Tentamentsskrivning i **Mathematisk statistik TMA290, 4p.**

Tid: tisdagen den 10 april, 14.00-18.00
Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736 907 613, rum MC 1421.
Hjälpmedel: valfri rknare, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.

There are six questions with the total number of marks 30. Attempt as many questions, or parts of the questions, as you can. Preliminary grading system
grade "3" for 12 to 17 marks,
grade "4" for 18 to 23 marks,
grade "5" for 24 and more marks.

---

1. (5 marks) Of the people passing through an airport metal detector 10% activate it. Let X be the number of people activating the detector among 280 passengers flying from Gothenburg to Frankfurt.

a. Compute the mean μ and standard deviation σ of X and plot the probability density function (frekvensfunktionen) for its normal approximation. Show clearly on the plot the meaning of μ and σ.

b. Find the probability $P(20 \leq X \leq 35)$.

c. State cleary the assumptions you make when computing the answers to the above questions 1a and 1b.

2. (5 marks) The urinary fluoride concentration (measured in ppm) was determined for 11 randomly chosen livestock both at the beginning of and in the middle of their grazing period in a region previously exposed to fluoride pollution:

	1	2	3	4	5	6
Beginning	24.7	46.1	18.5	29.5	26.3	33.9
Middle	12.4	14.1	7.6	9.5	19.7	10.6

	7	8	9	10	11
Beginning	23.1	20.7	18.0	19.3	23.0
Middle	9.1	11.5	13.3	8.3	15.0

a. The eleven differences have sample mean 13.8 and sample variance 66.8. Explain how the number 69.8 is computed and what does it measure.

b. Estimate the population average difference between the two urinary fluoride concentrations. What is the standard error of this estimate?

c. Compute an exact 95% confidence interval for the average difference. What assumption is required here. Does it seem to be a reasonable assumption?

3. (5 marks) True or false? Explain!

a. The center of a 95% confidence interval for the population mean μ is a random variable.

b. A 95% confidence interval for μ contains the sample mean with probability 0.95.

c. A 95% confidence interval contains 95% of the population.

d. Out of one hundred 95% confidence intervals for μ, 95 will contain μ.

4. (5 marks) The relationship between energy consumption and household income was studied, yielding the following data on household income X (in units of \$1000/year) and energy consumption Y (in units of 10^8 Btu/year).

	Energy consumption		Household income
	1.8		20.0
	3.0		30.5
	4.8		40.0
	5.0		55.1
	6.5		60.3
	7.0		74.9
	9.0		88.4
	9.1		95.2
mean	5.8		58.1
$\sum x_i^2$	315		32090
st.dev	2.6		27.1
cov		70.2	

The last three raws in the table summarize the data in a convenient way.

a. Draw a scatterplot of these data.

b. Fit a linear regression model to the data using the least squares method. In your calculations show clearly how the summary statistics in the table give the answer. Draw your regression line on top of the scatter plot to check the answer.

c. Compute the sample correlation coefficient. Its square is the determination coefficient which gives the proportion of variation found in Y explained by the variation in X. What percentage of variation in Y is due to the noise factors (that is other than the main explanatory variable X)?

5. (5 marks) The gamma distribution with parameters α and λ has density $f(x) = \frac{1}{\Gamma(\alpha)}(\lambda x)^{\alpha-1}\lambda e^{-\lambda x}$ for $x > 0$. Its mean and variance are $\mu = \alpha/\lambda$ and $\sigma^2 = \alpha/\lambda^2$. The Figure 2 depicts four gamma distribution curves with parameters (1,1), (1,2), (2,1), and (2,2) (not necessarily in this order).

a. Which of the two parameters should be called a shape parameter? Why the other parameter is called a scale parameter. Explain by referring to the Figure 2.

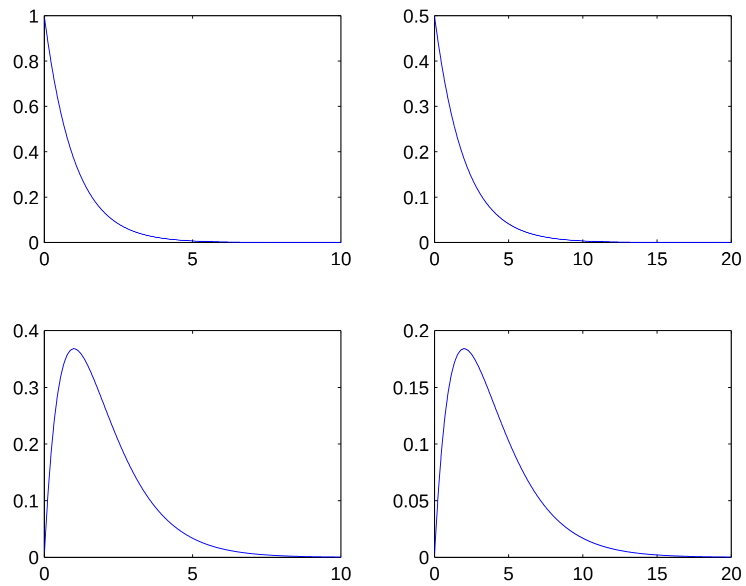

Figure 1: *Gamma distribution.*

b. Express the second moment $E(X^2)$ of the gamma distribution in terms of the parameters α and λ.

c. Using the method of moments estimate parameters α and λ for the household income data in Question 4 under the assumption that the household income has gamma distribution.

