Exam: MVE051/MSG810 – Matematisk statistik och diskret matematik, MSG820 – Statistik för fysiker
Time and place: Thursday January 14, 2016, 14:00-18:00, Hörsalsvägen
Examinator: Marcin Lis, tel. 73 75 54 583
Allowed help: Chalmers approved calculator, Beta handbook, Swedish-English dictionary
Grades: Chalmers: 3: 12 points, 4: 18 points, 5: 24 points. GU: G: 12 points, VG: 21 points. Maximal amount of points is 30

You should always justify your answer except in problem 1, where you do not need to do that.
Lycka till!

1. ($6\times 0.5 = 3$ points) This quiz tests your understanding of fundamental notions of probability theory. Decide whether the following statements are **true** or **false**. You do not need to justify your answer.

   (a) For any two events A and B such that $\mathbf{P}(A) = \mathbf{P}(B) = 0.9$, it is true that
   $$\mathbf{P}(A \cap B) \geq 0.8$$
   Answer: True. Since $\mathbf{P}(A\cap B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cup B) \geq 0.9 + 0.9 - 1 = 0.8$.

   (b) For any two events A and B such that $\mathbf{P}(A) = \mathbf{P}(B) = 0.3$, it is true that
   $$\mathbf{P}(A \cup B) \geq 0.6$$
   Answer: False. Take $A = B$. Then, $\mathbf{P}(A \cup B) = \mathbf{P}(A) = 0.3 < 0.6$.

   (c) For any random variable X, it is true that
   $$\mathbf{E}\left[\frac{1}{2}(X + 2016)\right] \geq \mathbf{E}[X]$$
   Answer: False. This inequality is equivalent to $\mathbf{E}[X] \leq 2016$ which is e. g. not true for the constant random variable $X = 2017$.

   (d) For any random variable X, it is true that
   $$\mathrm{Var}\left[\frac{1}{2}(X + 2016)\right] \leq \mathrm{Var}[X]$$
   Answer: True. This inequality is equivalent to $\mathrm{Var}[X] \geq 0$ which is always true.

   (e) If $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ for two events A and B such that $\mathbf{P}(B) > 0$, then A and B are independent.
   Answer: True. The equality is equivalent to $\mathbf{P}(A\cap B) = \mathbf{P}(A)\mathbf{P}(B)$ which is the definition of independence.

   (f) If $\mathbf{E}[XY] = \mathbf{E}[X]\,\mathbf{E}[Y]$ for two random variables X and Y, then X and Y are independent.
   Answer: False. A counterexample was provided during the lecture. Let $X \sim \mathrm{Unif}[-1, 1]$. Then X and X^2 are not independent, but $\mathbf{E}[X^3] = 0 = \mathbf{E}[X]\,\mathbf{E}[X^2]$.

2. ($2 + 2 = 4$ points)

   (a) Compute the expected number of points that an unprepared student gets for solving problem 1 by guessing at random. You can assume that

   - each question is worth 0.5 point,
   - the probability of guessing right for each question is 50%,
   - guesses for different questions are independent of each other.

Solution: Let X_i be the number of points the student gets for answering question number i. The distribution of X_i is given by

$$\mathbf{P}(X_i = 0.5) = 0.5 \qquad \text{and} \qquad \mathbf{P}(X_i = 0) = 0.5.$$

Hence, $\mathbf{E}[X_i] = 0 \times 0.5 + 0.5 \times 0.5 = 0.25$. The total number of points is $X = X_1 + \ldots + X_6$, and hence

$$\mathbf{E}[X] = \mathbf{E}[X_1 + \ldots + X_6] = \mathbf{E}[X_1] + \ldots + \mathbf{E}[X_6] = 6 \times 0.25 = 1.5.$$

(b) Suppose that there are 100 unprepared students who are solving problem 1 independently of each other. Using the central limit theorem, compute the probability that the total number of points that they score is at least 200.

Solution: Let $X_i^{(j)}$ be the number of points the jth unprepared student gets for answering question number i. By the computation above, $\mathbf{E}[X_i^{(j)}] = 0.25$. Moreover

$$\mathrm{Var}[X_i^{(j)}] = \mathbf{E}[(X_i^{(j)})^2] - \mathbf{E}[(X_i^{(j)})]^2 = 0^2 \times 0.5 + 0.5^2 \times 0.5 - 0.25^2 = 1/16.$$

Let $Y = \sum_{j=1}^{100} \sum_{i=1}^{6} X_i^{(j)}$ be the total number of points all students get. We have $\mathbf{E}[Y] = 600 \times 0.25 = 150$, and since all the variables $\mathbf{E}[X_i^{(j)}]$ are independent, $\mathrm{Var}[Y] = 600 \times 1/16 = 37.5$. By the central limit theorem Y is approximately normal with expectation 150 and variance 37.5. Hence,

$$\mathbf{P}(Y \geq 200) = \mathbf{P}(Y - 150 \geq 50) = \mathbf{P}\Big(\frac{Y - 150}{\sqrt{37.5}} \geq \frac{50}{\sqrt{37.5}}\Big) = \mathbf{P}(Z \geq 8.165) < 0.0002 \sim 0,$$

where Z is a standard normal variable.

