Solutions for the exam Matematisk statistik och diskret matematik (MVE050/MVE051/MSG810). Statistik ör fysiker (MSG820). 19 December 2013.

1. (2p) Assume that X and Y are independent random variables and that $\mathbf{E}[X] = \mathbf{E}[Y] = \mathbf{Var}[X] = \mathbf{Var}[Y] = 0.5$. Choose the correct statement, and motivate your choice:

   (a) $\mathbf{E}[2 * X + 2 * Y] = 2, \mathbf{Var}[2 * X - 2 * Y] = 0,$

   (b) $\mathbf{E}[2 * X - 2 * Y] = 0, \mathbf{Var}[2 * X + 2 * Y] = 2,$

   (c) $\mathbf{E}[2 * X + 2 * Y] = 2, \mathbf{Var}[2 * X - 2 * Y] = 4,$

   (d) none of the above is correct.

   Solution: (a) is incorrect: $\mathbf{Var}[2 * X - 2 * Y] = 4\,\mathbf{Var}[X] + 4\,\mathbf{Var}[Y] = 2 + 2 = 4$.
   (b) is incorrect: $\mathbf{Var}[2 * X + 2 * Y] = 4\,\mathbf{Var}[X] + 4\,\mathbf{Var}[Y] = 2 + 2 = 4$.
   (c) is correct: $\mathbf{E}[2 * X + 2 * Y] = 2\mathbf{E}[X] + 2\mathbf{E}[Y] = 1 + 1 = 2, \mathbf{Var}[2 * X - 2 * Y] = 4\,\mathbf{Var}[X] + 4\,\mathbf{Var}[Y] = 2 + 2 = 4$.
   (d) is incorrect.

2. (2p) Assume that 100 people have answered Problem 1 independently, each choosing one of the four possible options uniformly at random, so that every option has the same probability to be chosen. What can you say about the distribution of X, the total number of people, who guessed the correct answer? Choose the two correct statements (there are exactly two):

   (a) X is Binomially distributed with parameters $n = 100$, $p = 0.25$.

   (b) X is Binomially distributed with parameters $n = 4$, $p = 0.5$.

   (c) X is Binomially distributed with parameters $n = 100$, $p = 0.5$

   (d) X is Binomially distributed with some other parameters.

   (e) The distribution of X can be approximated by a Normal distribution with parameters $\mu = 100$, $\sigma = 0.25$

   (f) The distribution of X can be approximated by a Normal distribution with parameters $\mu = 25$, $\sigma^2 = 18.75$

   (g) The distribution of X can be approximated by a Normal distribution with parameters $\mu = 50$, $\sigma^2 = 25$

   (h) The distribution of X can be approximated by a Normal distribution with some other parameters.

   Solution: (a) is correct: X is the total number of successes in a sequence of 100 independent trials, each having a probability of success 0.25.
   (f) is correct: the Binomial distribution with parameters $n = 100$, $p = 0.25$ can be approximated by a Normal distribution with parameters $\mu = np = 25$ and $\sigma^2 = np(1-p) = 100 * 0.25 * 0.75 = 18.75$.

3. (2p) Find the probability of guessing a correct answer for a Problem 2, if one chooses the answer uniformly at random among all possible combinations of two options. (Hint: use the classical definition of probability)

Solution: The total number of combinations of two options out of eight possible is given by $\binom{8}{2} = \frac{8!}{6!\cdot 2!} = \frac{8\cdot 7}{2\cdot 1} = 28$. Only one combination is correct, so by classical definition of probability, the probability of guessing an answer randomly in the conditions specified is $1/28$.

4. (4p) Assume that 100 students answer the Problem 1 independently of each other, and that each has the same probability of getting a correct answer. Denote that probability by p. Out of curiosity, Anton wants to test a hypothesis $H_0 : p = 0.25$, corresponding to the situation where everybody attempts to guess the answer, choosing one of the options uniformly at random, against the alternative $H_1 : p > 0.25$.

   a) Anton uses the test statistic $\hat{p}$. Find the critical region for it on the significance levels $\alpha = 0.05$ and $\alpha = 0.01$.
   b) Assume that out of 100 students, 34 have answered correctly. Find the results of the hypothesis test on $\alpha = 0.05$ and $\alpha = 0.01$. What is the p-value of the corresponding tests?

