EXAM: Matematisk statistik och diskret matematik D (MVE055/MSG810)
Time and place: Tuesday 19 October 2010, kl. 08.30–12.30, V.
Jour: Krzysztof Bartoszek, tel. 0700-771 093.
Aids: Chalmers approved calculater and at most one (double–sided) A4 page of own notes. Tables of appropriate statistical distributions are provided.
Grades: 3: 12 points, 4: 18 points, 5: 24 points. Maximal points : 30.
Motivations: All answers/solutions must be motivated.
Language: There is a Swedish and English version of the questions. You may write your answers in either of these two languages.

1. (3p) Let A and B be two independent events, with $P(A) = 0.6$ and $P(B) = 0.5$.
   a) What is the definition of independence between two events? What is the statistical meaning of independence?
   b) Calculate $P(A \cup B)$ and $P(A \cap B^C)$.
   c) What is the definition of the conditional probability of event G given event H, *i.e.* $P(G|H)$? What is the conditional probability if the two events are independent?

2. (3p) Let X be the outcome of a throw of a six sided dice, *i.e.* X is a random variable taking values in the set $\{1, 2, 3, 4, 5, 6\}$. However the dice is not fair, we have $P(X = 1) = 0.5$, $P(X = 2) = 0.15$, $P(X = 3) = 0.05$ and the probabilities of observing 4, 5 and 6 are equal.
   a) Calculate $P(X = 4)$.
   b) What is $P(X \in \{3, 6\})$?
   c) Let us make two independent throws of this dice and then sum the result. What is the chance of observing an even outcome of at most 5?

3. (2p) Let X be a random variable with density $f_X(x) = Cx^{-\alpha}$ for $X > 0$ and $f_X(x) = 0$ for $X \leq 0$, where $\alpha > 1$ and C is the normalizing constant.
   a) What is the definition of a cumulative distribution function? What is the relationship between a cumulative distribution function and a density function of a continuous random variable?
   b) Calculate the cumulative distribution function of the random variable X.

4. (4p) Let X be an exponential random variable with parameter λ.
   a) What is the density of X? Calculate the mean value of X if $\lambda = 1$.
   b) Let X be an exponential random variable with parameter $\lambda_X = 1$ and Y be an exponential random variable with parameter $\lambda_Y = 1$. Calculate $\mathbf{E}[3X - 7Y]$.
   c) Let X and Y be as in the previous question and in addition assume that they are independent. Calculate the variance of $3X - 7Y$, knowing that $\mathrm{Var}[X] = \mathrm{Var}[Y] = 1$. Clearly state the properties of the variance that you are using. What is (without doing any calculations) the covariance between X and Y?
   d) In general if you know the covariance between two random variables what does this tell you about the dependence between them?

5. (4p) Studies have shown that the random variable X, the processing time required to do a multiplication on a new 3–D computer, is normally distributed with mean μ and standard deviation 2 microseconds. A random sample (in microseconds) of 16 observations is taken. The observed data is :

42.65	45.15	39.32	44.44
41.63	41.54	41.59	45.68
46.50	41.35	44.37	40.27
43.87	43.79	43.28	40.70.

   The sample average is $\overline{x} = 42.88$.

a) Give the definition of an unbiased estimator. Is $\overline{X}$ an unbiased estimator of μ and why or why not? What is the distribution of $\overline{X}$ here?

b) (2p) Derive the formula for a $1-\alpha$ level confidence interval for μ (with the variance known). What is the interpretation of a confidence interval?

c) Using your previous calculations find a 95% confidence interval for μ based on the observed data. Based on this would you be surprised to read that the average time required to process a multiplication on this system is 42.2 microseconds and why?

