is thought to follow the Poisson distribution Pois$(3.5)$ (see Task 5 / Exercise 5). In order to study the matter a sample, size of $n = 100$, was taken, from which the following results were obtained:

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>≥ 7</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_i$</td>
<td>5</td>
<td>16</td>
<td>13</td>
<td>21</td>
<td>17</td>
<td>14</td>
<td>6</td>
<td>8</td>
</tr>
</tbody>
</table>

a) What do you conclude according to the sample? Use the P value.

b) How does this kind of test differ from the test of last week’s Exercise 6?

The density and cumulative distribution functions of the Poisson distribution can be obtained from web-calculators, for example:

http://stattrek.com/Tables/Poisson.aspx
1. The density function of the exponential distribution is
\[ f(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{when } x \geq 0 \\ 0 & \text{otherwise.} \end{cases} \]
Find the maximum likelihood estimate \( \hat{\lambda} \) for the parameter \( \lambda \) calculated from the sample \( x_1, \ldots, x_n \).

2. The density function of the binomial distribution is
\[ f(x; p, n) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x}, & \text{when } x = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases} \]
The parameter \( n \) is supposed to be known — as it usually is. Find the maximum likelihood estimate \( \hat{p} \) for the parameter \( p \) based on
a) one realized value \( x \),
b) a sample of \( m \) values \( x_1, \ldots, x_m \).

3. If the parameter \( \mu \) of the normal distribution \( N(\mu, \sigma^2) \) is known, then what is the maximum likelihood estimate \( \hat{\sigma}^2 \) obtained from the sample \( x_1, \ldots, x_n \) for the parameter \( \sigma^2 \)?

4. The density function of the random variable \( X \) is
\[ f(x; \theta) = \begin{cases} \theta, & \text{when } 0 \leq x < 1 \\ 1 - \theta, & \text{when } 1 \leq x \leq 2 \\ 0 & \text{otherwise.} \end{cases} \]
What is the maximum likelihood estimator \( \hat{\theta} \) obtained from the sample \( x_1, \ldots, x_n \) for the parameter \( \theta \)?

5. The density function of the general Cauchy distribution is
\[ f(x; \theta) = \frac{1}{\pi((x - \theta)^2 + 1)}, \]
where \( \theta \) is a parameter of the distribution. The distribution is symmetric about the value \( x = \theta \). See the figure on the right.
It would seem to be natural to choose the realized sample mean as the estimate for \( \theta \). The Cauchy distribution however doesn’t have an expectation, so that doesn’t work. Form the equations from which it’s possible to (numerically) solve the maximum likelihood estimate for the parameter \( \theta \) from the sample \( x_1, \ldots, x_n \).
1. Fit the model \( y = \beta_0 + \beta_1 x + \epsilon \) to the given data. Additionally, plot the points \((x_i, y_i)\) and the line \( y = b_0 + b_1 x \) on the same plot.

(There are a zillion web-calculators to fit such simple regression model, for example http://www.easycalculation.com/statistics/regression.php. They don’t however necessarily test anything. The data, by the way, are the results of a study of abrasion.)

2. When studying the dependence of dye absorption on treatment time \( x_1 \) (min) and temperature \( x_2 \) (°C), the data on the right was obtained in nine experiments (the response \( y \) is in mg).

\[
\begin{array}{c|ccc}
 y & x_1 & x_2 \\
\hline
136 & 15 & 105 \\
153 & 30 & 105 \\
186 & 60 & 105 \\
182 & 15 & 120 \\
175 & 30 & 120 \\
187 & 60 & 120 \\
170 & 15 & 135 \\
179 & 30 & 135 \\
183 & 60 & 135 \\
\end{array}
\]

a) Estimate the parameters \( \beta_0, \beta_1, \beta_2 \) of the model \( y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \) and the variance of the error term \( \sigma^2 \).

b) Is the model significant at the risk \( \alpha = 0.05 \)? (The error term of the model is assumed to be normally distributed.)

c) Furthermore, test at the risk \( \alpha = 0.05 \) the parameters one by one (that is, whether the parameter = 0 or not).

d) Calculate the coefficient of determination \( R^2 \).

e) What do you conclude from the model overall?

(There is a good regression calculator on the web-page http://www.wessa.net/esteq.wasp that prints a lot of information, including the answers to this question.)