3. ($2+1+1 = 4$ points) Let X be a continuous random variable with the probability density function

$$f(x) = \begin{cases} (1+x)/2 & \text{for } -1 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Compute the following:

(a) $\mathbf{E}[2X+1]$,

Solution:

$$\mathbf{E}[X] = \int_{-1}^{1} \frac{x(1+x)}{2} dx = \int_{-1}^{1} \frac{x+x^2}{2} dx = \left[\frac{x^2}{4} + \frac{x^3}{6}\right]\Bigg|_{-1}^{1} = 1/3.$$

Hence, $\mathbf{E}[2X+1] = 2\,\mathbf{E}[X] + \mathbf{E}[1] = 2\,\mathbf{E}[X] + 1 = \frac{5}{3}$.

(b) $\mathbf{P}(X > 0)$,

Solution:

$$\mathbf{P}(X > 0) = \int_{0}^{1} \frac{1+x}{2} dx = \left[\frac{x}{2} + \frac{x^2}{4}\right]\Bigg|_{0}^{1} = \frac{1}{2} + \frac{1}{4} = \frac{3}{4}.$$

(c) $\mathbf{P}(X = -1/2 \text{ or } X = 1/2)$.

Solution: Since $\{X = -1/2\}$ and $\{X = 1.2\}$ are disjoint, and by the property of continuous variables:

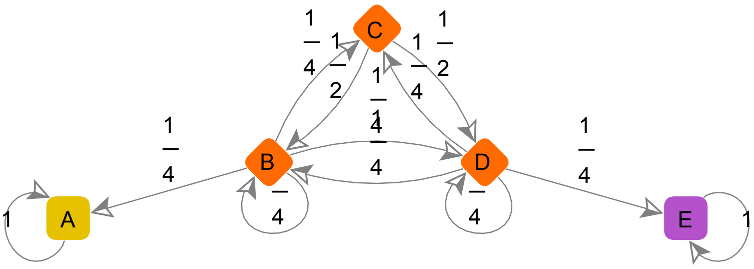
$$\mathbf{P}(X = -1/2 \text{ or } X = 1/2) = \mathbf{P}(X = -1/2) + \mathbf{P}(X = 1/2) = 0 + 0 = 0.$$

4. ($2+1 = 3$ points) Let

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/4 & 1/4 & 1/4 & 1/4 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 1/4 & 1/4 & 1/4 & 1/4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

be the transition matrix of a Markov chain on the state space $\{A, B, C, D, E\}$.

(a) Draw the state diagram of the chain. Identify all transient and absorbing states. Is this chain absorbing?
Solution: The state diagram is:



The absorbing states are A and E, and the transient states are B, C, and D. The chain is absorbing since from every transient state there is a path to an absorbing state.

(b) Argue without computations why the probability that the chain started in state C is absorbed in state E is $1/2$.

Solution: For every trajectory of the chain starting in C and ending in A, there is a mirror-image trajectory from C to E with the same probability. This means that the probability of absorption in A and E are the same. Since they sum up to 1, they are equal to 1/2.

5. (3 points) Compute the generating function of the sequence $(a_n)_{n=1}^{\infty}$ which is given by the following recursive formula:

$$a_0 = 0, \qquad a_1 = 1, \qquad \text{and} \qquad a_n = 3a_{n-1} + 2a_{n-2}, \qquad \text{for } n \geq 2.$$

You **do not need** to compute the sequence $(a_n)_{n=1}^{\infty}$ itself.
Solution: From the recursive formula we get a sequence of equations:

$$\begin{aligned} a_2x^2 &= 3x \cdot a_1x + 2x^2 \cdot a_0 \\ a_3x^3 &= 3x \cdot a_2x^2 + 2x^2 \cdot a_1x \\ a_4x^4 &= 3x \cdot a_3x^3 + 2x^2 \cdot a_2x^2 \\ &\vdots \\ a_nx^n &= 3x \cdot a_{n-1}x^{n-1} + 2x^2 \cdot a_{n-2}x^{n-2} \\ &\vdots \end{aligned}$$

Summing everything up, we get $A(x) - a_1x - a_0 = 3x(A(x) - a_0) + 2x^2A(x)$. We put $a_0 = 0$ and $a_1 = 1$ to get $A(x) - x = A(x)(3x + 2x^2)$. We get

$$A(x) = \frac{x}{1 - 3x - 2x^2}.$$

6. $(1 + 3 = 4$ points) A doctor claims that the weight of a newborn baby has grown compared to the last year and is now 3.7 kg. He weighs a sample of six babies and gets the following results

$$3.5, \quad 3.6, \quad 4.0, \quad 4.1, \quad 3.8, \quad 4.2.$$

(a) Write down the research and null hypothesis. Describe the assumptions that you have to make concerning the data.