6. (5 marks) An urn contains two balls. The prior information about the colors is such that three possibilities are considered to be equally likely

A_0 = no white balls

A_1 = one ball is white the other ball is not white

A_2 = two white balls
Random experiment: place an additional white ball in the urn and choose one ball out of three at random. Suppose the experiment resulted in a white ball = event W.

a. Compute the conditional probability $\mathrm{P}(W|A_2)$.

b. Compute the posterior probability $\mathrm{P}(A_0|W)$.

c. Are events W and A_1 independent? Explain.

d. Draw a Venn diagram containing all four events A_0, A_1, A_2, and W in such a way that probabilities are proportional to the areas covered by the events.

Statistical tables supplied:
1. Normal distribution. 2. t-distribution.

Good luck!

ANSWERS

1a. Using a Binomial distribution Bin(280,0.1) we get the mean $\mu = 28$ and standard deviation $\sigma = \sqrt{280 \cdot 0.1 \cdot 0.9} = 5$. The plot below shows the probability density function for the normal approximation of the binomial distribution. The arrows indicates the points $\mu \pm \sigma = 28 \pm 5$, which are the inflection points of the bell curve.

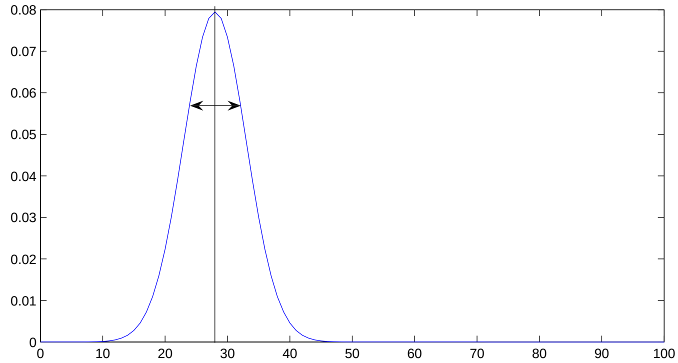

Figure 2: *Normal approximation*

1b. With help of the normal approximation we get

$$P(20 \le X \le 35) \approx P(-1.6 \le Z \le 1.4) = 0.9452 - 1 + 0.9192 = 0.86.$$

1c. To justify the Binomial distribution model we assume that every person activates the detector independently from others and with the same probability. Normal approximation is justified by the large number of trials.

2a. Sample variance

$$\frac{1}{10}\sum_{i=1}^{11}(D_i - 13.8)^2 = 66.8,$$

where $D_1, \ldots, D_{11}$ are the differences in question. This is a measure of the variation of these differences around their mean.

2b. The sample mean 13.8 is an unbiased and consistent estimate of the average difference in question. Its standard error is $\sqrt{66.8/11} = 2.5$.

2c. A 95% CI for the average difference is $13.8 \pm 2.228 \cdot 2.5 = 13.8 \pm 5.5$. This formula assumes that the differences are normally distributed. Looking at the deviations from the sample mean

$$-9.1, -7.2, -5.8, -4.6, -2.9, -2.8, -1.5, 0.2, 6.2, 9.5, 18.2$$

and comparing them to the sample standard deviation $s = 8.2$ we see that even the distribution is somewhat asymmetric there is no major contradiction with the normal distribution model.

3a. True. The center of a CI is the sample mean $\bar{X}$ which is a random variable due to random nature of sampling.

3b. False. The CI always contains the sample mean.

3c. False. The CI concerns the population mean μ and not the population distribution as a whole.

3d. True with the following correction. Out of one hundred 95% confidence intervals for μ, ON AVERAGE 95 will contain μ.

4b. First compute the sample correlation coeeficient $r = \frac{70.18}{2.63 \cdot 27.08} = 0.985$. The linear regression model $y = 0.096x + 0.23$ is based on the least square estimates

$$b_1 = r \cdot \frac{2.63}{27.08} = 0.096, \ b_0 = 5.78 - 0.096 \cdot 58.05 = 0.23.$$

4c. The determination coefficient $r^2 = 0.97$ says that 97% of the variation in the energy consumption is explained by the main explanatory variable - the household income. The remaining three procent of the variation is explained by other factors.

5b. $E(X^2) = \sigma^2 + \mu^2 = \frac{\alpha(1+\alpha)}{\lambda^2}$

5c. Two sample moments $\bar{X} = 58.05$, $\bar{X^2} = 4011.25$ bring two equations

$$\frac{\alpha}{\lambda} = 58.05, \ \frac{\alpha(1+\alpha)}{\lambda^2} = 4011.25.$$

Solving them we get method of moment estimates $\tilde{\alpha} = 5.3$, $\tilde{\lambda} = 0.09$.

6a. If initially there were two white balls, then obviously, after adding another white ball the chosen ball is necessarily white $P(W|A_2) = 1$.

6b. The total probability of chosing a white ball is

$$P(W) = \frac{1}{3}P(W|A_0) + \frac{1}{3}P(W|A_1) + \frac{1}{3}P(W|A_2) = \frac{1/3 + 2/3 + 1}{3} = \frac{2}{3}.$$

Thus

$$P(A_0|W) = \frac{\frac{1}{3}P(W|A_0)}{2/3} = \frac{1}{6}.$$

6c. Similarly,

$$P(A_1|W) = \frac{\frac{1}{3}P(W|A_1)}{2/3} = \frac{1}{3} = P(A_1)$$

implying independence of the events A_1 and W.

6d. The next Vien diagram summarizes our calculations