   Solution:

   a) The alternative H_1 has a form $p > 0.25$, so the critical region is one-sided. The critical region for the significance level α is an interval $[L, 1]$, where L is such that

   $$\mathbf{P}(\hat{p} > L | H_0) = \alpha$$

   Due to the Central Limit Theorem, we can approximate the distribution of $\hat{p}$ with $Normal(\mu = p_0, \sigma^2 = p_0(1-p_0)/n)$, so that

   $$\mathbf{P}(\hat{p} > L | H_0) = \mathbf{P}(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0}/n} > \frac{L - p_0}{\sqrt{p_0(1-p_0)/n}}) = \mathbf{P}(Z > \frac{L - p_0}{\sqrt{p_0(1-p_0)/n}}) = \alpha,$$

   where $Z \sim \text{Normal}(0, 1)$, so

   $$\frac{L - p_0}{\sqrt{p_0(1-p_0)/n}} = \zeta_\alpha$$

   and, finally, $L = p_0 + \zeta_\alpha \sqrt{\frac{p_0(1-p_0)}{n}}$ which becomes $0.25 + 1.645 \cdot \sqrt{\frac{0.25\cdot 0.75}{100}} = 0.32$ for $\alpha = 0.05$ and $0.25 + 2.33 \cdot \sqrt{\frac{0.25\cdot 0.75}{100}} = 0.35$ for $\alpha = 0.01$.
   Answer: $C = [0.32, 1]$ for $\alpha = 0.05$ and $C = [0.35, 1]$ for $\alpha = 0.01$.
   b) The value of $\hat{p}$ obtained from data is then $\hat{p} = 0.34$, which is inside the critical region for $\alpha = 0.05$, and outside of the critical region for $\alpha = 0.01$. That means, we reject H_0 at significance level $\alpha = 0.05$, but fail to reject H_0 at significance level $\alpha = 0.01$. The correspondent p-value is equal to:

   $$\text{p-val} = \mathbf{P}(\hat{p} > 0.34 | H_0) = \mathbf{P}\left(\frac{\hat{p} - 0.25}{\sqrt{0.25(1-0.25}/100} > \frac{0.34 - 0.25}{\sqrt{0.25(1-0.25)/100}} \Big| H_0 \right)$$

   $$= P(Z > 2.08) = 0.0188,$$

   where $Z \sim \text{Normal}(0, 1)$.

5. (3p) Assume the total proportion of the students that fail the exam is p. Anton starts grading the exams one by one, works until he grades the first failed exam, then goes for a coffee.

   a) What is the distribution of the number of works Anton is going to grade before his first cup of coffee? (you can assume that different students' results are independent of each other, and that the total ammount of exams to grade is absurdly immense, or even infinite)
   b) Find the probability (write a formula) that Anton will grade at least 3 works before having a coffee.

   Solution:

   a) This number (denote it by X) has Geometric distribution with parameter p: it is a distribution of a number of trials before the first success in the series of independent Bernoulli(p) trials.
   b) $\mathbf{P}(X \geq 3) = 1 - \mathbf{P}(X \geq 2) = 1 - \mathbf{P}(X = 1) - \mathbf{P}(X = 2) = 1 - p - (1-p)p = 1 - 2p + p^2$.

6. (5p) Assume that the time Y needed to finish all the 9 questions of the exam is distributed normally, with parameters $\mu = 3.5$ hr, $\sigma = 1$ hr.

   a) If 4 hours are given to complete the exam, what is the probability to be on time?
   b) Again, assume $n = 100$ people are taking the exam. Given the probability p from part 'a)', what is the exact distribution of X, the number of students who finish on time? (Hint: let 'finishing on time' correspond to the 'success' in the series of 100 independent experiments)
   c) If X denotes the number of people who finish on time, find $\mathbf{P}(X > 50)$. (Hint: use the Normal approximation)

   Solution:

   a)
   $$\mathbf{P}(Y < 4) = \mathbf{P}\left(\frac{Y - \mu_Y}{\sigma_Y} < \frac{4 - \mu_Y}{\sigma_Y}\right) = \mathbf{P}(Z < 0.5) = 0.6915,$$
   where $Z \sim \text{Normal}(0, 1)$.
   b) X is distributed Binomially with parameters $n = 100$ and $p = 0.6915$.
   c) Using the Central Limit Theorem, we can approximate the distribution of X by a Normal distribution with parameters $\mu = np = 69.15$ and $\sigma = \sqrt{np(1-p)} = \sqrt{21.33} = 4.62$. Then
   $$\mathbf{P}(X > 50) = \mathbf{P}(\frac{X - 69.15}{4.62} < \frac{50 - 69.15}{4.62}) = \mathbf{P}(Z < -4.15) = \Phi(-4.15) < 0.0001$$
   where $Z \sim \text{Normal}(0, 1)$ and Φ is the cdf of the Standard Normal distribution.

7. (4p) Beth is not sure about the answer for the Problem 1. Her mind goes wondering according to a Markov chain (X_n), starting in $X_0 =$ 'a)', governed by the following

transition matrix:

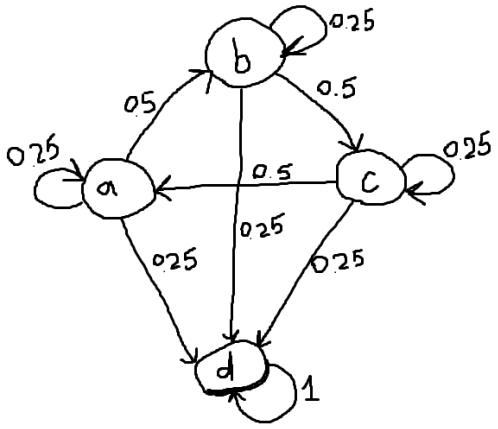
$$A = \begin{array}{c} \\ \text{a)} \\ \text{b)} \\ \text{c)} \\ \text{d)} \end{array} \begin{array}{c} \begin{array}{cccc} \text{a)} & \text{b)} & \text{c)} & \text{d)} \end{array} \\ \begin{pmatrix} 0.5 & 0.25 & 0 & 0.25 \\ 0 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

