1. Mean and standard deviation

5.7) Calculate $E\left[e^X\right]$ if $f_X(x) = 2e^{-2x}$, $x \geq 0$. □

5.12) The r.v. X has density function $f_X(x) = 3x^{-4}$, $x \geq 1$. Calculate its mean and variance. □

5.13) The r.v. X has the density function $f_X(x) = 2x$ for $0 \leq x \leq 1$.
a) Calculate the mean μ and standard deviation σ for X.
b) Calculate $P(\mu - 2\sigma < X < \mu + \sigma)$.
c) Calculate $P(\mu - \sigma < X < \mu + 2\sigma)$. □

5.22) The random variables X_1, X_2 and X_3 are independent, all with mean 2 and standard deviation 3. Let $Y = 3X_1 - 2X_2 + X_3 - 6$. Find $E[Y]$ and $D[Y]$. □

6.1) X is $N(0, 1)$. Find
a) $P(X \leq 1.82)$,
b) $P(X \leq -0.35)$,
c) $P(-1.2 < X < 0.5)$,
d) a such that $P(X > a) = 5$ %,
e) a such that $P(|X| < a) = 95$ %. □

5.5) The discrete random variable X has the probability mass function $p_X(k) = 6k^{-2}/\pi^2, k = 1, 2, ...$ Show that $E(X)$ doesn't exist. □

6.9) In a packaging machine, packets of margarine is cut so that the weight (unit: kg) is a r.v. X. Suppose that with good approximation it is reasonable to suppose that X is $N(0.5, 0.003)$. What is the probability that a packet of margarine weighs at least 495 grams? Also find limits $0.5 \pm d$ such that in the long run
a) 50 %
a) 95 %
a) 99 %
of all packets of margarine has a weight between $0.5 - d$ and $0.5 + d$. □

ANSWERS

5.7 2.

5.12 3/2 respectively 3/4.

5.13 a) 2/3 respectively $1/\sqrt{18} = \sqrt{2}/6$; b) $6\mu\sigma - 3\sigma^2 = 2\sqrt{2}/3 - 1/6$;
c) $1 - (\mu - \sigma)^2 = 1/2 + 2\sqrt{2}/9$.

5.22 $E[Y] = -2$, $D[Y] = 3\sqrt{14} \approx 11.2$.

6.1 a) $\Phi(1.82) \approx 0.966$; b) $1 - \Phi(0.35) \approx 0.363$ c) $\Phi(0.5) + \Phi(1.2) - 1 \approx 0.576$;
d) $\lambda_{0.05} \approx 1.64$; e) $\lambda_{0.025} \approx 1.96$.

6.9 0.9522 a) 0.002; b) 0.006; c) 0.008.

2. Parameter estimates

1∼) A computer manufacturer has received a very large shipment of electrical components. The lifetimes of each component measured in years are independent and exponential-distributed with mean 5 years. It is suspected that the seller mixed in an unknown part, a, of components which have independent lifetimes exponentially distributed with mean 1. Thus the lifetime of a randomly selected component has the density function

$$f(x) = \frac{1}{5}e^{-x/5}(1-a) + e^{-x}a, \quad \text{for } x \geq 0.$$

We wish to estimate a.

(a) Method 1: We randomly pick n_1 units of the shipment and measure the lifetimes $x_1, \ldots, x_{n_1}$. Estimate a using the method of moments.

(b) Method 2: We randomly pick n_2 units of the shipment, use them for half a year and find that y units have stopped working during these 6 months. Estimate a in a suitable manner.

(c) Is there any practical benefit in using method 2? □

2∼) A certain type of transistors have a lifetime that follows an exponential distribution. We put 400 of them to use and note that after one unit of time only 109 works. Estimate the mean lifetime and median lifetime. □

3∼) Let $x_1, \ldots, x_n$ be a sample with density function

$$f(x;\theta) = \begin{cases} \frac{1}{1-\theta} & \theta < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Calculate the moment estimator of θ. Investigate if the estimator is unbiased. □

10.1) The caries index was calculated for 22 people, i.e. the number of tooth surfaces attacked by caries among the 100 tooth surfaces you get if you disregard the wisdom teeth and the lingual surface (surface against the tongue) of all teeth. Result:

41	47	66	73	48	52	49	54	61	62	47
52	65	61	69	31	54	53	50	47	36	69

Calculate mean, median, variance, standard deviation and coefficient of variation for this data. □

10.4) In two materials $x_1, x_2, \ldots, x_{10}$ respectively $y_1, y_2, \ldots, y_5$ the means and standard deviations were calculated with the results

$$\bar{x} = 5313, \quad s_x = 5.2, \quad \bar{y} = 5309, \quad s_y = 3.0.$$

If the 15 numbers had been considered as a single material, what mean and standard deviation would have been obtained? □

11.6a) We have a random sample $x_1, x_2, \ldots, x_n$ from $N(\mu, \sigma)$ where μ and σ are unknown. We make the estimates

$$\mu^*_{\text{obs}} = \bar{x} = (x_1 + x_2 + \ldots + x_n)/n \quad \text{and} \quad \hat{\mu}_{\text{obs}} = (x_1 + x_n)/2.$$

a) Show that both estimates are unbiased.
b) Which of the estimates μ^*_{obs} and $\hat{\mu}_{\text{obs}}$ is the most effective? □

11.8) Let θ^*_{obs} and $\hat{\theta}_{\text{obs}}$ be two independent unbiased estimates of θ with known variances σ_1^2 respectively σ_2^2.
a) Show that $\tilde{\theta}_{\text{obs}} = a\theta^*_{\text{obs}} + (1-a)\hat{\theta}_{\text{obs}}$ is an unbiased estimate of θ for all numbers a.
b) For what value of a do you get the most effective estimate? □

11.9) With a certain method for performing measurements, the distance between two points are measured. The result (unit: meters)

1456.3 1458.5 1457.7 1457.2

can be viewed as a random sample from $N(\mu, \sigma)$ where μ is the real distance and the standard deviation σ is a measure of the precision of the method. Calculate an unbiased estimate of the variance σ^2 if
a) you know that $\mu = 1457.0$,
b) μ is unknown.
c) Is the assumption of normal distribution essential for the unbiasedness? □

ANSWERS

1 ∼ a) $\hat{a} = \frac{5-\bar{x}}{4}$; b) $\hat{a} = \left(\frac{y}{n_2} + e^{-1/10} - 1\right) \Big/ (e^{-1/10} - e^{-1/2}) \approx \left(\frac{y}{n_2} - 0.0952\right) \Big/ 0.2983$; c) In practice, Metod 2 is much more useful because: in Metod 1 in order to get $\bar{x}$; one needs to collect the lifetime of each element, which means that one has to wait until all elements break(this may take along time); in Metod 2, in order to get y one just needs to wait for a half year.

2 ∼ $\hat{\mu} = \frac{1}{\ln(400/109)} \approx 0.77$ $\hat{M} = \mu \ln 2 \approx 0.77 \ln 2 \approx 0.53$.

3 ∼ $\hat{\theta} = 2\bar{x} - 1$. It is unbiased.

10.1 Mean $\bar{x} = 53.95$, median $\tilde{x}_{0.50} = 52.5$, variance $s^2 = 118.71$, standard deviation $s = 10.90$ and coefficient of variation $c_v = 0.2 = 20\%$.

10.4 Mean 5312, standard deviation 4.87.

11.6 b) μ^*_{obs} is more efficient when $n > 2$.

11.8 b) $a = \sigma_2^2/(\sigma_1^2 + \sigma_2^2)$.

11.9 a) 0.82; b) 0.85; c) No.

3. ML-estimation

11.10) The discrete random variable X has the probability function $p_X(k) = \theta(1-\theta)^{k-1}$ for $k = 1, 2, 3, \ldots$, where $0 < \theta < 1$. We have a random sample $4, 5, 4, 6, 4, 1$ from this distribution.
a) Write down the likelihood function $L(\theta)$.
b) For which θ is $L(\theta)$ largest? That is, what is the ML-estimator of θ? □

11.11) The random variable X has density function $f_X(x) = \theta(1+x)^{-\theta-1}$ for $x \geq 0$. It is known that the parameter θ is $2, 3$ or 4. We have a random sample from this distribution with the two values 0.2 and 0.8.
a) Find the values of the L-function for the three different values of θ.
b) Find the ML-estimate of θ. □

11.12) The number of calls X to a telephone switch during the most busy time of the day is $Po(\mu)$. The number of calls during different days are independent. During a period of n days $x_1, \ldots, x_n$ calls have been registered.
a) Find the ML-estimate of μ.
b) Find the mean and standard deviation of the estimate.
c) Estimate μ if during a period of 8 days the following number of calls have been received:

115 82 108 106 118 87 99 92

d) Calculate the standard error for the estimate in c). □

11.13) The time between errors in a complicated technical equipment can be thought of as independent and Exp(λ). We registered times between errors: $x_1, \ldots, x_n$.
a) Find the ML-estimate of λ. □

11.14) We have a random sample $x_1, x_2, \ldots, x_n$ from a distribution with density function $f_X(x) = \theta x^{\theta-1}$ for $0 < x < 1$. Find the ML-estimate of θ. □

11.15) The r.v. X is Rayleigh-distributed with density function

$$f_X(x) = (x/a)e^{-x^2/2a} \quad \text{for } x \geq 0,$$

where a is an unknown positive parameter. We have a sample $x_1, x_2, \ldots, x_n$ from this distribution. Find the ML-estimate of a. □

11.16) A measuring error of glitch type has a uniform distribution on $[-\theta, \theta]$. Suppose that while measuring known quantities we got the measuring errors $x_1, \ldots, x_n$.
a) Find the ML-estimator of θ.
b) Is the ML-estimator unbiased?
Hint: Find the density function of the estimator by first calculating the distribution function. □

11.22) To test a measuring equipment the following three series of measurement were done on different substances:

Serie 1	0.16	0.18	0.19	0.18	0.21
Serie 2	0.22	0.21	0.24	0.24	0.25
Serie 3	0.17	0.15	0.15	0.18	0.14

The result of a series of measurements is though of as a random sample from a normal distribution with unknown mean and with a variance that is unknown but the same for the three different distributions.
a) Estimate the means for the distributions.
b) Estimate the joint standard deviation. □

11.23) We have an observation $x = 16$ of $X \in \text{Bin}(25, p)$.
a) Estimate p.
b) Find the standard deviation for the estimator.
c) Estimate the standard error of the estimator. □

11.25) The number of ships that passes Helsingborg on the way south through Östersund under a time interval of length t is thought to be $\text{Po}(\lambda t)$. The number of ships in disjoint intervals is thought to be independent. A person wants to estimate λ and counts the number of passing ships during three different time periods.

Observation	30	30	40
Number of ships	10	12	18

a) Find the ML-estimate of λ.
b) How big is the standard deviation for the estimator? □

11.28) To find the properties of a packing machine, the contents of 12 cans packed by the machine has been weighed. The cans are supposed to contain 250g.

248.0	250.6	249.8	251.6	247.2	251.0
248.8	249.0	254.4	248.4	253.5	251.5

The weights can be thought of as independent observations of a $N(\mu, \sigma)$-distributed random variable. Find the ML-estimates of μ and σ. □

ANSWERS

11.10 a) $L(\theta) = \theta(1-\theta)^{4-1} \cdot \ldots \cdot \theta(1-\theta)^{1-1} = \theta^6(1-\theta)^{18}$;
b) $L(\theta)$ is largest when $\theta = 1/4$; this is the ML-estimate.

11.11 a) $L(\theta) = \theta^2 \cdot 2.16^{-\theta-1}$; $L(\theta) = 0.40,\ 0.41$ and 0.34 for $\theta = 2,\ 3$ respectively 4;
b) $L(\theta)$ is largest when $\theta = 3$; this is the ML-estimate.

11.12 a) $\mu^*_{\text{obs}} = \bar{x}$; b) $E[\mu^*] = \mu$, $D[\mu^*] = \sqrt{\mu/n}$; c) $\mu^*_{\text{obs}} = 100.9$; d) 3.55.

11.13 a) $\lambda^*_{\text{obs}} = 1/\bar{x}$; b) $d(\lambda^*) = 1/(\bar{x}\sqrt{n})$.

11.14 The ML-estimate is $\theta^*_{\text{obs}} = -n/\sum \ln x_i$.

11.15 The ML-estimate is $a^*_{\text{obs}} = \sum x_i^2/(2n)$.

11.16 a) $\theta^*_{\text{obs}} = \max(|x_1|, \ldots, |x_n|)$; b) No.