3. In a rather artificial situation, the data matrix of a simple regression model \( y = \beta_0 + \beta_1 x_1 + \epsilon \) is

\[
X = \begin{pmatrix}
1 & 1 + \delta \\
1 & 1 \\
1 & 1 - \delta
\end{pmatrix}.
\]

With small absolute values of \( \delta \) the columns of the data matrix are “nearly linearly dependent” and of course linearly dependent if \( \delta = 0 \). Calculate the matrix \( C = (X^T X)^{-1} \) and determine how small values of \( \delta \) affect the variances of the parameter estimates \( b_0 \) and \( b_1 \).

(Everyone remembers the inverse formula of a \( 2 \times 2 \)-matrix \( \begin{pmatrix} a & b \\ c & d \end{pmatrix} \)^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \), right?)

4. a) Show that \( X^T P = O \) and that as a consequence \( X^T e = 0_{k+1} \). (Here \( O \) is a \((k + 1) \times n \) null matrix and \( 0_{k+1} \) is a \( k + 1 \) null vector.)

(The residual vector \( e \) is diagonal to the columns of the data matrix \( X \). The regression model explains only that part of the response that is in the column space of the data matrix.)

b) Furthermore, conclude that the sum of the elements of the residual vector \( e \) is \( = 0 \).

(As a consequence of this the realized sample mean \( \bar{e} \) of the residuals is also always \( = 0 \).)

5. In the model

\[
y = \gamma_0 + \gamma_1 \sin(\omega t + \phi) + \epsilon
\]

\( \gamma_0 \) and \( \gamma_1 \) are unknown coefficients, \( t \) is a time variable and \( \epsilon \) is a noise term. The angular velocity \( \omega \) is known and \( \neq 0 \). The angle of phase \( \phi \) is unknown and is included in the interval \([0, \pi]\).

a) What are the regressors, parameters and responses and how is the data matrix \( X \) obtained when the model is changed into a linear regression model and the response samples are obtained at times \( t_1, \ldots, t_n \)?

b) How are \( \gamma_0, \gamma_1 \) and \( \phi \) obtained from the estimated parameters of the regression model?
1. Here $n = 12$ blood pressure exam results are shown. The response $y$ is the change of blood pressure after a certain procedure $z_1$. The levels of the categorical regressor $z_1$ are HE ("healthy food and exercise program"), M ("medication") and N ("no change"). Additionally, the weight indexes $x_1$ of the subjects were recorded. Fit the regression model and examine its significance and residuals. Which one is better, healthy food and exercise or medication?

2. Consider a regression model with two categorical regressors $z_1$ and $z_2$ and no continuous regressors. The regressor $z_1$ has two levels and the regressor $z_2$ has three levels. Three experiments are performed for each of the six possible combinations of levels. What kind of encoded data matrix results?

3. Two encodings from categorical to continuous regressors were presented in the lecture notes: the “ordinary” encoding and the encoding used by the JMP program. Explain how to convert the JMP encodings to the “ordinary” encodings.

(On the web-page http://www.ats.ucla.edu/stat/sas/webbooks/reg/chapter5/sasreg5.htm other encoding systems are presented.)

4. Why can’t the categorical regressor $z_i$, whose levels are $A_{i,1}, \ldots, A_{i,m_i}$, be encoded with $m_i$ “ordinary” regressors $z_{i,1}, \ldots, z_{i,m_i}$ and the method given by the enclosed table, which is possibly the first method that comes to mind?

<table>
<thead>
<tr>
<th>$z_i$</th>
<th>$z_{i,1}$</th>
<th>$z_{i,2}$</th>
<th>…</th>
<th>$z_{i,m_i-1}$</th>
<th>$z_{i,m_i}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_{i,1}$</td>
<td>1</td>
<td>0</td>
<td>…</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A_{i,2}$</td>
<td>0</td>
<td>1</td>
<td>…</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>$A_{i,m_i}$</td>
<td>0</td>
<td>0</td>
<td>…</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

5. In a study of the effect of smoking on the occurrence of coronary thromboses (heart attacks), data on patients in a certain hospital were collected. The regressor $x_1$ is the (estimated) mean of daily consumption of cigarettes. Fit the data to a logistic regression model for the proportion of coronary thrombosis hospital patients as a function of the amount of smoking.