6. (2p) The time in seconds required to connect to the Internet via a dial–in service is influenced by a variety of factors such as number of phone lines available in the local calling area, time of day, day of the week, number of users in the area, and so on. These data (in seconds) are obtained in a given area at two different times of the day but always on the same day of the week :

X_1 Morning (9 am to 11 am)

10	20	31	42	44
44	15	33	35	47
47	45	22	33	21
51	53	52	37	28
35	56	63	60	48
49	44	57	63	61

X_2 Night (10 pm to midnight)

10	11	21	31	15
40	27	20	32	38
36	22	41	39	24
33	25	35	36	42
43	52	51		

The appropriate averages in the sample are $\overline{x_1} = 41.53$, $\overline{x_1^2} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{1_i}^2 = 1929.00$, $\overline{x_2} = 31.09$ and $\overline{x_2^2} = \frac{1}{n_2}\sum_{i=1}^{n_2} x_{2_i}^2 = 1100.65$

a) Find a 99% confidence interval on the difference in the average time required to access the Internet during these two time periods. Which time period appears to give the fastest average access time and why? The formula for a $1-\alpha$ level confidence interval for $\mu_1 - \mu_2$ is,

$$\hat{\mu}_1 - \hat{\mu}_2 \quad \pm \quad t_{\alpha/2}(n_1+n_2-2)\sqrt{S_p^2\left(\tfrac{1}{n_1}+\tfrac{1}{n_2}\right)},$$

where

$$S_p^2 \quad := \quad \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}.$$

b) What would you need to observe from your calculations for you to conclude that there would be no difference between the two time periods?

7. (3p) A study of the computer market is conducted. Random samples are drawn from among the users of the two leading mainframes. The purpose of the study is to estimate the proportion of users in each population that either do use or would like to use the small office system built by the mainframe supplier. These data result :

Type I $n_1 = 200$, $x_1 = 62$

Type II $n_2 = 190$, $x_2 = 76$

where n_i is the sample size in population i and x_i the number of users that use or would like to use the system in population i, $i = 1, 2$. Denote by p_i the probability of a user from population $i = 1, 2$ that uses or would like to use the small office system built by the mainframe supplier.

a) Find point estimates for p_1, p_2 and $p_1 - p_2$. Are the estimators you used unbiased and why?

b) The formula for a confidence interval for p_i is $\hat{p}_i \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. What approximations were used in the derivation of this confidence interval?

c) The formula for an $1-\alpha$ level confidence interval for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/n_1 + \hat{p}_2(1-\hat{p}_2)/n_2}$. Calculate the 90% confidence interval for $p_1 - p_2$. Based on this would you be surprised to discover that $p_1 = p_2$ and why?

8. (3p) (Ehrenfest Model) The following is a special case of a model called the *Ehrenfest model*, that has been used to explain diffusion of gases. We have two urns that, between them, contain four balls. At each step, one of the four balls is chosen at random (without any preferences) and moved from the urn that it is in into the other urn. We choose, as states, the number of balls in the first urn. Notice that this is a Markov chain that takes values in the state space $\{0, 1, 2, 3, 4\}$.

   a) Draw the graph (with probabilities) representing this Markov chain and write the transition matrix $\mathbf{P}$.

   b) What is an absorbing state and an absorbing Markov chain? Is this Markov chain absorbing? How can you quickly tell this from $\mathbf{P}$?

   c) Calculate $\mathbf{P}^2$. What is the interpretation of $\mathbf{P}^n$?

9. (6p)

   a) State the density function of the binomial distribution with parameters n and p. How is the binomial distribution related to the Bernoulli distribution with parameter p?

   b) Derive the mean value and variance of the binomial distribution with parameters n and p.

   c) (2p) State and prove the Chebyshev inequality.

   d) (2p) Let S_n be the number of success in n Bernoulli trials with probability p for success in each trial. Show that for any $\epsilon > 0$,

   $$P\left(|\tfrac{S_n}{n} - p| \geq \epsilon\right) \leq \tfrac{p(1-p)}{n\epsilon^2}.$$

   Discuss what you think this inequality could be useful for.

Lycka till! Good luck!