11.22 a) 0.184, 0.232 respectively 0.158; b) 0.017.

11.23 a) $p^*_{\text{obs}} = 0.64$; b) $D[p^*] = \sqrt{p(1-p)/25}$; c) $d(p^*) = 0.096$.

11.25 a) The ML-estimate is $\lambda^*_{\text{obs}} = 0.40$; b) $\sqrt{\lambda}/10$.

11.28 $\mu^*_{\text{obs}} = 250.32$, $\sigma^*_{\text{obs}} = 2.134$.

4. Confidence intervals

12.1) The random variable X is χ^2-distributed with 24 degrees of freedom. Find numbers a, b, c such that $P(X < a) = 95\%$ and $P(b < X < c) = 95\%$. □

12.4) The random variable X is t-distributed with f degrees of freedom.
a) For $f = 9$ find numbers a, b such that $P(|X| \leq a) = 99\%$ and $P(X > b) = 5\%$.
b) Using a table show that $t_\alpha(f) \to \lambda_\alpha$ as $f \to \infty$. How large f is needed for $t_{0.025}(f) = \lambda_{0.025}$ if both numbers are rounded to one decimal? □

12.10) We have a random sample from $N(\mu, 2)$:

44.3 45.1 46.1 45.3.

Find a 95% confidence interval for μ. □

12.11) Four measurements have been made on a solution with unknown pH-value μ.

8.24 8.18 8.15 8.23.

Model: The pH-meter has a systematic error Δ and a random error that is $N(0, \sigma)$. The four measurements are thus a sample from $N(\mu + \Delta, \sigma)$. Furthermore it is known that $\Delta = 0.10$ and $\sigma = 0.05$. Find a confidence interval for μ with confidence level 99%. □

12.18) Brazilian pine is delivered to Sweden in the form of planks of standard width and varying length for usage in e.g. ceilings. The lengths are varying as independent normal distributed random variables. Sixteen random planks had the following lengths (meter):

5.8 5.9 5.1 3.5 4.2 4.9 5.3 5.3 4.7 3.9 4.5 4.1 4.0 4.2 4.7 4.8.

a) Find a 95% confidence interval for the expected length of a plank (= the mean length of all planks).
b) Find a 95% confidence interval for the standard deviation. □

12.19) In many contexts only an upper limit of the standard deviation σ is needed.
a) Find a 95% upper limited confidence interval for σ in the case of n observations from a normal distributed random variable.
b) When manufacturing screws the variations in the diameters of the heads of the screws must not be too large. The diameter has been measured for 50 screws and $s = 0.021$ mm has been calculated. Give an upper limit of the true standard deviation, σ, that is correct with 95% certainty. □

12.23) Four measurements have been made on a solution with unknown pH-value μ.

4.32 4.22 4.23 4.37.

Furthermore with the same meter, six measurements have been made on a solution with the known pH-value 4.84:

4.71 4.63 4.69 4.76 4.58 4.83.

Model: The pH-meter has a systematic error Δ and a random error that is $N(0, \sigma)$; Δ and σ are unknown. The measurement on a solution with pH-value a is thus a sample from $N(a + \Delta, \sigma)$.
a) Find an estimate μ^*_{obs} of μ.
b) Find $D[\mu^*]$.
c) Estimate the standard error $d(\mu^*)$.
d) Find a 95% confidence interval for μ. □

ANSWERS

12.1 $a = 36.4$, $b = 12.4$, $c = 39.4$, where symmetry has been used in order to obtain b and c.

12.4 a) $a = 3.25$, $b = 1.83$; b) $f = 30$.

12.10 $I_\mu = (43.2, 47.2)$.

12.11 $I_\mu = (8.04, 8.16)$.

12.18 a) $\bar{x} \mp t_{\alpha/2}(n-1) \cdot s/\sqrt{n} = 4.86 \mp 0.36 \quad (s = 0.6853)$;
b) $\left(s \cdot \sqrt{(n-1)/\chi^2_{\alpha/2}(n-1)}, s \cdot \sqrt{(n-1)/\chi^2_{1-\alpha/2}(n-1)}\right) = (0.51, 1.06)$.

12.19 a) $\left(0, s \cdot \sqrt{(n-1)/\chi^2_{0.95}(n-1)}\right)$;
b) $\sigma < 0.025$ (interpolation in table gives $\chi^2_{0.95}(49) = 33.97$).

12.23 a) $\mu^*_{\text{obs}} = 4.425$; b) $D[\mu^*] = \sigma\sqrt{5/12}$; c) $d(\mu^*) = 0.054$; d) $I_\mu = (4.30, 4.55)$.

5. Two samples

4∼) Surface irregularity has been measured for four materials used for encapsulation. Result:

Material	Surface irregularity						$\bar{x}_i$	s_i
EC10	0.50	0.55	0.55	0.36			0.4900	0.0898
EC10A	0.31	0.07	0.25	0.18	0.56	0.20	0.2617	0.1665
EC4	0.20	0.28	0.12				0.2000	0.0800
EC1	0.10	0.16					0.1300	0.0424

Do pairwise comparisons between the materials by constructing suitable confidence intervals, each with confidence level 0.99. You may assume that the data come from normal distributions with the same variance. Low surface irregularity is preferred. □

5∼) To compare the effects of three different types of blood pressure lowering medicines, three groups each with 10 people were treated with the different medicines. After a month the decrease in blood pressure was measured. Result:

Material	Mean $\bar{x}_i$	Standard deviation s_i
Medicine 1:	17.3	6.19
Medicine 2:	21.1	7.26
Medicine 3:	10.8	5.23

Model: We have three samples from $N(\mu_i, \sigma)$, $i = 1, 2, 3$.
a) Construct a confidence interval for σ of type $(0, a)$ with confidence level 0.95.
b) Is it likely that $\mu_2 > 1.4\mu_3$? Answer the question by constructing a suitable confidence interval with confidence level 0.95. □

6∼) A small industry has three different ovens for heating metal objects. It has been assumed that the ovens have the same temperature setting but it seems to not be the case. Repeated independent measurements of the temperatures are made for each oven with the following results:

						$\bar{x}_i$	s_i
Oven 1:	492.4	493.6	498.5	488.6	494.0	493.42	3.55
Oven 2:	488.5	485.3	482.0	479.4		483.80	3.96
Oven 3:	502.1	496.8	497.5			498.80	2.88

Model: We have three samples from $N(\mu_i, \sigma)$ where μ_i is the temperature setting for oven i.
a) Construct a 95% upper bounded confidence interval for σ.
b) Perform pairwise comparisons between the μ_is by constructing suitable confidence intervals, each with confidence level 98%. □

7∼) In a sample of 500 units from a very large supply, 87 were found to be defect. Find a 95% confidence interval for the proportion of faulty units. □

8∼) The service times for a queue system follows an exponential distribution with mean μ. 80 different service times have been observed with mean $\bar{x} = 4.5$ minutes.
a) Find a confidence interval for μ with approximate confidence level 95%.
b) Let p be the probability that a service time is greater than 10 minutes. Construct a confidence interval for p with approximate confidence level 0.95. □

9∼) A company has a warehouse where goods are transported by forklifts. At 500 different randomly chosen intervals of length one hour the number of forklifts arriving was observed. Result:

Number of forklifts	0	1	2	3	4	5	6	7	8
Frequency	52	151	130	102	45	12	5	1	2

Model: $x_1, \ldots, x_{500}$ are observations from $Po(\mu)$. Find a confidence interval for μ with approximate confidence level 0.95. □

12.21) Superelevation is a sometimes critical property of concrete units. A study was made to see if there were any difference in this regard concerning units from two different factories A and B. In this study 9 and 16 random samples were made from the production of factory A respectively factory B. The observations can be thought as independent samples from independent random variables X and Y that are $N(\mu_A, \sigma)$ respectively $N(\mu_B, \sigma)$. The following values were obtained:

A: $\bar{x} = 18.1 \quad s_1 = 5.0 \quad n_1 = 9$
B: $\bar{y} = 14.6 \quad s_2 = 7.1 \quad n_2 = 16.$

Find a 99% confidence interval for $\mu_A - \mu_B$. □

12.22) A study was made to see the change in blood pressure (unit: mm Hg) from using a certain substance. The blood pressure of 10 people was measured after which each person was given a certain dose of the substance (same dose for everyone). After 20 minutes the blood pressure of each person was measured again. Model: The result before and after on person number i is $N(\mu_i, \sigma_1)$ and $N(\mu_i + \Delta, \sigma_2)$ respectively.
a) Interpret the parameters $\mu_1, \mu_2, \ldots, \mu_{10}$ and Δ.
b) The results were:

Person no.	1	2	3	4	5	6	7	8	9	10
Blood pressure before:	75	70	75	65	95	70	65	70	65	90
Blood pressure after:	85	70	80	80	100	90	80	75	90	100

Find a 95% confidence interval for Δ. □

12.25) a) To see if a certain medicine has as primary side effect to increase a certain liver value, this value was measured on 50 people that had not been treated with the medicine (measurements $x_1, \ldots, x_{50}$), and on 25 people that had been treated with the medicine (measurements $y_1, \ldots, y_{25}$). It was obtained:

$\bar{x} = 148.2 \quad \bar{y} = 151.7$
$s_x = 10.0 \quad s_y = 8.0$

Find a 95% confidence interval for the difference of expected liver value for the two groups. Write down all assumptions made about distribution and independence.
b) The result of the study in a) was bad in the sense that the confidence interval was too wide to be able to draw any interesting conclusions. A consulting statistician suggested a new trial were the liver value before and after would be measured on 25 patients (measurements x_i respectively y_i, $i = 1, \ldots, 25$). It was obtained:

$\bar{x} = 149.0 \quad \bar{y} = 150.9$
$s_x = 8.1 \quad s_y = 9.5 \quad s_z = 1.6$ where $z_i = y_i - x_i,\ i = 1, \ldots, 25.$

Find a 95% confidence interval for the difference between expected liver value before and after the treatment. Write down all assumptions made about distribution and independence. □

12.27) The samples $x_1, \ldots, x_6$ and $y_1, \ldots, y_{12}$ are from $N(\mu_1, \sigma)$ respectively $N(\mu_2, \sigma)$, where μ_1, μ_2, and σ are unknown. Mean and variance of the samples are $\bar{x} = 49.2$, $s_x^2 = 8.80$ respectively $\bar{y} = 37.4$, $s_y^2 = 3.04$. Find a 90% confidence interval for $\mu_1 - \mu_2$. □

12.28) With the help of a random sample with five observations from $N(\mu_1, \sigma_1)$, where μ_1 is unknown but σ_1 is known, a 95% confidence interval for μ_1 has been constructed in the usual way with the result $(1.37, 1.53)$. In the same manner, with the help of a random sample with seven observations from $N(\mu_2, \sigma_2)$, where μ_2 is unknown but σ_2 is known, the interval $(1.17, 1.29)$ was obtained as a 95% confidence interval for μ_2. Find a 95% confidence interval for $\mu_1 - \mu_2$. □

12.30) A chemist is studying the amount of pollution in Motala ström. Among other things he is interested in the pollution coming from a certain industry along the stream. During a period of 70 different days he takes 30 samples upstream and 40 samples downstream from the industry in question, measuring the amount of a certain pollutant in these samples. Since the samples were taken on different days they can be assumed to be independent. The following data was obtained:

	Mean	Standard deviation
Upstream	13.2	2.8
Downstream	86.1	38.7

The measurements downstream was sometimes small, 10-15, and sometimes very large, 80-150, which means that the observations can't be assumed to come from a normal distribution. Find an approximately 95% confidence interval that can be used to estimate the pollution from the industry in question. □

12.31) From a large batch of goods 600 units were selected. Out of these, 24 units were found to be defect. Construct a confidence interval with an approximate confidence level of 95% for
$p =$ fraction of broken units in the batch. □

12.32) (Cont. from problem 12.31) How large sample is needed to be able to with 95% confidence level be able to estimate p with an error less than 0.005 if
a) p is unknown?
b) it is known that $0 < p < 0.04$? □

12.33) At a comparison of two opinion surveys it was seen that out of 1704 participants 46.5% supported the middle class parties in October. In November, 45.6% out of 1689 participants supported the middle class parties. Find a confidence interval with approximate confidence level 95% for the change of the support proportion between the two surveys. The number of eligible voters can be thought of as infinite compared to the sample sizes. □

12.34) Let p denote the relative frequency of defect units among the 100000 units of a certain kind in a storage. 1000 units were randomly selected for testing and 36 of these were found to be broken.
a) Find a 95% two-sided confidence interval for p.
b) Find a 95% confidence interval for the total number of defect units in the supply. □

12.36) The traffic at a certain location can be modeled: The number of cars passing a certain point during a time of length t (unit: min) is $Po(\lambda t)$. At a traffic count it was found that 400 cars passed in 10 minutes.
a) Find an approximately 95% confidence interval for 10λ.
b) Find an approximately 95% confidence interval for λ. □

12.37) The number of calls X to a telephone operator during the most busy time of the day is $Po(\mu)$. During a period of 8 days the following observations of X were made:

115 82 108 106 118 87 99 92

Find an approximately 95% confidence interval for μ. □

ANSWERS

4 ∼ $I_{\mu_i-\mu_j} = (\bar{x}_i - \bar{x}_j) \mp t_{0.005}(df) \cdot s \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$. Here $s = 0.1270$, $df = 11$. It follows
$I_{\mu_1-\mu_2} = (-0.03, 0.48)$, $I_{\mu_1-\mu_3} = (-0.01, 0.59)$, $I_{\mu_1-\mu_4} = (0.02, 0.70)$.
We can conclude $\mu_1 > \mu_4$ (Material 1 is worse than Material 4), the other results are not significant.

