Tentamentsskrivning i **Mathematisk statistik TMA321, 3p.**

Tid: Lördagen den 21 maj, 2005 kl 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 (including eventual bonus points):
grade "3" for 12 to 17 marks,
grade "4" for 18 to 23 marks,
grade "5" for 24 and more marks.

---

1. (5 marks) Cast iron is an alloy composed primarily of iron together with smaller amounts of other elements. In malleable cast iron the carbon is present as discrete graphite particles. Assume that in a particular casting these particles average 20 per square inch.

a. Would it be unusual to see a 1/4-inch-square area of this casting with fewer than two graphite particles? Explain, based on the probability involved.

b. Justify the parametric model you use in 2a.

2. (5 marks) A test has been developed to detect a particular type of arthritis in individuals over 50 years old. From a national survey it is known that approximately 10% of the individuals in this age group suffer from this form of arthritis. The proposed test was given to individuals with confirmed arthritic disease, and a correct result was obtained in 85% of the cases. When the test was administered to individuals of the same age group who were known to be free of the disease, 4% were reported to have the disease.

a. Consider a random experiment consisting of two steps. First, choose at random an individual over 50 years old. Second, apply the medical test. Four events are of interest: A_1 = the person is free of the disease, A_2 = the person has the disease, A_3 = the test result is positive (indicates the presence of the disease), A_4 = the test result is negative. Find the probabilities $P(A_1)$, $P(A_2)$, $P(A_3|A_1)$, $P(A_3|A_2)$. Draw a Venn diagram.

b. Find $P(A_3)$.

c. What is the probability that the randomly chosen individual has the disease given a positive test result?

3. (5 marks) The Elbe River is important in ecology of central Europe, as it drains much of this region. Due to increased industrialisation, it is feared that the mineral content in the soil is being depleted. This will be reflected in an increase in the level of certain minerals in the water of the Elbe. A study of the river conducted in 1982 indicated that the mean silicon level was 4.6 mg/l.

a. Set up the appropriate null and alternative hypotheses needed to gain evidence to support the contention that the mean silicon concentration in the river has increased.

b. A sample size 28 yields the sample mean $\bar{x} = 5.2$ with the sample standard deviation $s = 1.6$. Find the P value for the test. Do you think that H_0 should be rejected?

4. (5 marks) The paper "Root regeneration and early growth of red oak seedlings: influence of soil temperature" reports the results of a regression analysis in which the independent variable x was daily degree hours of soil heat and the dependent y was shoot elongation per seedling (cm).

x	300	350	400	400	450	450	480	480
y	5.8	4.5	5.9	6.2	6.0	7.5	6.1	8.6

x	530	530	580	580	620	620	670	700
y	8.9	8.2	14.2	11.9	11.1	11.5	14.5	14.8

$\sum x_i = 8140$, $\sum x_i^2 = 4,340,600$
$\sum y_i = 145.7$, $\sum y_i^2 = 1505.01$, $\sum x_i y_i = 79,574$

a. Assuming that a simple linear reghression model is appropriate, fit a regression line to the data using the least squares method. Clearly show your calculations and assumptions.

b. What proportion of variation in shoot elongation is explained by variation in soil heat?

5. (5 marks) A bivariate normal distribution $\mathrm{N}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$ is given by the joint density

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \\ &\times \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2}\right]\right\}. \end{aligned}$$

Suppose a pair of random variables (X, Y) has joint distribution $\mathrm{N}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)$.

a. Show that in the case of zero correlation $\rho = 0$ these two random variables X and Y are independent and normally distributed.

b. Give the formula defining the covariance for two random variables. Using this formula explain why a scatterplot sloped downwards gives a negative covariance.

6. (5 marks) In a recent survey 250 among 1000 interviewed swedes expressed support to an embryonic feminist party.

a. Let p be the all Sweden proportion of the supporters of the feminist party at the period the survey was performed. Find a 99% CI for p.

b. Explain what is random about the interval obtained in 6a, and what is the exact meaning of the 99% degree of confidence.

