Solutions for the exam for Matematisk statistik och diskret matematik (MVE050/MSG810). Statistik för fysiker (MSG820).
December 15, 2012.

1. (3p) The joint distribution of the discrete random variables X and Y is given by the density table:

$x\backslash y$	0	1
1	0.4	0.2
2	0.1	0.3

   a) Find $\mathbf{E}[X]$, $\mathbf{E}[Y]$, $\mathbf{E}[XY]$
   b) Find $\mathbf{Cov}(X, Y)$
   c) Find $\mathbf{E}[5X - Y + 17]$

   Solution:

   a)

$$\mathbf{E}[X] = 1 * (0.4 + 0.2) + 2 * (0.1 + 0.3) = 1.4,$$

$$\mathbf{E}[Y] = 0 * (0.4 + 0.1) + 1 * (0.2 + 0.3) = 0.5,$$

$$\mathbf{E}[XY] = 0.4 * 0 * 1 + 0.2 * 1 * 1 + 0.1 * 0 * 2 + 0.3 * 1 * 2 = 0.8$$

   b) $\mathbf{Cov}(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] = 0.8 - 1.4 * 0.5 = 0.1$
   c) $\mathbf{E}[5X - Y + 17] = 5 * \mathbf{E}[X] - \mathbf{E}[Y] + 17 = 5 * 1.4 - 0.5 + 17 = 23.5$

2. (3p) Draw a card at random from a standard 52-card deck. Denote $A = \{\text{draw a nine}\}$, $B = \{\text{draw a diamond}\}$, $C = \{\text{draw a number}\}$.

   a) Are the events B and C independent? A and C? Why?
   b) Find $\mathbf{P}(A \cap B|C)$.
   c) Remove one card (the Ace of spades) from the deck. Are A and B independent now? Why?

   Solution:

   a) B and C are independent:

$$\mathbf{P}(B \cap C) = \mathbf{P}(\text{draw a number which is a diamond}) = 9/52,$$

$$\mathbf{P}(B) = 1/4, \mathbf{P}(C) = 9/13,$$

   so $\mathbf{P}(B \cap C) = \mathbf{P}(B)\mathbf{P}(C)$ – a definition of independent events.
   Similarly, show that A and C are not independent:

$$\mathbf{P}(A \cap C) = 1/52 \neq 1/52 * 9/52 = \mathbf{P}(A) * \mathbf{P}(C).$$

   b)

$$\mathbf{P}(A \cap B|C) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(C)} = \frac{1/52}{9/13} = \frac{1}{36}.$$

c) No, if we remove one card, A and B are not independent anymore. Now:

$$\mathbf{P}(A \cap B) = 1/51 \neq 4/51 * 13/51 = \mathbf{P}(A) * \mathbf{P}(B).$$

3. (2p) The continuous random variable X has a density $f_X(x)$:

$$f_X(x) = \begin{cases} \frac{x}{a}, & 20 < x < 40, \\ 0, & \text{otherwise.} \end{cases}$$

a) Find the value of a so that $f_X(x)$ becomes a proper density function.
b) Using the obtained value of a, find the probability $\mathbf{P}(X < 25)$.

Solution:

a) For $f_X(x)$ to be a proper density function, it must be non-negative and the total area under the curve must be equal to 1:

$$\int_{-\infty}^{+\infty} f_X(x)\,dx = \int_{20}^{40} \frac{x}{a}\,dx = \frac{x^2}{2a}\bigg|_{20}^{40} = \frac{1600 - 400}{2a} = \frac{600}{a} = 1,$$

so $a = 600$.

b)

$$\mathbf{P}(X < 25) = \int_{-\infty}^{25} f_X(x)\,dx = \int_{20}^{25} \frac{x}{600}\,dx = \frac{x^2}{1200}\bigg|_{20}^{25}$$

$$= \frac{625 - 400}{1200} = \frac{3}{16} = 0.1875$$

4. (4p) Assume the total proportion p of red-haired people in Sweden is 0.015. The dentist examines $n = 10$ people every day, which you can assume to be independent of each other.

a) What is the probability for the dentist to meet at least 2 red-haired individuums during the same day?
b) What is the probability to examine at least 1 red-haired person every day during the whole working week (5 days)?
c) What is the expected number of days a new dentist has to work before meeting the first red-haired person?