(On the web-page http://statpages.org/logistic.html there is a logistic regression calculator.)

6. The F test statistic of the regression model is sometimes presented in the form

$$F = \frac{(n - k - 1)R^2}{k(1 - R^2)} = \frac{(n - k - 1)R^2_{\text{adj}} + k}{k(1 - R^2_{\text{adj}})}.$$ 

Show that it’s correct.
1. The resistance of a certain component is claimed to have a median less than 0.150 \( \Omega \). (In other words, the probability that the resistance is < 0.150 \( \Omega \) is at least 0.5.) When \( n = 8 \) randomly chosen components were measured, the following results were obtained (ohms):

\[
0.148, 0.147, 0.151, 0.146, 0.151, 0.148, 0.147, 0.150.
\]

Use the sign test at the risk level \( \alpha = 0.05 \) to decide whether or not the claim is true.

(The web-page http://www.graphpad.com/quickcalcs/binomial1.cfm has a nice sign test calculator.)

2. The table shows the weights of \( n = 5 \) people (kilograms) before they quit smoking and five months after quitting. Test with the signed rank test at the risk level \( \alpha = 0.05 \) whether weight increases after quitting.

(The web-page http://faculty.vassar.edu/lowry/vsord.html has calculators for nonparametric tests.)

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>66</td>
<td>71</td>
</tr>
<tr>
<td>2</td>
<td>80</td>
<td>82</td>
</tr>
<tr>
<td>3</td>
<td>69</td>
<td>68</td>
</tr>
<tr>
<td>4</td>
<td>52</td>
<td>56</td>
</tr>
<tr>
<td>5</td>
<td>75</td>
<td>73</td>
</tr>
</tbody>
</table>

3. Two lubricants A and B were compared by using them in a certain type of bearings until it overheated. The times (hours) according to the table below were obtained. Test at the risk \( \alpha = 0.05 \) with the Mann-Whitney test whether the lubricant A is more durable (in the sense of medians), as it is claimed.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>427</td>
<td>463</td>
</tr>
<tr>
<td>1</td>
<td>542</td>
<td>390</td>
</tr>
<tr>
<td>2</td>
<td>491</td>
<td>361</td>
</tr>
<tr>
<td>3</td>
<td>201</td>
<td>419</td>
</tr>
<tr>
<td>4</td>
<td>468</td>
<td>424</td>
</tr>
<tr>
<td>5</td>
<td>673</td>
<td>345</td>
</tr>
<tr>
<td>6</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

4. The table shows the number of hours of use per single charge measured for a certain consumer device with three different types of rechargeable batteries A, B and C. Test with the Kruskal-Wallis test whether the medians of times of use of different types are equal at the risk level \( \alpha = 0.01 \).

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>4.9</td>
<td>5.5</td>
<td>6.4</td>
</tr>
<tr>
<td>1</td>
<td>6.1</td>
<td>5.4</td>
<td>6.8</td>
</tr>
<tr>
<td>2</td>
<td>4.3</td>
<td>6.2</td>
<td>5.6</td>
</tr>
<tr>
<td>3</td>
<td>4.6</td>
<td>5.8</td>
<td>6.5</td>
</tr>
<tr>
<td>4</td>
<td>5.2</td>
<td>5.5</td>
<td>6.3</td>
</tr>
<tr>
<td>5</td>
<td>–</td>
<td>4.8</td>
<td>6.6</td>
</tr>
</tbody>
</table>

5. This is about calculating the rank correlation coefficient when the sample size is \( n \).

a) What is the difference between the sums of squares of the sequence numbers \( \sum_{i=1}^{k} r_{1,i}^2 \) in the cases

(1) there are no duplicate values,
(2) there’s exactly one duplicate value?

b) What’s the difference between the cases

(1) there are no duplicate sample values,
(2) there’s exactly one triplicate value?

c) What’s the difference between the cases

(1) there are no duplicate sample values,
(2) all the sample values are the same?

(The moral is that no matter how much you’d like to, you shouldn’t use the “easy” formula

\[
r_{S} = 1 - \frac{6}{n(n^2 - 1)} \sum_{i=1}^{n} d_i^2
\]

to calculate the Spearman’s rank correlation coefficient unless neither sample has duplicate values.)