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

Good luck!

ANSWERS

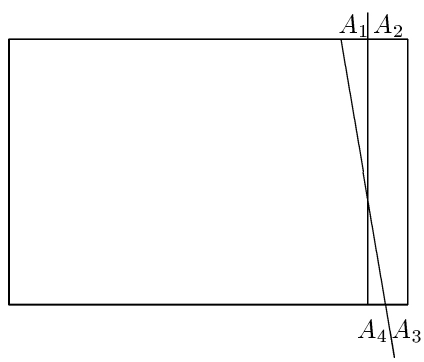
1a. Use the Poisson distribution with parameter $\lambda/4 = 5$, where $\lambda = 20$ particles per square inch. Probability to see less then two particles is then

$$p_0 + p_1 = e^{-5} + 5e^{-5} = 0.04.$$

Quite an unusual result.

1b. Think of a one square inch area as a lattice of huge number n of spots either free or occupied by a particle. If particles are allocated independently at random with a small probability p per a spot, then the number of particles is binomial with parameters (n, p). This is approximely a Poisson number with parameter $\lambda = np$.

2a. $P(A_1) = 0.90$, $P(A_2) = 0.10$, $P(A_3|A_1) = 0.04$, $P(A_3|A_2) = 0.85$.



2b. Law of total probability $P(A_3) = P(A_3|A_1)P(A_1) + P(A_3|A_2)P(A_2) = 0.12$.

2c. Bayes formula $P(A_2|A_3) = \frac{P(A_3|A_2)P(A_2)}{P(A_3)} = 0.70$,

3a. $H_0 : \mu = 4.6$, $H_1 : \mu > 4.6$.

3b. Observed test statistic $T = \frac{5.2-4.6}{1.6/\sqrt{28}} = 1.98$. An one-sided P-value of the test is approximately $P(Z > 1.98) = 1 - 0.9761 = 0.024$. Reject the H_0 at 5% significance level.

4a. Sample size $n = 16$. Sample moments $\bar{x} = 508.75$, $\overline{x^2} = 271,287.5$, $\bar{y} = 9.11$, $\overline{y^2} = 94.06$. Sample standard deviations $s_x = 115.29$, $s_y = 3.44$. Sample covariance $c_{xy} = 361.24$, sample correlation $r = 0.91$. Least squares estimates $b_1 = 0.027$, $b_0 = -4.63$. Fitted regression line $y = 0.027x - 4.63$.

4b. $r^2 = 0.83$.

5a. If $\rho = 0$, then the joint density

$$\begin{aligned} f(x,y) &= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\} \\ &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{\sigma_1^2}\right\} \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(y-\mu_2)^2}{\sigma_2^2}\right\} \end{aligned}$$

is a product of two normal densities

$$\begin{aligned} f_x(x) &= \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left\{-\frac{(x-\mu_1)^2}{\sigma_1^2}\right\} \\ f_y(y) &= \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left\{-\frac{(y-\mu_2)^2}{\sigma_2^2}\right\}. \end{aligned}$$

These are two marginal distributions $\mathrm{N}(\mu_1, \sigma_1^2)$ and $\mathrm{N}(\mu_2, \sigma_2^2)$.

5b. Draw a scatterplot with a negative slope. Cut it into four quadrants through the central point of the plot

2	1
3	4

Points in the quadrants 2 and 4 give positive products $(X - \mu_x)(Y - \mu_y) > 0$, while the points of scatterplot falling in 1 and 3 bring $(X - \mu_x)(Y - \mu_y) < 0$. To finish this picture proof it remains to observe that with a negative slope the majority of points are in the quadrants 2 and 4 leading to a positive covariance $Cov(X,Y) = E((X - \mu_x)(Y - \mu_y)) > 0$.

6a. Point estimate $\hat{p} = 0.25$, standard error $s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} = 0.014$. Approximate 99% CI for p is 0.25 ± 0.035.

6b. Other samples of the same size will produce different CIs. On average 99% of these CIs will cover the true value of p.