5 ∼ a) $I_\sigma = (0, \sqrt{65.76}) = (0, 8.11)$;
b) $I_{\mu_2-1.4\mu_3} = \left(\bar{x}_2 - 1.4\bar{x}_3 \mp t_\alpha(n_1+n_2+n_3-3) \cdot s \cdot \sqrt{\frac{1^2}{n_2} + \frac{1.4^2}{n_3}}, \infty\right) = (0.17, \infty)$.
Since $0.17 > 0$ we can conclude $\mu_2 > 1.4\mu_3$.

6 ∼ a) $I_\sigma = (0, 5.86)$;
b) $I_{\mu_1-\mu_2} = (2.89, 16.35)$, $I_{\mu_1-\mu_3} = (-12.71, 1.95)$, $I_{\mu_2-\mu_3} = (-22.67, -7.33)$.
We can conclude $\mu_1 > \mu_2$ and $\mu_3 > \mu_2$. Difference between μ_1 and μ_3 is not significant.

7 ∼ $I_p = \hat{p} \mp \lambda_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = (0.141, 0.207)$.

8 ∼ a) $I_\mu = (3.69, 5.76)$; b) $I_p = (0.067, 0.176)$.

9 ∼ $I_\mu = \bar{x} \mp \lambda_{\alpha/2}\sqrt{\bar{x}/n} = 2.02 \mp 1.96\sqrt{2.02/500} = (1.90, 2.14)$.

12.21 Two independent samples:
$I_{\mu_A-\mu_B} = \bar{x} - \bar{y} \mp t_{\alpha/2}(n_1+n_2-2) \cdot s\sqrt{1/n_1 + 1/n_2} = (-4.05, 11.05)$ $(s = 6.4476)$.

12.22 b) $I_\Delta = (5.5, 16.5)$, it seems the substance increases the blood pressure.

12.25 a) Two samples:
$x_1, \ldots, x_{50}$ are observations on $X_i \in N(\mu_1, \sigma_1)$, $i = 1, \ldots, 50$,
$y_1, \ldots, y_{25}$ are observations on $Y_i \in N(\mu_2, \sigma_2)$, $i = 1, \ldots, 25$.
The two samples are assumed independent
$I_{\mu_y-\mu_x} = \bar{y} - \bar{x} \mp t_{\alpha/2}(n_1+n_2-2) \cdot s\sqrt{1/n_1 + 1/n_2} = 3.5 \mp 4.58$ where $s = 9.3896$;

b) Observations in pairs:
$x_1, \ldots, x_{25}$ are observations on $X_i \in N(\mu_i, \sigma_1)$, $i = 1, \ldots, 25$,
$y_1, \ldots, y_{25}$ are observations on $Y_i \in N(\mu_i + \Delta, \sigma_2)$, $i = 1, \ldots, 25$,
$z_i = y_i - x_i$ are observations on $Z_i \in N(\Delta, \sigma)$, $i = 1, \ldots, 25$,
$I_\Delta = \bar{z} \mp t_{\alpha/2}(n-1) \cdot s_z/\sqrt{n} = 1.9 \mp 0.66$.

12.27 $I_{\mu_1-\mu_2} = (9.9, 13.7)$.

12.28 $I_{\mu_1-\mu_2} = (0.12, 0.32)$.

12.30 $I_{\mu_2-\mu_1} = \bar{y} - \bar{x} \mp \lambda_{\alpha/2} \cdot \sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}$. Pick e.g. $\alpha = 0.05$ to get $I_{\mu_2-\mu_1} = 72.9 \mp 12.0$.

12.31 $I_p = (0.024, 0.056)$.

12.32 a) $n \approx 40000$; b) $n \approx 6000$.

12.33 $I_{\text{change}} = -0.009 \mp 0.034 = (-0.043, 0.025)$.

12.34 a) $I_p = (0.024, 0.048)$; b) $I_{\text{total number}} = (2400, 4800)$.

12.36 a) $I_{10\lambda} = (361, 439)$; b) $I_\lambda = (36.1, 43.9)$.

12.37 $I_\mu = \bar{x} \mp \lambda_{\alpha/2} \cdot \sqrt{\bar{x}/n} = 100.9 \mp 7.0$.

6. Hypothesis testing, P-value method

10∼) A certain measuring equipment measures the radioactive background radiation at a location. It is reasonable to suppose that the number of registered particles during t minutes is $Po(\lambda t)$ where $\lambda = 5$ (unit: min^{-1}). After a radioactive spill it is suspected that the radiation has increased. For how long is it needed to measure the radiation to test

$$H_0 : \lambda = 5 \quad \text{vs} \quad H_1 : \lambda > 5$$

on the level $\alpha = 0.01$ with a test having a power of 0.99 if the intensity of radiation is $\lambda = 7.5$. □

13.4) A certain type of light bulbs have a life time (unit: hours) that is an exponential r.v. X with mean θ. The manufacturer claims that $\theta = 1000$. Per doubts that θ is that large. He intends to try the null hypothesis $H_0 : \theta = 1000$ by buying a light bulb and observe its life time x_1. If x_1 is small, say $x_1 < a$, he will reject H_0. To find a he solves, for $\theta = 1000$, the equation $P(X < a) = \alpha$, where α is the chosen significance level.
a) Find a as a function of α.
b) Suppose that Per gets $x_1 = 75$. Is this significant on the 5%-level?
c) Same question for $x_1 = 50$.
d) Suppose $x_1 = 45$. Test the hypothesis H_0 with the P-value method. □

13.5) Per plays on a slot machine that gives a prize with the unknown probability p. The number of games X including the one when the first prize is awarded has the probability function

$$p_X(k) = p(1-p)^{k-1}, \quad \text{for } k = 1, 2, 3, \ldots$$

It is claimed that $p = 0.2$ but Per doubts that p is that large and wishes to test the hypothesis $H_0 = p = 0.2$ against $H_1 : p < 0.2$. Is it possible, using the significance level 0.10, to reject H_0 if he loses the first ten games and wins on the eleventh? Use the P-value method. □

13.24) A person claiming to be able to find water using a dowsing rod was tested in the following manner. He was brought to a place containing 10 covered containers at great distance from each other and was told that five of the containers contained water while five were empty. He identified four of the five water filled containers correctly and one wrong. Test the hypothesis that his dowsing rod had no effect, i.e. he was guessing. Find the explicit P-value. □

13.25) The r.v. X is $Po(\mu)$. Using 50 observations one wishes to test the hypothesis $H_0 : \mu = 0.2$ against $H_1 : \mu > 0.2$. The sum of the observations is 19. Can H_0 be rejected? Use the P-value method.
Hint: The sum of n observations from $Po(\mu)$ is distributed $Po(n\mu)$. □

13.26) A random sample of n values are to be taken from $Po(\mu)$. Using this sample the null hypothesis $H_0 : \mu = 4$ against the alternative hypothesis $H_1 : \mu = 5$ in a way such that the probabilities for error of the 1st and 2nd kind are 0.001 and 0.01 respectively. How should n be chosen and how should the test be designed? □

ANSWERS

10 $\sim$ $t = 21.428$.

13.4 a) $a = -1000\ln(1-\alpha)$; b) No, since $x_1 = 75$ is larger than $a = -1000\ln(0.95) = 51.3$; c) Yes;
d) $P = P(X \leq 45, \text{ when } \theta = 1000) = e^{-45/1000} = 0.044$. This is less than 0.05 and thus the result is significant on the 0.05 level.

13.5 Since H_1 contains small p-values, Per should reject H_0 if it takes a long time until he wins. We have $P = \sum_{k=11}^{\infty} 0.2 \cdot 0.8^{k-1} = 0.107$. Since $P > 0.10$, H_0 can not be rejected.

13.24 $26/252 = 0.103$.

13.25 $P = \sum_{k=19}^{\infty} \frac{10.0^k}{k!} e^{-10.0} = 0.0072$ (can use table). H_0 is rejected on the 0.01 level.

13.26 $n = 130$. Reject H_0 if the sum of the 130 observations is larger than $520 + 3.09\sqrt{520} = 590.5$

7. Hypothesis testing

12∼) Tensile strength for three different lines have been measured. The result:

Type	Measured values								$\bar{x}_i$	s_i
A:	3.72	2.93	4.73	3.90	4.57	4.27	5.38	3.24	4.09	0.809
B:	13.27	16.48	9.54	16.08	20.57	15.87	13.57	9.63	14.38	3.70
C:	10.42	11.98	11.50	7.85	5.71				9.49	2.65

Is it possible to determine which line is the best?
a) While answering the question we cannot use a model where we view the data as three samples from normal distributions with the same standard deviation. Show this with the help of suitable tests each on the level 0.01. It is enough if you write one of the tests.
b) Instead suppose that taking the logarithms of the tensile strengths gives three samples from $N(\mu_i, \sigma)$ and try to answer the above question by constructing suitable confidence intervals each with confidence level 95%. Before the measurements were made there was no information on what kind would be better. □

13∼) 25 independent measurements of a quantity has yielded the results

$$\bar{x} = 11.2 \quad s = 2.1$$

and we suppose $X_i \sim N(\mu, \sigma)$.
a) Try the hypothesis $H_0 : \mu = 10$ against $H_1 : \mu \neq 10$ on the significance level 1%.
b) Find a 99% confidence interval for μ. □

13.8) A medicine manufacturer sometimes use a certain coloring additive. One wishes to see how the coloring additive affects the color of the produced medicines. Therefore ten packages of medicine are chosen randomly and the muddiness was measured after time of storage. Result:

3.9 4.1 4.4 4.0 3.8 4.0 3.9 4.3 4.2 4.4

Without a coloring additive the muddiness is on average 4.0. One now wonders if the result seems to imply that the muddiness increases. Model: The data is a random sample from $N(\mu, 0.2)$. Test the hypothesis $H_0 : \mu = 4.0$ against $H_1 : \mu > 4.0$ with a test on the 0.05 level. □

13.9) (Continued from 13.8) If μ is the correct value, what is the distribution for the r.v. that the test statistic is an observation of? Find the power function for the test, i.e. find $P(H_0 \text{ is rejected})$ if μ is the real value. What is the power of the test for $\mu = 3.8$? For $\mu = 4.3$? □

13.10) A researcher has developed a new alloy and calculated its theoretical melting point to $1050°C$. To check the result she measured the melting point of 10 samples of the alloy and obtained the following measurements:

1054.8	1052.9	1051.0	1049.8	1051.6
1047.9	1051.8	1048.5	1050.2	1050.7

The variations in measurements are effects of imperfections in the thermometer. Experience from earlier tests makes it reasonable to suppose that the measuring errors are independent and normal distributed with mean zero and standard deviation 2.3.
a) Test the hypothesis that the melting point is $\mu = 1050°C$ on the level 5%. As alternative hypothesis use that the melting point is different from $1050°C$.
b) Find the power function of the test and calculate the power for the alternatives $\mu = 1051$ and $\mu = 1053$. □

13.11) Using the arithmetic mean $\bar{x}$ out of n independent observations from $N(\mu, 2)$ the hypothesis $H_0 : \mu = 1$ is tested against $H_1 : \mu < 1$ by the test

$$\text{reject } H_0 \text{ if } \bar{x} < 1 - 2\lambda_{0.05}/\sqrt{n}.$$

How large must n be so that the power of the test for $\mu = 0$ is at least 0.99? □