Beth's mind wonders for a little while (2 steps of the Markov chain), and the answer she chooses is then given by X_2, the state of the Markov chain after 2 steps.

a) Sketch a state diagram of the corresponding Markov chain.
b) Which answer will Beth pick with the maximal probability? (Hint: find the distribution vector of a MC's state after two steps, pick the state with the maximal probability)
c) Which answer would Beth pick eventually, if her mind was given an infinite amount of time to wonder (that is, if the corresponding Markov chain made a very large number of steps), and why?

Solution:

a)



b) Let us find a distribution vector after two steps:

$$\vec{u}^2 = \vec{u}^0 A^2 = (1,0,0,0,) \begin{pmatrix} 0.5 & 0.25 & 0 & 0.25 \\ 0 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0 & 1 \end{pmatrix}^2$$

$$= (0.5, 0.25, 0, 0.25,) \begin{pmatrix} 0.5 & 0.25 & 0 & 0.25 \\ 0 & 0.5 & 0.25 & 0.25 \\ 0.25 & 0 & 0.5 & 0.25 \\ 0 & 0 & 0 & 1 \end{pmatrix} = (0.25, 0.25, 0.0625, 0.4375),$$

so the option Beth will pick with maximum probability 0.4375 is (d).

c) The Markov chain in question would always eventually end up in its only absorbing state (d).

8. (4p) Assume that p is the total proportion of statistical problems that Beth can solve.

   a) How many questions does the final exam have to contain in order to build a 95% confidence interval for that proportion of length at most 0.2? You can assume that Beth solves different questions independently of each other, each with probability p.
   b) Would that number increase or decrease if we had the prior estimate of p?

   Solution:

   a) A 95% two-sided confidence interval for that proportion has the following form:

$$\left[\hat{p} - \zeta_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + \zeta_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right],$$

   so its length is given by

$$2 \cdot \zeta_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

   Solve the inequality for n:

$$2 \cdot \zeta_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < 0.2,$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < \frac{0.1}{\zeta_{0.025}},$$

$$\frac{\hat{p}(1-\hat{p})}{n} < \frac{0.1^2}{\zeta_{0.025}^2},$$

$$n > \frac{\zeta_{0.025}^2 \cdot \hat{p}(1-\hat{p})}{0.1^2}.$$

   We don't have a prior estimate for p, but the expression $\hat{p}(1-\hat{p})$ can't be larger than 0.25, so $n > 1.96^2 * 0.25/0.01 = 96.04$ should be enough (which is, probably, slightly too many for practical purposes).
   b) Given some prior estimate of p, we could possibly decrease the number of questions needed, since in that case we would not have to use the worst-case estimate $\hat{p}(1-\hat{p}) < 0.25$, but instead could plug in the numerical value $\hat{p}$.

9. (4p) Alice, Bob, Claire and Dean decide to cooperate, dividing the 9 questions between themselves. Alice wants to get between 3 and 9 questions, Bob wants to get between 0 and 4, and Clair and Dean have no preferences whatsoever.

   a) Denote by y_1, y_2, y_3, y_4 the respective number of questions each of the four students gets, and write the corresponding diophantine equation with the constraints.
   b) Write the generating function for that combinatorical problem.
   c) How many ways is there to make a division so that everyone is happy?

   Solution:

a)

$$\begin{cases} y_1 + y_2 + y_3 + y_4 = 9, \\ y_1 \geq 3, \\ 0 \leq y_2 \leq 4, \\ y_3 \geq 0, \\ y_4 \geq 0. \end{cases}$$

b)

$$A(x) = (x^3 + x^4 + ...)(1 + x + x^2 + x^3 + x^4)(1 + x + x^2 + ...)(1 + x + x^2 + ...)$$

$$= \frac{x^3}{1-x}\frac{1-x^5}{1-x}\frac{1}{1-x}\frac{1}{1-x} = \frac{(x^3 - x^8)}{(1-x)^4}$$

c) Let us represent $A(x)$ as power series $\sum_{n=0}^{\infty} a_n x^n$ and find a coefficient a_9. Use the formula $\frac{1}{(1-x)^{k+1}} = \sum_{n=0}^{\infty} \binom{n+k}{k} x^n$

$$A(x) = \frac{(x^3 - x^8)}{(1-x)^4} = (x^3 - x^8) \sum_{n=0}^{\infty} \binom{n+3}{3} x^n,$$

so $a_9 = \binom{6+3}{3} - \binom{1+3}{3} = \frac{9 \cdot 8 \cdot 7}{3 \cdot 2 \cdot 1} - \frac{4}{1} = 84 - 4 = 80$, so there is 80 ways to make a division.