Solution: $H_0 : \mu \leq 3.7$ (or $H_0 : \mu = 3.7$), $H_1 : \mu > 3.7$. We have to assume that the data is normally distributed.

(b) Compute the p-value of the observed results.
Solution: The test statistic that we will use is

$$\frac{\overline{X} - \mu_0}{S}\sqrt{n} \sim T_{n-1} = T_5.$$

We compute the sample mean $\overline{x} = 3.8666$ and the sample standard deviation $s = 0.2804$. Hence, the observed value of the test statistic is

$$t = \frac{3.87 - 3.7}{0.2804}\sqrt{6} = 1.456.$$

The p-value is the probability of seeing results as extreme as the observed ones. From the tables for the distribution of T_5 we get $\mathbf{P}(T_5 > 0.727) = 0.25$ and $\mathbf{P}(T_5 > 1.476) = 0.9$.
The p-value is hence between 0.1 and 0.25 (closer to 0.1).

7. $(2 + 3 = 5$ points) John wants to know how much fuel his car uses on average. He makes five trips and writes down the distance traveled and the amount of fuel used. Here are his results:

distance [km]	100	150	300	350	400
used fuel [l]	5.9	7.6	14.6	17.2	19.8

(a) Write down the assumptions of the linear regression model.

Solution: The assumptions of the linear regression model say that the mean value of the dependent variable is a linear function of the independent variable, the distribution of the dependent variable is normal with constant variance, and that the dependent variables corresponding to different values of the independent variable are independent (as random variables) of each other.

(b) Estimate the amount of fuel used per 100km by performing linear regression.
Hint: Use 100km as the unit of distance.

Solution: The amount of fuel used per 100km is the slope β_1 of the regression line. We use the formula from the lecture/book

$$\beta_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

with y_i being the amount of fuel and x_i the distance. We use 100km as the unit of distance and get

$$n = 5, \quad \sum x_i y_i = 200.5, \quad \sum x_i = 13, \quad \sum x_i^2 = 40.5, \quad \sum y_i = 65.1.$$

We plug in the numbers into the formula to get $\beta_1 = 4.66$.

8. $(2 + 2 = 4$ points)

(a) Let $X_1, X_2, X_3, \ldots$ be a sequence of independent random variables with the same dsitribution X ($X_i \sim X$ for $i = 1, 2, 3, \ldots$). Assume that

$$\mathbf{E}[X] = 0 \quad \text{and} \quad \mathrm{Var}[X] = 2016,$$

and let

$$S_n = \frac{X_1 + X_2 + \ldots + X_n}{n}.$$

Using Chebyshev's inequality, find N such that

$$\mathbf{P}(|S_n| \geq 10) \leq 0.01 \qquad \text{for all } n \geq N.$$

Solution: We have

$$\mathbf{E}[S_n] = \mathbf{E}\,\big[\frac{1}{n}\sum_{i=1}^{n} X_i\big] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}[X_i] = 0,$$

and since X_i are independent

$$\mathrm{Var}[S_n] = \mathrm{Var}\,\big[\frac{1}{n}\sum_{i=1}^{n} X_i\big] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}[X_i] = 2016/n.$$

Applying Chebyshev's inequality to S_n, we get

$$\mathbf{P}(|S_n - \mathbf{E}[S_n]| \geq 10) = \mathbf{P}(|S_n| \geq 10) \leq \frac{\mathrm{Var}[S_n]}{10^2} = \frac{2016}{100n}.$$

This is smaller than 0.01 for $n \geq 2016 = N$.

(b) Let X be continuous random variable satisfying $\mathbf{E}[X] = 0$. Prove that for all $\varepsilon > 0$,

$$\mathbf{P}(|X| \geq \varepsilon) \leq \frac{\mathbf{E}[X^4]}{\varepsilon^4}.$$

Hint: Modify the proof of Chebyshev's inequality.

Proof. Let f be the pdf of X. We have

$$\begin{aligned}\mathbf{E}[X^4] &= \int_{-\infty}^{\infty} x^4 f(x)dx = \int_{|x|\geq\varepsilon} x^4 f(x)dx + \int_{|x|<\varepsilon} x^4 f(x)dx \geq \int_{|x|\geq\varepsilon} x^4 f(x)dx \\ &\geq \int_{|x|\geq\varepsilon} \varepsilon^4 f(x)dx = \varepsilon^4 \int_{|x|\geq\varepsilon} f(x)dx = \varepsilon^4 \mathbf{P}(|X| \geq \varepsilon).\end{aligned}$$

□