13.12) (Continued from 13.10) Suppose in problem 13.10 that the standard deviation is not known in advance. Test the hypothesis that the melting point is $1050°C$ on the 5% level. As alternative hypothesis use that the melting point is different from $1050°C$. □

13.14) At a physics class, 18 independent measurements were made of the acceleration due to gravity, g (cm/s^2), and calculated the mean and standard deviation for the results: $\bar{x} = 972$, $s = 6.0$. Suppose that a normal distribution with mean g and unknown standard deviation describes the results of the measurements. Test on significance level 0.05 the hypothesis $H_0 : g = 981$ against $H_1 : g \neq 981$. □

13.18) At the manufacturing of magnecyl the weight of each pill varies as a r.v. with mean μ and standard deviation $\sigma = 0.02$. As a control 35 pills are weighed and the average weight of the 35 pills $\bar{x} = 0.69$ (unit: grams) is used as a point estimate for μ. Test the hypothesis $H_0 : \mu = 0.65$ against $H_1 : \mu \neq 0.65$ with a test with approximate significance level 0.05. (Note: there is no assumption that the weights for the pills follows a normal distribution). □

13.21) We have two random samples each with 10 observations from $N(\mu_1, 0.3)$ and $N(\mu_2, 0.4)$ respectively, where μ_1 and μ_2 are unknown parameters. We wish to test the hypothesis $\mu_1 = \mu_2$ with a suitable two-sided test on the significance level 0.01.
a) Find the power of the test if $\mu_1 - \mu_2 = 0.6$.
b) We wish to increase the number of each sample with the same amount of additional samples such that the power function takes the value 0.99 for $\mu_1 - \mu_2 = 0.6$. Approximately how many additional samples are needed? □

ANSWERS

12 ∼ a) $H_0 : \sigma_1^2 = \sigma_2^2$, $u = s_1^2/s_2^2 = 0.0478$,
critical region $R = \big(0, F_{1-\alpha/2}(7,7)\big) \cup \big(F_{\alpha/2}(7,7), \infty\big) = (0, 0.1125) \cup (8.89, \infty)$,
$u \in R$ so we reject H_0;
b) $I_{\mu_i - \mu_j} = (\bar{y}_i - \bar{y}_j) \mp t_{0.025}(df) \cdot s \cdot \sqrt{\frac{1}{n_i} + \frac{1}{n_j}}$. Here $s = 0.25527$, $df = 18$. It follows
$I_{\mu_1-\mu_2} = (-1.512, -0.9754)$, $I_{\mu_1-\mu_3} = (-1.129, -0.5170)$, $I_{\mu_2-\mu_3} = (0.1151, 0.7265)$.
We can conclude that μ_2 is the largest and thus the Type 2 line is the best.

13 ∼ a) $u = \frac{\bar{x}-\mu_0}{s/\sqrt{n}} = \frac{11.2-10}{2.1/\sqrt{25}} = 2.857$,
critical region $R = \big(-\infty, -t_{\alpha/2}(n-1),\big) \cup \big(t_{\alpha/2}, \infty\big) = \big(-\infty, -2.80,\big) \cup \big(2.80, \infty\big)$,
$u \in R$ so we reject H_0;
b) $I_\mu = \bar{x} \mp t_{\alpha/2}(n-1) \cdot \frac{s}{\sqrt{n}} = 11.2 \mp 2.80 \cdot \frac{2.1}{\sqrt{25}} = (10.024, 12.376)$.

13.8 A suitable test statistic is the arithmetic mean $\bar{x}$. Critical region is all $\bar{x}$ such that $u = \frac{\bar{x}-4.0}{0.2/\sqrt{10}} > \lambda_{0.05} = 1.64$. In this case $\bar{x} = 4.10$ and $u = 1.58$; thus the result is not significant on the significance level 0.05.

13.9 $h(\mu) = 1 - \Phi\Big(1.64 - (\mu - 1.49)/(0.2/\sqrt{10})\Big)$; $h(3.8) \approx 0$; $h(4.3) \approx 0.9990$.

13.10 a) $\frac{\bar{x}-1050}{2.3/\sqrt{10}} = 1.2649 < 1.96 = \lambda_{0.025}$; H_0 can not be rejected.
b) $h(\mu) = 1 - \Phi\Big(1.96 + (1050 - \mu)\frac{\sqrt{10}}{2.3}\Big) + \Phi\Big(-1.96 + (1050 - \mu)\frac{\sqrt{10}}{2.3}\Big)$,
$h(1051) = 0.28$, $h(1053) = 0.98$.

13.11 ≥ 64.

13.12 $\Big|\frac{\bar{x}-1050}{2.02802/\sqrt{10}}\Big| = 1.4345 < 2.26 = t_{0.025}(9)$; Do not reject H_0.

13.14 $u = -6.36$; we get $|u| > t_{0.025}(17) = 2.11$; reject H_0.

13.18 $u = 11.8$; we get $|11.8| > \lambda_{0.025} = 1.96$; reject H_0.

13.21 a) 0.89; b) 7 in each sample.

8. Multidimensional normal distribution

G1.1) Let X_1 and X_2 be independent $N(0,1)$ and define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} = \begin{pmatrix} X_1 + X_2 \\ 2X_1 - X_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \mathbf{AX}.$$

Find the density function for $\mathbf{Y}$.
Hint: For a multidimensional normal distribution $\mathbf{Z}$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{C}$ with $\det \mathbf{C} \neq 0$, $\mathbf{Z}$ has the density function

$$f(z_1, \ldots, z_n) = \frac{1}{(\sqrt{2\pi})^n \sqrt{\det \mathbf{C}}} e^{-\frac{1}{2}\left[(\mathbf{z}-\boldsymbol{\mu})^T \mathbf{C}^{-1} (\mathbf{z}-\boldsymbol{\mu})\right]}$$

□

G1.2) a) At an employment interview the applicants take three different tests with the results X_1, X_2, and X_3. In this branch of industry it is reasonable to suppose

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim N\left(\begin{pmatrix} 60 \\ 60 \\ 60 \end{pmatrix}, \begin{pmatrix} 100 & 80 & 20 \\ 80 & 100 & 10 \\ 20 & 10 & 80 \end{pmatrix} \right)$$

for a randomly chosen applicant. To make the decision process easier the results are summarized in one value $Y = (X_1 + X_2 + 2X_3)/4$. Find the distribution for Y and a number a such that $P(Y > a) = 0.90$.
b) Let the r.v. X_1, X_2, and X_3 be independent and $N(0,1)$. Let

$$U = X_1 - 2X_2 + X_3$$
$$V = c_1 X_1 + c_2 X_2 + c_3 X_3.$$

Find a necessary and sufficient condition on c_1, c_2, and c_3 so that U and V are independent. □

G1.3) At a certain moment in a communications system the received signal Y can be written on the form $Y = X + Z$, where X is the transmitted signal and Z is noise independent of X. Furthermore $X \sim N(10, 2)$ and $Z \sim N(0, 1)$.
a) Find the distribution for the random vector with the components X and Y.
b) We wish to reconstruct X by using a linear function $aY + b$ of the received signal. Find constants a and b such that $E[aY + b] = E[X]$ and such that $V[X - aY - b]$ is minimized. □

G1.4) For a random sequence $X_1, X_2, X_3, \ldots$ it holds

$$\begin{pmatrix} X_{n-2} \\ X_{n-1} \\ X_n \end{pmatrix} \sim \left(\begin{pmatrix} 20 \\ 20 \\ 20 \end{pmatrix}, \begin{pmatrix} 4 & -3.2 & 2.56 \\ -3.2 & 4 & -3.2 \\ 2.56 & -3.2 & 4 \end{pmatrix} \right)$$

This covariance structure implies that if we have two adjacent components, usually one is 'large' and the other is 'small'.
a) How can this be seen in the parameters?
b) To get less random fluctuations, so called moving averages are introduced

$$Y_{n-1} = \frac{1}{2}(X_{n-2} + X_{n-1})$$
$$Y_n = \frac{1}{2}(X_{n-1} + X_n)$$

Find the distribution of the random vector with components Y_{n-1} and Y_n.
c) Find the correlation $\rho[Y_{n-1}, Y_n]$. □

G1.6) The random variables $X_1, X_2, \ldots, X_5$ are independent and $N(10, 4)$. Consider

$$Y_1 = \frac{1}{5}(X_1 + X_2 + X_3 + X_4 + X_5),$$
$$Y_2 = 4X_1 + X_2 - X_3 - X_4 - X_5.$$

a) Find the joint distribution for (Y_1, Y_2).
b) Find $P(Y_1 > Y_2)$.
c) Calculate the correlation coefficient between Y_1 and Y_2. □

G1.7) Let X_1 and X_2 be the results for two psychological tests for a person and X_3 be the grading for how well the person performs a certain task at the a company. From experience it is known that $(X_1 X_2 X_3)^T$ has a three dimensional normal distribution with mean vector and covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} 71 \\ 53 \\ 18 \end{pmatrix}; \quad \mathbf{C} = \begin{pmatrix} 100 & 64 & 38 \\ 64 & 64 & 28.8 \\ 38 & 28.8 & 16 \end{pmatrix}.$$

We wish to obtain information about X_3 by means of X_1 and X_2. It is possible to determine a best linear predictor

$$\hat{X}_3 = a + bX_1 + cX_2$$

with the constraints

$$(i) \quad E[\hat{X}_3] = E[X_3]$$
$$(ii) \quad V[X_3 - \hat{X}_3] \quad \text{is minimized.}$$

a) Find this best linear predictor $\hat{X}_3$.
b) Show that the prediction error $\varepsilon = X_3 - \hat{X}_3$ is independent of X_1. (The prediction error is independent of X_2 too but you don't have to show that). □

G1.8) The random variable $\begin{pmatrix} X_1 & X_2 & X_3 \end{pmatrix}^T$ has a three dimensional normal distribution with mean vector and covariance matrix

$$\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{respectively} \quad \begin{pmatrix} 7/2 & 1/2 & -1 \\ 1/2 & 1/2 & 0 \\ -1 & 0 & 1/2 \end{pmatrix}$$

a) Find the mean vector and covariance matrix for $\begin{pmatrix} Y_1 & Y_2 & Y_3 \end{pmatrix}^T$, where

$$Y_1 = X_2 + X_3 \qquad Y_2 = X_1 + X_3 \qquad Y_3 = X_1 + X_2$$

b) Calculate $P(Y_2 > 2Y_3)$.
c) Is Y_1 and Y_2 independent? The answer must be explained. □

G1.9) The disturbances $\varepsilon_1, \varepsilon_2, \varepsilon_3$ at three consecutive signal transmissions in a communications system can be considered as the components of a normal distributed vector with mean vector and covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}; \quad \mathbf{C} = \begin{pmatrix} 1.5 & 0.9 & 0.54 \\ 0.9 & 1.5 & 0.9 \\ 0.54 & 0.9 & 1.5 \end{pmatrix}$$

Find the probability that the average

$$\varepsilon = (\varepsilon_1 + \varepsilon_2 + \varepsilon_3)/3$$

of the three disturbances has an absolute value greater than 2 units. □

ANSWERS

G1.1 $f(y_1, y_2) = \frac{1}{6\pi} e^{-\frac{1}{18}(5y_1^2 - 2y_1y_2 + 2y_2^2)}$.

G1.2 a) $Y \sim N(60, \sqrt{50})$, $a \approx 50.949$; b) $c_1 - 2c_2 + c_3 = 0$.

G1.3 a) $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 10 \\ 10 \end{pmatrix}, \begin{pmatrix} 4 & 4 \\ 4 & 5 \end{pmatrix}\right)$ b) $a = 4/5 \quad b = 2$.

G1.4 a) $\rho(X_{n-1}, X_n) = -0.8$. A large negative correlation means that one variable is 'large' when the other is 'small' and vice versa.

b) $\begin{pmatrix} Y_{n-1} \\ Y_n \end{pmatrix} \sim N\left(\begin{pmatrix} 20 \\ 20 \end{pmatrix}, \begin{pmatrix} 0.4 & 0.04 \\ 0.04 & 0.4 \end{pmatrix}\right)$.

c) $\rho = 0.10$.

G1.6 a) $\mathbf{Y} \sim N\left(\begin{pmatrix} 10 \\ 20 \end{pmatrix}, \begin{pmatrix} 0.8 & 1.6 \\ 1.6 & 80 \end{pmatrix}\right)$ b) $1 - \Phi(1.135) \approx 0.13$ c) $\rho = 0.2$.

G1.7 $\hat{X}_3 = -10.45 + \frac{23}{90} X_1 + \frac{7}{36} X_2$.

G1.8 a) $\mathbf{Y} = \mathbf{A}\mathbf{X} = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix} \mathbf{X}$. Thus $E[\mathbf{Y}] = \mathbf{A} \cdot E[\mathbf{X}] = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

and $C_{\mathbf{Y}} = \mathbf{A} C_{\mathbf{X}} \mathbf{A}^T = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 3 \\ 0 & 3 & 5 \end{pmatrix}$.

b) $P(Y_2 > 2Y_3) = P(Y_2 - 2Y_3 > 0)$, $Y_1 - 2Y_3 = \begin{pmatrix} 0 & 1 & -2 \end{pmatrix} \mathbf{Y}$,

$Y_2 - 2Y_3 \sim N\left(\begin{pmatrix} 0 & 1 & -2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & -2 \end{pmatrix} C_{\mathbf{Y}} \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}\right) = N(1, 10)$,

$P(Y_2 - 2Y_3 > 0) = \Phi(1/\sqrt{10}) \approx 0.63$.

c) Y_1 and Y_2 are independent since they have a simultaneous normal distribution with $C[Y_1, Y_2] = 0$.

G1.9 $\bar{\varepsilon} \sim N(0, 1.02)$, $P(|\bar{\varepsilon}| > 2) = 0.0478$.

9. Simple linear regression

G2.1) Show that the vector of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ has the covariance matrix $(\mathbf{I} - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)\sigma^2$. □

G2.2) Consider the simple linear regression model

$$Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, 2, \ldots, n$$

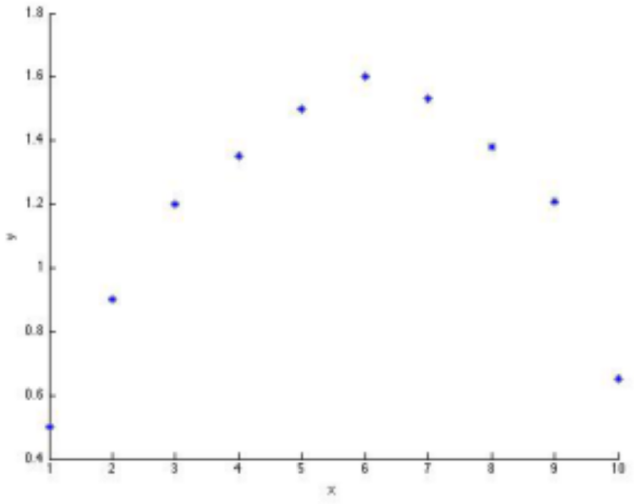
where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are independent $N(0, \sigma^2)$.
Show that the least square estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are independent if and only if

$$\sum_{j=1}^{n} x_j = 0.$$

□

G2.3) A new medicine against cancer was tested on ten mice, each with a tumor of the size 4 grams. The mice were given different doses (x) of the medicine and the reduction (y) of the tumor was measured on each mouse. Result:

x	1	2	3	4	5	6	7	8	9	10
y	0.50	0.90	1.20	1.35	1.50	1.60	1.53	1.38	1.21	0.65



Model: $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$ where $\varepsilon = N(0, \sigma^2)$.

An analysis of variance according to this model gave:

Analysis of variance
Estimated regression line: $y = -0.0283 + 0.551x - 0.0473x^2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-0.02833	0.07935
1	0.55150	0.03314
2	-0.047348	0.002936

	Degrees of freedom	Sum of squares
REGR	2	1.26130
RES	7	0.03186
TOT	9	1.29316

a) Test $H_0 : \beta_1 = \beta_2 = 0$ against $H_1 :$ at least one of β_1 and $\beta_2 \neq 0$, on the 0.001 level.
b) Find the optimal dose according to this regression analysis.
c) Is it reasonable to remove one of the explanatory variables? Briefly justify your answer. □

G2.4) In a study for the profitability of movie studies, 20 Hollywood movies were chosen randomly and for each movie the following values were obtained:

$$y = \text{gross revenue (unit: millions of USD)}$$
$$x_1 = \text{production cost (unit: millions of USD)}$$
$$x_2 = \text{marketing cost (unit: millions of USD)}.$$

There was special interest in considering whether there was any influence if the movie was based on a book that had been published before the movie was produced. To separate such movies from the others a so called dummy-variable was defined

$$x_3 = \begin{cases} 1 & \text{for movies based on a book} \\ 0 & \text{otherwise} \end{cases}$$

Result:

	x_1	x_2	x_3	y
1	4.2	1.0	0	28
2	6.0	3.0	1	35
3	5.5	6.0	1	50
4	3.3	1.0	0	20
5	12.5	11.0	1	75
6	9.6	8.0	1	60
7	2.5	0.5	0	15
8	10.8	5.0	0	45
9	8.4	3.0	1	50
10	6.6	2.0	0	34
11	10.7	1.0	1	48
12	11.0	15.0	1	82
13	3.5	4.0	0	24
14	6.9	10.0	0	50
15	7.8	9.0	1	58
16	10.1	10.0	0	63
17	5.0	1.0	1	30
18	7.5	5.0	0	37
19	6.4	8.0	1	45
20	10.0	12.0	1	72

Data has been analyzed according to the models

$$\text{Model 1: } Y = \beta_0' + \beta_2' x_2 + \tilde{\varepsilon}$$
$$\text{Model 2: } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where $\tilde{\varepsilon}$ respectively ε in the two models are independent and $N(0, \tilde{\sigma}^2)$ respectively $N(0, \sigma^2)$. Analysis of variance can be found below.

Analysis of variance no. 1
Estimated regression line: $y = 24.3 + 3.76x_2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	24.332	3.387
1	3.7606	0.4726

	Degrees of freedom	Sum of squares
REGR	1	5094.6
RES	18	1448.3
TOT	19	6542.9

Analysis of variance no. 2
Estimated regression line: $y = 7.84 + 2.85x_1 + 2.28x_2 + 7.17x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	7.836	2.333
1	2.8477	0.3923
2	2.2782	0.2534
3	7.166	1.818

	Degrees of freedom	Sum of squares
REGR	3	6325.2
RES	16	217.8
TOT	19	6542.9

a) Briefly explain why model 2 explains the data better than model 1. Justify your answer with suitable parameters from the data analysis.
b) Does seem to affect the revenue if the movie is based on a book and in that case in which way is it affected? Justify your answer with a suitable 95% confidence interval.
c) A movie based on a just published book is about to be produced. The production cost is estimated to be 11 million USD and 9 million will be spent on marketing the movie. Estimate the expected gross revenue from the movie using model 2. You don't have to construct an interval.

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.399978 & -0.051254 & 0.006225 & -0.010691 \\ -0.051254 & 0.011308 & -0.004290 & -0.014215 \\ 0.006225 & -0.004290 & 0.004719 & -0.003020 \\ -0.010691 & -0.014215 & -0.003020 & 0.242792 \end{pmatrix}.$$

□

G2.5) Suppose that the time Y for a chemical reaction has a linear regression with respect to the temperature x. We have the following data:

x_i	15	18	21	24	27
y_i	13.8	11.5	9.2	7.6	5.4

a) Calculate point estimates for β_0, β_1 and σ^2.
b) Plot the data points as well as the regression line in a coordinate system. □

14.2) For the numerical data $(x_1, y_1), (x_2, y_2), \ldots, (x_{10}, y_{10})$ the following has been calculated

$$\sum x_i = 12.0, \quad \sum x_i^2 = 18.40, \quad \sum y_i = 15.0, \quad \sum y_i^2 = 27.86, \quad \sum x_i y_i = 20.40.$$

The data is described by a regression model $y_i = \alpha + \beta x_i + z_i$ where $z_1, \ldots, z_{10}$ are independent observations from $N(0, \sigma)$. Find 95% confidence intervals for β and α. □

14.4) To see if a certain dimension y at a manufactured item depends on the setting x on a certain machine, y was measured for 7 different settings of the machine and the following data was obtained:

x :	1.0	2.0	3.0	4.0	5.0	6.0	7.0
y :	0.9	1.4	2.2	2.7	3.2	4.3	4.2

a) Plot the data in a coordinate system and determine if it is reasonably to assume that y depends linearly on x (with some random variations).
b) Estimate the parameters of the regression model and plot the regression line in the coordinate system.
c) If you wish to have 2.5 as the expected value for the dimension, what should the setting of x be?
d) Find a 95% confidence interval for the intercept α and the slope β.
e) Find a 95% confidence interval for $\mu_0 = \alpha + \beta x_0$ and plot the boundaries as well as the estimated regression line in a coordinate system. □

14.7) A testing facility had the task to investigate how the nicotine content (y) depends on the contents of carbon (x_1) and chloride (x_2). A multiple regression model approach was made

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where ε is $N(0, \sigma)$. After computer processing the data the following table was obtained:

	Coefficient	Standard error	p-value	Lower 95%	Upper 95%
Intercept	-0.471	0.626	0.459	-1.769	0.826
Carbon	1.423	0.247	8.3E-06	0.912	1.935
Chloride	-0.176	0.112	0.132	-0.409	0.057

In addition the covariance between the β_1- and the β_2- estimates was estimated to -0.00248.
a) Should the hypothesis $H_{\beta_1} : \beta_1 = 0$ respectively $H_{\beta_2} : \beta_2 = 0$ be rejected on the 5% significance level?
b) Estimate how much the expected nicotine contents is changed if both the carbon- and chloride-content is increased by 1 unit.
c) Calculate the standard error for the estimate in b). □

ANSWERS

G2.1 Write $\mathbf{e}$ as a linear transformation of the $\mathbf{Y}$ vector.

G2.3 a) $v = 138.55 > 21.69$ (from $F(2,7)$-table). H_0 is rejected.
b) $x = 5.82$ gives a maximum according to the estimated regression equation.
c) Taking into account the appearance of the curve, none of the explanatory variables should be excluded.

G2.4 a) Model 2 has the coefficient of determination $R^2 = 96.7\%$ which is significantly better than $R^2 = 77.9\%$ for model 1.
b) $I_{\beta_3} = (7.166 \mp 2.12 \cdot 1.818) = (3.312, 11.020)$. It seems that a manuscript based on a book gives higher gross revenue.
c) $\hat{\beta}_0 + 11\hat{\beta}_1 + 9\hat{\beta}_2 + \hat{\beta}_3 = 66.83$. The expected gross revenue is approximately 67 million dollars.

G2.5 a) $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where

$$\mathbf{X}^T = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 15 & 18 & \dots & 27 \end{pmatrix}, \quad \mathbf{y}^T = (13.8 11.5 \dots 5.4),$$

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \begin{pmatrix} 5 & 105 \\ 105 & 2295 \end{pmatrix}^{-1} \begin{pmatrix} 47.5 \\ 935.4 \end{pmatrix}$$

$$= \frac{1}{450} \begin{pmatrix} 2295 & -105 \\ -105 & 5 \end{pmatrix} \begin{pmatrix} 47.5 \\ 935.4 \end{pmatrix} = \begin{pmatrix} 23.99 \\ -0.6900 \end{pmatrix}$$

$\hat{\beta}_0 + \hat{\beta}_1 x_i : 13.64 \;\; 11.57 \;\; 9.50 \;\; 7.43 \;\; 5.36.$
σ^2 is estimated by $s^2 = \frac{1}{3}\sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = 0.050$.

14.2 $I_\beta = 0.60 \mp 0.81$ respectively $I_\alpha = 0.78 \mp 1.09$.

14.4 a) It seems reasonable to assume that y depends linearly on x;
b) $\alpha^* = 0.3143, \quad \beta^* = 0.5964$;
c) $x = 3.66$;
d) $I_\alpha = \alpha^* \mp t_{p/2}(n-2) \cdot s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$ gives $I_\alpha = 0.3143 \mp 0.5138$,
$I_\beta = \beta^* \mp t_{p/2}(n-2) \cdot s \Big/ \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$ gives $I_\beta = 0.5964 \mp 0.1149$;
e) $I_{\mu_0} = \mu_0^* \mp t_{p/2}(n-2) \cdot s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}}$.

14.7 a) The first hypothesis is rejected but the other is not since the p-value is less than 5% in the first case but not in the other;
b) The expected increase is $\beta_1^* + \beta_2^* = 1.247$;
c) $V[\beta_1^* + \beta_2^*] = V[\beta_1^*] + V[\beta_2^*] + 2C[\beta_1^*, \beta_2^*]$ which is estimated with
$0.247^2 + 0.112^2 - 2 \cdot 0.00248 = 0.0686$. The standard error is the square root of this value $d(\beta_1^* + \beta_2^*) = 0.26$.

10. Confidence- and prediction- intervals

G2.6) At a test of the brakes on a car, the car was repeatedly stopped from a speed of around 100 km/h on dry tarmac. At each test the speed was measured at the start of the slowdown (in practice it was hard to keep the speed 100 km/h) and the stopping distance. Results:

Starting speed	103.5	98.0	95.5	102.0	100.0
Stopping distance	57.0	51.5	50.0	56.0	54.0

a) Let Y_j, $j = 1, 2, \ldots, 5$ denote the stopping distance in the respective test. Suppose that $Y_1, Y_2, \ldots, Y_5$ are independent $N(\mu, \sigma^2)$, where μ denotes the expected stopping distance at starting speed 100 km/h. (We consider all variation in the stopping distance, even the one caused by different starting speed, as purely random). Thus we have $Y_j = \mu + \varepsilon_j$ where $\varepsilon_j \sim N(0, \sigma^2)$. Find a 95% confidence interval for μ.
b) A part of the variation in stopping distance is probably due to the starting distance not being exactly 100 km/h. By using a linear model you should be able to take that speed into account. In test j we had the starting speed x_j and the stopping distance y_j as an observation of $Y_j = \beta_0 + \beta_1 x_j + \tilde{\varepsilon}_j$ where $\tilde{\varepsilon}_j \sim N(0, \tilde{\sigma}^2)$. An analysis of variance found:

Analysis of variance
Estimated regression line: $y = -38.4 + 0.923x$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-38.423	6.169
1	0.92308	0.06179

	Degrees of freedom	Sum of squares
REGR	1	34.338
RES	3	0.462
TOT	4	34.800

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 247.347 & -2.47643 \\ -2.47643 & 0.02481 \end{pmatrix}.$$

Find a 95% confidence interval for the expected stopping distance at the speed 100, i.e. for $\beta_0 + 100\beta_1$.
c) Compare the results in a) and b). Which method do you prefer? □

G2.7) A factory produces nitric acid by oxidizing ammonia. During a period of 21 days the corresponding values of

x_1 = air flow
x_2 = inlet temperature of cooling water
x_3 = concentration of HNO_3 in the absorbing liquid
$y = 10\times$ halt (in %) of NH_3 that is lost, i.e. a reverse measure of the exchange

were measured. Data from Operation of a Plant for the Oxidation of Ammonia to Nitric Acid:

Run no.	Air flow x_1	Cooling water inlet temp. x_2	Acid concentration x_3	Stack loss y
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Analysis of variance according to the following models are found below.

$$\underline{\text{Model 1}}\text{: } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$
$$\underline{\text{Model 2}}\text{: } Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Using model 2, construct a 95% confidence interval for $E[Y]$ when $x_1 = 50$ and $x_2 = 18$. Are you happy with the interval?

Analysis of variance no. 1
Estimated regression line: $y = -39.9 + 0.716x_1 + 1.30x_2 - 0.153x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-39.92	11.90
1	0.7156	0.1349
2	1.2953	0.3680
3	-0.1521	0.1563

	Degrees of freedom	Sum of squares
REGR	3	1890.41
RES	17	178.83
TOT	20	2069.24

Analysis of variance no. 2
Estimated regression line: $y = -50.4 + 0.671x_1 + 1.30x_2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-50.359	5.138
1	0.6712	0.1267
2	1.2954	0.3675

	Degrees of freedom	Sum of squares
REGR	2	1880.44
RES	18	188.80
TOT	20	2069.24

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 2.51724 & -0.01926 & -0.06189 \\ -0.01926 & 0.00153 & -0.00347 \\ -0.06189 & -0.00347 & 0.01288 \end{pmatrix}.$$

□

G2.8) We wish to use a regression model for finding prices on passenger aircraft. As dependent variable we have

$$Y = \text{the aircraft price/number of passenger seats (unit: 1000s of USD)}$$

and as explanatory variables

$$x_1 = \text{starting weight/number of passenger seats}$$
$$x_2 = \ln \text{ (speed)}.$$

Observed values:

x_1	x_2	y
249.3	5.44	172.00
272.3	5.59	194.44
219.6	5.65	190.00
213.7	5.50	129.55
216.8	5.59	148.91
290.6	5.66	135.16
226.8	5.56	116.07
233.9	5.66	166.67
220.6	6.12	150.00
222.4	6.12	177.57
225.7	5.61	178.57
236.0	5.50	115.39
199.9	5.59	154.41
252.6	5.95	198.86
224.1	5.95	181.37
212.9	5.36	127.78
211.1	6.14	169.23

Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Analysis of variance

Estimated regression line: $y = -283 + 0.688x_1 + 50.3x_2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-282.7	142.4
1	0.6881	0.2765
2	50.29	21.73

	Degrees of freedom	Sum of squares
REGR	2	5119.3
RES	14	6605.7
TOT	16	11725.0

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 42.9585 & -0.04120 & -5.8932 \\ -0.04120 & 0.0001621 & 0.000825 \\ -5.8932 & 0.000825 & 1.0004 \end{pmatrix}.$$

Construct a 95% prediction interval for the price of an airplane with 60 seats, starting weight 15000 and the speed 287. Are you happy with the interval? □

G2.9) In the middle of the 19th century the Scottish physicist James D. Forbes wanted to estimate the height above sea level by measuring the boiling point of water. He knew that the height above sea level could be measured with the help of air pressure. In a series of experiments he studied the relation between air pressure and boiling point. The motivation for this solution to the problem was that the

19th century barometers were fragile and thus hard for travelers to transport. Forbes collected data in the Alps and Scotland, and published the following data in 1857:

Case no.	Boiling point ($^\circ F$)	Pressure (in. Hg)	Log(Pressure)	100×Log(Pressure)
1	194.5	20.79	1.3179	131.79
2	194.3	20.79	1.3179	131.79
3	197.9	22.40	1.3502	135.02
4	198.4	22.67	1.3555	135.55
5	199.4	23.15	1.3646	136.46
6	199.9	23.35	1.3683	136.83
7	200.9	23.89	1.3782	137.82
8	201.1	23.99	1.3800	138.00
9	201.4	24.02	1.3806	138.06
10	201.3	24.01	1.3805	138.05
11	203.6	25.14	1.4004	140.04
12	204.6	26.57	1.4244	142.44
13	209.5	28.49	1.4547	145.47
14	208.6	27.76	1.4434	144.34
15	210.7	29.04	1.4630	146.30
16	211.9	29.88	1.4754	147.54
17	212.2	30.06	1.4780	147.80

An analysis according to
Model 1: $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ where $\varepsilon \sim N(0, \sigma^2)$, x_j = temperature and $y_j = 100 \cdot \log(\text{pressure})$, is found below. The plot of residuals show one deviating observation, a so called outlier, namely observation no. 12. We wish to investigate if it is likely that this deviating sample has occurred out of pure chance.

a) The data has also been analyzed according to
Model 2: $Y_j = \beta_0 + \beta_1 x_j + \beta_2 u_j + \varepsilon'_j$ where

$$u_j = \begin{cases} 1 & \text{for } j = 12 \\ 0 & \text{otherwise.} \end{cases}$$

Use a suitable test or confidence interval to determine if observation 12 can be considered to be deviant on the 0.05 level.

b) In analysis no. 3, observation 12 was removed and the rest of the data was analyzed according to model 1. Use this analysis to construct a 95% interval that can be used to find out if observation 12 can be considered deviant. What is your conclusion?

Remark: It is possible to show that the methods in a) and b) are equivalent.

Analysis of variance no. 1
Estimated regression line: $y = -42.2 + 0.896x$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-42.164	3.341
1	0.89562	0.01646

	Degrees of freedom	Sum of squares
REGR	1	425.76
RES	15	2.16
TOT	16	427.91

Analysis of variance no. 2
Estimated regression line: $y = -41.3 + 0.891x + 1.45u$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-41.335	1.003
1	0.891110	0.004944
2	1.4528	0.1174

	Degrees of freedom	Sum of squares
REGR	2	427.73
RES	14	0.18
TOT	16	427.91

Analysis of variance no. 3
Estimated regression line: $y = -41.3 + 0.891x$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-41.335	1.003
1	0.891110	0.004944

	Degrees of freedom	Sum of squares
REGR	1	419.19
RES	14	0.18
TOT	15	419.37

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 78.0093 & -0.3843 \\ -0.3843 & 0.001894 \end{pmatrix}.$$

□

G2.10) A psychologist was interested in studying the effects of sleep deprivation. 20 people with similar sleeping habits were randomly put into five groups with four individuals in each. A test was performed during two days. The night between day 1 and day 2 the sleep assignments were:

Group 1: 0 hours of sleep
Group 2: 2 hours of sleep
Group 3: 4 hours of sleep
Group 4: 6 hours of sleep
Group 5: 8 hours of sleep

Day 1: In the morning each person had to do a test consisting of addition of numbers for ten minutes.
Day 2: In the morning each person had to do a similar test in which they performed addition.
For each person the difference, between the number of correct additions day 1 and the number of correct additions on day 2, was calculated. Result:

Group no.	Observed data			
1	39	33	41	40
2	25	29	34	26
3	10	18	14	17
4	4	6	-1	9
5	-5	0	-3	-8

Model: For person no. i the difference in test result y_i is an observation of $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ where $\varepsilon_1, \ldots, \varepsilon_{20}$ is independent $N(0, \sigma^2)$ and x_i = number of hours of sleep. An analysis of variance is found below.
a) Construct a 95% confidence interval for β_1 and test on the 0.05 level

$$H_0 : \beta_1 = -4 \quad \text{against} \quad H_1 : \beta_1 \neq -4.$$

b) Find out using a suitable test or confidence interval if the regression line intersects the x-axis at $x = 8$, i.e. if the people sleeping eight hours has unchanged ability to pass the test. Level 5%.

Analysis of variance
Estimated regression line: $y = 38.1 - 5.43x$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	38.100	1.420
1	-5.4250	0.2898

	Degrees of freedom	Sum of squares
REGR	1	4708.9
RES	18	241.9
TOT	19	4950.8

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.150000 & -0.025000 \\ -0.025000 & 0.006250 \end{pmatrix}.$$

□

G2.11) The following table shows the expenses for private consumption (y) and the disposable income (x_1) both expressed in billions of USD. The variable x_2 denotes the war state

$$x_1 = \begin{cases} 1 & \text{when the country is at war} \\ 0 & \text{otherwise} \end{cases}$$

This data is for USA during the years $1935 - 1949$.

x_1	x_2	y
58.5	0	56
66.3	0	62
71.2	0	67
65.5	0	64
70.3	0	67
75.7	0	71
92.7	0	81
116.9	1	89
133.5	1	99
146.3	1	108
150.2	1	120
160.0	0	144
169.8	0	162
189.2	0	175

An analysis of the data according to the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$:

Analysis of variance
Estimated regression line: $y = 1.00 + 0.92x_1 - 23.34x_2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	1.00157	2.38554
1	0.924056	0.0196041
2	-23.3432	2.06080

	Degrees of freedom	Sum of squares
REGR	2	25868.1
RES	12	139.677
TOT	14	26007.777

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.488909 & -0.003625 & 0.006729 \\ -0.003625 & 0.000033 & -0.000889 \\ 0.006729 & -0.000889 & 0.364862 \end{pmatrix}.$$

a) In which way is the private consumption affected by the war state according to this analysis? Explain your answer by constructing a suitable two-sided 99% confidence interval.

b) Construct a 99% prediction interval for the private consumption a year when $x_1 = 150$ and the country is not at war. □

G2.12) The table below is data for 20 businessmen's investments and corresponding profits. If we wish to invest 20 capital, what profit can we expect? Find a point estimate and 95% interval estimate.

Investment (x)	14	8	7	26	8	2	3	22	6	23
Profit (y)	83	65	71	140	135	30	30	128	80	68
Investment (x)	29	4	13	14	7	5	13	6	5	8
Profit (y)	139	88	121	125	56	98	101	96	73	116

Analysis of variance
Estimated regression line: $y = 57.5 + 3.55x$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	57.541	9.637
1	3.5524	0.0784

	Degrees of freedom	Sum of squares
REGR	1	15327
RES	18	10972
TOT	19	26299

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.152361 & -0.009180 \\ -0.009180 & 0.000823 \end{pmatrix}.$$

□

G2.13) A university has a computer that is used by teachers, students as well as external users. Let

$x_1 =$ number of university users
$x_2 =$ number of external users
$x_1 =$ average response time in hundredths of seconds

Corresponding values for x_1, x_2 and y has been observed:

x_1	x_2	y	x_1	x_2	y	x_1	x_2	y	x_1	x_2	y
10	0	8	48	4	52	23	3	13	69	12	106
36	8	59	21	1	13	66	7	81	58	10	74
75	5	77	66	10	88	10	2	15	26	2	23
16	4	21	30	3	28	70	10	100	14	4	18
35	5	45	55	9	70	44	0	28	62	9	72
50	8	56	49	6	63	42	4	43	63	3	55

An analysis according to the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

has been made. Analysis of variance can be found below.
a) Test on the 0.001 level $H_0 : \beta_1 = \beta_2 = 0$ against $H_1 :$ at least one of β_1 and β_2 is $\neq 0$.
b) Are internal and external users having the same effect on the load of the computer, i.e. is it possible that $\beta_1 = \beta_2$? Construct a confidence interval for $\beta_1 - \beta_2$ with confidence level 95%.

Analysis of variance
Estimated regression line: $y = -7.67 + 0.857x_1 + 3.90x_2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	-7.668	2.788
1	0.85692	0.08144
2	3.8957	0.4889

	Degrees of freedom	Sum of squares
REGR	2	19198.3
RES	21	707.0
TOT	23	19905.3

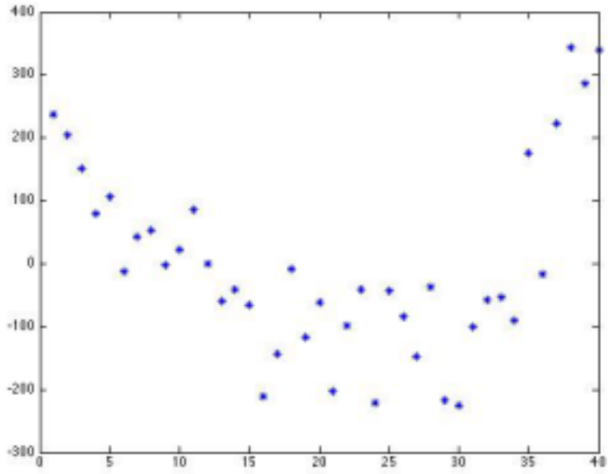
$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 2.230838 & -0.004077 & -0.002392 \\ -0.004077 & 0.000197 & -0.000827 \\ -0.002392 & -0.000827 & 0.007098 \end{pmatrix}.$$

□

G3.1) A time series y_t, $t = 1, 2, \ldots, 40$ has been observed. To find a suitable model for Y_t an analysis is made with the preliminary model

$$Y_t = \beta_0 + \beta_1 t + \tilde{\varepsilon}_t$$

where the $\tilde{\varepsilon}_t$-variables are independent and $N(0, \tilde{\sigma}^2)$. The residuals were plotted against t.



a) After studying the residual plot it is decided to analyze the data according to the model

$$Y_t = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon_t$$

where the ε_t-variables are independent and $N(0, \sigma^2 t)$. State <u>two</u> reasons why this model is chosen.
b) Provide a formula estimating $\beta_0, \beta_1, \beta_2$ in the model in a). Any matrices in the formula should be defined in detail. (Hint: Transform the data in a suitable way). □

G4.1) To compare three different types of railroad tracks, two miles of each type was laid out in five different districts and during a period of two years it was observed

x = average number of trains per day, passing this section of rail

y = number of cracks in the rail.

Result:

Type A		Type B		Type C	
x	y	x	y	x	y
16.9	8	17.8	5	19.6	9
23.6	11	24.4	9	25.4	8
14.4	7	13.5	5	35.5	16
17.2	10	20.1	6	16.8	7
9.1	4	11.0	4	31.2	11

Model: $Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 x + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$ and where

$$z_1 = \begin{cases} 1 & \text{for type B} \\ 0 & \text{otherwise} \end{cases} \qquad z_2 = \begin{cases} 1 & \text{for type C} \\ 0 & \text{otherwise.} \end{cases}$$

An analysis of variance can be found below.
a) Does the average number of trains per day seem to affect the number of cracks? Perform a suitable test on the 0.05 level for the above model.
b) Are there differences between the track types considering the forming of cracks? Perform a suitable test on the 0.05 level.

Analysis of variance
Estimated regression line: $y = 1.39 - 2.66x_1 - 1.64x_2 + 0.41x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	1.39097	1.13679
1	-2.65579	0.823806
2	-1.64984	0.999121
3	0.406960	0.0601822

	Degrees of freedom	Sum of squares
REGR	3	125.462
RES	11	18.5381
TOT	14	144.000

Analysis of variance
Estimated regression line: $y = 0.41 + 0.38x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	0.414181	1.29130
1	0.383768	0.0615918

	Degrees of freedom	Sum of squares
REGR	1	107.877
RES	13	36.1277
TOT	14	144.000

□

G4.2) A company has compared the demand of a certain type of hygiene articles in twelve different sales districts. In five of the districts the article has only been sold through the retail chain A and in the seven other districts it has been sold in many different retail stores. The following result was obtained:

District no.	Distribution	Degree of urbanization	Relative income	Sales volume per resident
1	Only A	42.2	31.9	167
2	-	48.6	33.2	185
3	-	42.6	28.7	170
4	-	39.0	26.1	152
5	-	34.7	30.1	150
6	Many retailers	44.5	28.5	192
7	-	39.1	24.3	183
8	-	40.1	28.6	180
9	-	45.9	20.4	191
10	-	36.2	24.1	171
11	-	39.3	30.0	168
12	-	46.1	34.3	189

We wish to investigate if the distribution form affects the sales volume by using a regression model. Let

$$x_1 = \begin{cases} 1 & \text{if distributed only by A} \\ 0 & \text{if distributed by a large number of retailers} \end{cases}$$

$x_2 =$ degree of urbanization

$x_3 =$ relative income

$Y =$ sales volume per resident.

A regression analysis with the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ gave the following result:

Analysis of variance

Estimated regression line: $y = 83.8 - 16.2x_1 + 2.50x_2 - 0.208x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	83.84	16.87
1	-16.163	3.257
2	2.4953	0.3872
3	-0.2078	0.4313

	Degrees of freedom	Sum of squares
REGR	3	2026.17
RES	8	211.49
TOT	11	2237.67

a) Does the choice of form of distribution seem to have an effect?
b) Below follows some analyses for different combinations of explanatory variables. Which explanatory variable would you choose if you could only choose one variable?
Can you be sure that the best pair of explanatory variables contains the variable you picked above? How many explanatory variables (one, two or three) do you think is reasonable to use in this example?

Analysis of variance 1

Estimated regression line: $y = 182 - 17.2x_1$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	182.000	4.432
1	-17.200	6.866

	Degrees of freedom	Sum of squares
REGR	1	862.9
RES	10	1374.8
TOT	11	2237.67

Analysis of variance 2
Estimated regression line: $y = 71.7 + 2.48x_2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	71.72	30.44
1	2.4833	0.7296

	Degrees of freedom	Sum of squares
REGR	1	1201.0
RES	10	1036.6
TOT	11	2237.67

Analysis of variance 3
Estimated regression line: $y = 181 - 0.23x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	181.23	31.82
1	-0.226	1.112

	Degrees of freedom	Sum of squares
REGR	1	9.2
RES	10	2228.5
TOT	11	2237.67

Analysis of variance 4
Estimated regression line: $y = 80.6 - 16.8x_1 + 2.44x_2$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	80.58	14.78
1	-16.761	2.880
2	2.4381	0.3524

	Degrees of freedom	Sum of squares
REGR	2	2020.0
RES	9	217.6
TOT	11	2237.67

Analysis of variance 5
Estimated regression line: $y = 164 - 19.0x_1 + 0.645x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	164.48	26.56
1	-19.024	7.570
2	0.6449	0.9630

	Degrees of freedom	Sum of squares
REGR	2	928.1
RES	9	1309.5
TOT	11	2237.67

Analysis of variance 6
Estimated regression line: $y = 89.4 + 2.76x_2 - 1.02x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	89.35	32.06
1	2.7572	0.7303
2	-1.0233	0.7591

	Degrees of freedom	Sum of squares
REGR	2	1375.17
RES	9	862.50
TOT	11	2237.67

□

G4.3) A certain type of bus has only one door which must be used both of the passengers disembarking and the ones going into the bus. At 20 occasions the time y (in seconds) where measured from the moment the bus stopped at a bus stop until it was moving again. At the same time the number of passengers boarding (x_1) and the number of passengers getting off (x_2) was noted.

x_1	x_2	y
0	1	4
2	3	24
1	0	23
1	0	12
2	1	20
4	0	45
5	8	60
1	1	18
0	1	5
1	0	15
1	3	18
8	3	88
5	0	50
1	3	24
1	1	12
0	3	8
0	6	14
1	4	16
2	0	32
0	8	25

We wish to describe the stop time of the bus y with a linear regression model on x_1 and/or x_2. An analysis of variance can be found below.
a) If you could only use one variable in the regression, which one would you choose? Explain your answer. Is it good as an explanatory variable? Perform a suitable test on the 1% level.
b) Using the model in a) would there be any use in also taking into account the other explanatory variable in the regression? Perform a test on the 1% level.
Estimated correlations are given by

	x_1	x_2
x_2	0.012	
y	0.960	0.192

Analysis of variance
Estimated regression line: $y = 8.74 + 9.40x_1$

	Degrees of freedom	Sum of squares
REGR	1	7523.0
RES	18	635.5
TOT	19	8158.5

Analysis of variance
Estimated regression line: $y = 5.41 + 9.38x_1 + 1.46x_2$

	Degrees of freedom	Sum of squares
REGR	2	7789.2
RES	17	369.3
TOT	19	8158.5

□

G4.4) In a study of the survival time for patients with prostate cancer, for each patient the treatment type (x_1), age in years (x_2) at the time of the start of the treatment, the halt (x_3) of a certain characteristic substance, AP, in the blood, the occurrence of of skeletal metastasis (x_4) and the survival time (y) from the start of the treatment, was observed. We have

$$x_1 = \begin{cases} 0 & \text{at placebo treatment (=no treatment)} \\ 1 & \text{at estrogen treatment.} \end{cases}$$

$$x_4 = \begin{cases} 0 & \text{for no skeletal metastasis} \\ 1 & \text{presence of skeletal metastasis.} \end{cases}$$

An analysis of variance for the data for the models

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

where $\varepsilon \sim N(0, \sigma^2)$ and

$$Y = \beta_0 + \beta_1 x_1 + \beta_3 x_3 + \tilde{\varepsilon}$$

where $\tilde{\varepsilon} \sim N(0, \tilde{\sigma}^2)$ yielded the estimated correlations

	x_1	x_2	x_3	x_4
x_2	-0.003			
x_3	-0.008	-0.323		
x_4	0.206	-0.207	0.242	
y	0.182	-0.117	-0.344	-0.087

and

Analysis of variance
Estimated regression line: $y = 104.1 + 10.0x_1 - 0.96x_2 2 - 0.03x_3 - 4.32x_4$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	104.076	37.7264
1	10.0219	6.69623
2	-0.956706	0.506024
3	-0.0262767	0.00907681
4	-4.31777	7.18044

	Degrees of freedom	Sum of squares
REGR	4	6665.0
RES	45	24082.1
TOT	49	30747.1

Analysis of variance
Estimated regression line: $y = 32.9 + 9.17x_1 - 0.022x_3$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	32.8860	4.97944
1	9.17159	6.65964
2	-0.0220795	0.00858002

	Degrees of freedom	Sum of squares
REGR	2	4692.56
RES	47	26054.6
TOT	49	30747.1

a) Consider model 1. How does the estrogen treatment seem to affect the survival time? Justify your answer by constructing a suitable confidence interval with confidence level 80%, which is really a very low confidence level.
b) The variable x_3 is the best single explanatory variable (why?), we want x_1 in the model since one of the goals is to see if the estrogen treatment is helping, but is it useful to also have x_2 and x_4? Perform a suitable test on the 5% level.
c) State some advantage with doing regression analysis (with many variables) in the way we did here,

compared to just splitting the patients in two treatment groups (treatment/no treatment) and from that investigate if there is a difference in expected survival time. □

G4.5) The following data illustrates the growth in a cultivation of bacteria

t	y
3	115000
6	147000
9	239000
12	356000
15	579000
18	864000

where t = number of days after inoculation and y = number of bacteria.
Model: $Y_t = e^{\beta_0+\beta_1 t+\varepsilon_t}$ where $\varepsilon_t = N(0, \sigma^2)$.
Construct a 95% prediction interval for the number of bacteria at the time 20, i.e. for Y_{20}.

Analysis of variance
Estimated regression line: $\ln(y) = 11.1 + 0.139t$

i	$\hat{\beta}_i$	$d(\hat{\beta}_i)$
0	11.1499	0.0619
1	0.138993	0.005300

	Degrees of freedom	Sum of squares
REGR	1	3.0428
RES	4	0.0177
TOT	5	3.0605

$$(\mathbf{X}^T\mathbf{X})^{-1} = \begin{pmatrix} 0.866667 & -0.066667 \\ -0.066667 & 0.006349 \end{pmatrix}.$$

□

ANSWERS

G2.6 a) $I_\mu = (50.0, 57.4)$;
b) $I_{\beta_0+100\beta_1} = (53.4, 54.4)$;
c) The linear model explains the data best, a plot and the variance estimates shows this.

G2.7 $I_{\mu_0} = (\hat{\mu}_0 \mp t \cdot s \cdot \sqrt{0.1153}) \approx (4.2, 8.8)$. The values in I_{μ_0} seems low, among other things this could be because the relation between y and x_2 is not linear.

G2.8 $U_0 = 60Y_0$ where $Y_0 = \beta_0 + 250\beta_1 + 5.659\beta_2 + \varepsilon_0$;
$I_{Y_0} = (123.79, 224.03)$ gives $I_{U_0} \approx (7400, 13400)$. The interval is so large that it is questionable whether it is useful or not.

G2.9 a) $I_{\beta_2} = (1.20, 1.70)$ i.e. $0 \notin I_{\beta_2}$ and thus observation 12 seems to be deviant.
b) Using analysis 3 we make a prediction interval for Y_{12} : $I_{Y_{12}} = (140.74, 141.23)$ i.e. $y_{12} = 142.44 \notin I_{Y_{12}}$ and observation 12 seems to be deviant.

G2.10 a) $I_{\beta_1} = (-5.4250 \mp t \cdot 0.2898) = (-6.03, -4.82)$ where $t = 2.10$ is from $t(18)$-table. $-4 \notin I_{\beta_1}$. Thus we can reject H_0 on the 0.05 level.
b) We construct a confidence interval for $\beta_0 + 8\beta_1$: $\hat{\beta}_0 + 8\hat{\beta}_1 = -5.30$; the random variable $\hat{\beta}_0 + 8\hat{\beta}_1 \sim N(\beta_0 + 8\beta_1, 0.15\sigma^2)$ since $V[\hat{\beta}_0 + 8\hat{\beta}_1] = \sigma^2 \begin{pmatrix} 1 & 8 \end{pmatrix} (\mathbf{X}^T\mathbf{X})^{-1} \begin{pmatrix} 1 \\ 8 \end{pmatrix} = 0.15\sigma^2$.
σ^2 is estimated by $s^2 = Q_{res}/18 = 3.666^2$, degrees of freedom: 18
The variable $\frac{\hat{\beta}_0+8\hat{\beta}_1-(\beta_0+8\beta_1)}{s\sqrt{0.15}} \sim t(18)$ and gives
$I_{\beta_0+8\beta_1} = (\hat{\beta}_0 + 8\hat{\beta}_1 \mp 2.10s\sqrt{0.15}) \approx (-8.28, -2.32)$. $0 \notin I_{\beta_0+8\beta_1}$. Thus we can reject the hypothesis that the regression line crosses the x-axis at $x = 8$. Since the interval contains only negative values it seems that eight hours of sleep improves the test result, this could mean that well rested people have an improved ability to learn.

G2.11 a) $I_{\beta_2} = (-23.3432 \mp t \cdot 2.0608) \approx (-29.6, -17.1)$. It seems that state of war gives a lower private consumption.
b) We want a prediction interval for $Y_0 + \beta_0 + 150\beta_1 + \varepsilon_0$:

$m_0 = \begin{pmatrix} 1 & 150 & 0 \end{pmatrix} \boldsymbol{\beta}$; $\hat{m}_0 = \begin{pmatrix} 1 & 150 & 0 \end{pmatrix} \hat{\boldsymbol{\beta}} = 139.61$; note: $\hat{\beta}_0$ = intercept estimate

$$V[\hat{m}_0] = \begin{pmatrix} 1 & 150 & 0 \end{pmatrix} \sigma^2 (\mathbf{X}^T\mathbf{X})^{-1} \begin{pmatrix} 1 \\ 150 \\ 0 \end{pmatrix} = 0.1439\sigma^2.$$

The random variable $Y_0 - \hat{m}_0 \sim N(0, 1.1439\sigma^2)$.
σ^2 is estimated by $s^2 = Q_{RES}/12 = 11.640$, $s = 3.411$, degrees of freedom: 12
The variable $\frac{Y_0-\hat{m}_0}{s\sqrt{1.1439}} \sim t(12)$ and gives
$I_{Y_0} = (\hat{m}_0 \mp t \cdot s \cdot \sqrt{1.1439}) \approx (128.5, 150.7)$ where $t = 3.05$.

G2.12 Prediction interval for the profit $I_{Y_0} = (74, 183)$.

G2.13 a) $v = 285.1 > 9.8$ (From $F(2, 21)$-table). H_0 is rejected.
b) $\beta_1 - \beta_2 = \begin{pmatrix} 0 & 1 & -1 \end{pmatrix} \boldsymbol{\beta}$. The random variable $\hat{\beta}_1 - \hat{\beta}_2 \sim N(\beta_1 - \beta_2, 0.008949\sigma^2)$. $I_{\beta_1-\beta_2} = \begin{pmatrix} -4.18 & -1.90 \end{pmatrix}$. $0 \notin I_{\beta_1-\beta_2}$. The external users seem to put higher load on the computer.

G3.1 a) The relation between y_t and t seems to be a 2nd degree polynomial curve. $V[\varepsilon_t]$ seems to increase as t increases.
b) Estimate $\boldsymbol{\beta}$ by using $Z_t = Y_t/\sqrt{t}$ which gets a regression model without increasing variance. $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Z}$ where $\mathbf{X}$ is the coefficient matrix corresponding to the $\mathbf{Z}$-vector.

G4.1 a) $I_{\beta_3} = (0.27, 0.54)$. $0 \notin I_{\beta_3}$. It seems the number of trains per day has an effect.
b) $H_0 : \beta_1 = \beta_2 = 0$ is tested with $v = 5.22 > 3.99$. H_0 is rejected. With large probability there is a difference between the track types.

G4.2 a) Yes, since the confidence interval for β_1 does not contain 0.
b) Question 1: x_2. Question 2: No. Question 3: The model with x_1 and x_2 as explanatory variables gives the lowest variance estimate.

G4.3 a) x_1 is the best single explanatory variable, since it has the greatest correlation with y. The test statistic $u = 213.069 > 8.29$, thus x_1 is useful as an explanatory variable.
b) The test statistic $v = 12.25 > 8.4$. Thus also use x_2 in the model.

G4.4 a) $I_{\beta_1} = (10.0219 \mp t \cdot 6.69623) = (10.0219 \mp 8.7051) \approx (1, 19)$ where $t = 1.30$ is given in $t(45)$-table. It seems estrogen treatment increases survival time.
b) x_3 has the greatest correlation with y and is therefore the best single explanatory variable. We test model 2 against model 1, i.e.

$$H_0 : \beta_2 = \beta_4 = 0 \quad \text{against} \quad H_1 : \text{at least one of } \beta_2 \text{ and } \beta_4 \neq 0$$

with an F-test. Test statistic

$$v = \frac{(26054.6 - 24082.1)/2}{24082.1/45} = 1.84$$

$F(2, 45)$-table gives critical point ≈ 3.2. $1.84 < 3.2$. H_0 can not be rejected. Model 2 is sufficient.
c) By adding extra explanatory variables in addition to the treatment type, some of the variation can be explained which gives a lower variance estimate and it is then easier to detect treatment effects. It is also possible to find variables that are of importance for the survival time etc.

G4.5 $I_{Z_0} = I_{\ln Y_0} = (13.6858, 14.1736)$ i.e. $I_{Y_0} = (878000, 1431000)$.

11. χ^2-TEST, TEST OF HOMOGENEITY

14∼) a) The lifetimes of 50 vacuum tubes were measured with the result $\bar{x} = 38.5$. We assume that the lifetimes are independent and exponentially distributed with an unknown mean μ. Find a two-sided confidence interval for μ with confidence level 95%.
b) After the observations there were doubts about the assumption of exponential distribution:

Interval	Absolute frequency
$0 \leq x < 20$	14
$20 \leq x < 40$	18
$40 \leq x < 60$	7
$60 \leq x < 80$	6
$80 \leq x$	5

Perform a χ^2-test on the 0.10 level to see if it is reasonable to assume an exponential distribution. □

13.30) The r.v. X takes the values $0, 1, 2, 3$. 4096 independent observations were made of X with the result:

Observation	0	1	2	3
Quantity	1764	1692	552	88

Test on the significance level 1% that $X \in Bin(3, 1/4)$. □

13.31) From each of three large populations of people, P_1, P_2, and P_3, a random sample was made in which the people were classified according to the following table:

	men	women
P_1	46	54
P_2	78	72
P_3	143	107

Perform a test on the 5% significance level that the distribution of genders is the same for the three populations. □

13.32) At some occasion for example when testing if a pattern is random or not, a problem arise in which you have to test if certain data comes from observations on a random variable that has a shifted geometric distribution with $p = 1/2$ i.e.

$$P(X = k) = (1 - p)^{k-1}p, \quad k = 1, 2, \ldots$$

At such an occasion 100 observations were made with the result:

x	1	2	3	4	5	6	7	8
Frequency	42	23	10	11	8	2	3	1

Test the hypothesis that the data comes from a shifted geometric distribution with $p = 1/2$ on the 5% level. □

13.34) At a traffic count, on one of the roads in Sweden, the number of cars passing a certain location in one direction at $9.00 - 9.10$ were counted during 81 weekdays. The following data was obtained:

Number of cars	0	1	2	3	4	5	6
Number of days	14	12	25	16	10	3	1

Test if the number of cars per ten minute period can be considered Poisson distributed. Choose the level 5%. □

13.35) The data in the table below has been collected from 500 traffic accidents on country roads:

Damage	Number of accidents seat belt	
	was used	was not used
None or small injuries	101	143
Severe injuries	58	198

Test on the 5% level if using a seat belt affects the amount of injury. □

ANSWERS

14 ∼ a) $I_u = \bar{x} \mp \lambda_{\alpha/2} \cdot \frac{\bar{x}}{\sqrt{n}} = 38.5 \mp 1.96 \cdot \frac{38.5}{\sqrt{50}} = (27.83, 49.17)$;
b) $u = \sum_{i=1}^{4} \frac{(N_i - np_i)^2}{np_i} = 4.897$,
$R = (\chi^2_\alpha(k - 1 - \#\text{unknowns}), \infty) = (\chi^2_{0.1}(4 - 1 - 1), \infty) = (4.6, \infty)$,
$u \in R$ so reject exponential distribution assumption.

13.30 $Q_{\text{obs}} = 11.5 > \chi^2_{0.01}(3) = 11.3$, H_0 is rejected.

13.31 $Q_{\text{obs}} = 3.77 < \chi^2_{0.05}(2) = 5.99$, H_0 can not be rejected.

13.32 $Q_{\text{obs}} = 15.2 > \chi^2_{0.05}(4) = 9.49$, reject the hypothesis. ($x \geq 5$ merged to one group).

13.34 H_0 : the number of cars per ten minute period is Poisson distributed. μ estimated with the mean 171/81. One parameter estimated! $x \geq 4$ merged (If we use $\mu = 2.1$). $Q_{\text{obs}} = 6.0 < \chi^2_{0.05}(3) = 7.81$. Do not reject the hypothesis.
Note: It is possible to only merge $x \geq 5$ (If we use $\mu = 2.111$ or similar). We then get $Q_{\text{obs}} = 6.61 < \chi^2_{0.05}(4) = 9.49$. Do not reject H_0.

13.35 H_0 : using seat belt does not affect the type of injury. $Q_{\text{obs}} = 20.22 > \chi^2_{0.05}(1) = 3.84$. Reject the hypothesis.