Solution:

a) Denote by X the total number of red-haired people a doctor examines during one day. X can be thought of as a Binomially$(n = 10,\, p = 0.015)$ distributed random variable. We are interested in the probability:

$$\mathbf{P}(X \geq 2) = 1 - \mathbf{P}(X \leq 1) = 1 - \mathbf{P}(X = 0) - \mathbf{P}(X = 1)$$

$$= 1 - \binom{10}{0}(0.015)^0(0.985)^{10} - \binom{10}{1}(0.015)^1(0.985)^9 = 0.009$$

b) First, find the probability to meet at least one red-haired person during one day.

$$\mathbf{P}(X \geq 1) = 1 - \mathbf{P}(X = 0) = 1 - \binom{10}{0}(0.015)^0(0.985)^{10} = 1 - 0.860 = 0.140$$

Now, assuming the different days being independent of each other, we can argue:

$$\mathbf{P}(\text{meet at least 1 on every one of 5 working days})$$

$$= \mathbf{P}(\bigcap_{i=1}^{5}\{\text{meet at least 1 red-haired person on } i\text{-th day}\})$$

$$= (\mathbf{P}(X \geq 1))^5 = (0.140)^5 \approx 0.0000538$$

c) There is several ways to approach this problem, here is one of them. Denote by Y the number of the first day the doctor meets a red-haired patient. We can see the sequence of days as a sequence of independent Bernoulli(0.140) trials, since the probability to meet at least one red-haired patient during one day is 0.140 (from part b). Therefore, Y has a Geometric distribution with parameter 0.140, hence

$$\mathbf{E}Y = \frac{1}{0.140} = 7.14$$

5. (3p) Bad thoughts are popping up in Peter's head according to a Poisson process with intensity $\lambda = 2$ [tph] (thoughts per hour). When bad thoughts are there, Peter becomes uneasy and can not eat.

a) It takes Peter 40 minutes to eat his lunch. What's the probability for him to eat the lunch in peace, without any bad thoughts?
b) If Peter wakes up 6:30 and has a breakfast at 7, what's the probability for him to get at least 3 bad thoughts before breakfast?

Solution:

a) The amount X of bad thoughts that come to Peter during lunch has a Poisson distribution with parameter λt, where $\lambda = 2$ and $t = 2/3$. We are interested in the probability

$$\mathbf{P}(X = 0) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-4/3} \approx 0.264$$

b) Again, the amount of bad thoughts $X \sim \text{Pois}(\lambda t)$. Now t is 0.5, $\lambda t = 1$, and we want the probability

$$\mathbf{P}(X \geq 3) = 1 - \mathbf{P}(X \leq 2) = 1 - e^{-1}\frac{1^0}{0!} - e^{-1}\frac{1^1}{1!} - e^{-1}\frac{1^2}{2!}$$

$$= 1 - e^{-1}(1 + 1 + 0.5) = 1 - 2.5e^{-1} \approx 0.0803$$

6. (3p) Find a generating function $A(x)$ of the sequence (a_n), defined by the following recursion:

$$\begin{cases} a_0 = 2, \\ a_1 = 2, \\ a_n = a_{n-1} + a_{n-2} - 1, \quad n = 2, 3, \dots \end{cases}$$

Note: you don't have to find the (a_n) itself, just the generating function.

Solution:
Start with writing out the recursion for $n = 2, 3, \dots$ and multiplying with the corresponding power of x:

$$\begin{aligned} &a_2 = a_1 + a_0 - 1, \quad *x^2 && a_2x^2 = xa_1x + x^2a_0 - x^2 * 1 \\ &a_3 = a_2 + a_1 - 1, \quad *x^3 && a_3x^3 = xa_2x^2 + x^2a_1x - x^2 * x \\ &a_4 = a_3 + a_2 - 1, \quad *x^4 && a_4x^4 = xa_3x^3 + x^2a_2x^2 - x^2 * x^2 \\ &\dots && \dots \end{aligned}$$

Sum up:

$$a_2x^2 + a_3x^3 + a_4x^4 + \dots = x(a_1x + a_2x^2 + a_3x^3 + \dots)$$
$$+x^2(a_0 + a_1x + a_2x^2 + \dots) - x^2(1 + x + x^2 + \dots)$$

Now, remembering that $\sum_{n=0}^{\infty} a_nx^n = A(x)$ and $\sum_{n=0}^{\infty} x^n = \frac{1}{1-x}$, we get:

$$A(x) - a_1x - a_0 = x(A(x) - a_0) + x^2A(x) - x^2\frac{1}{1-x},$$

or, putting in the values for a_0, a_1:

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

Combine the terms with $A(x)$ on the left and the rest on the right:

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

so

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

7. (4p) Stefan is on a quest to find out the average IQ of a seagull. Omitting details about how he measures the IQ of his involuntary subjects, we present his data:

7.5 7.0 8.5 4.0 6.0 7.0 5.5

a) Help Stefan construct the 95% confidence interval for μ – the actual average IQ of a seagull.
b) What's the length of your interval? Based on your estimate for the actual variance, how many seagulls does Stefan have to catch for his 95% confidence interval to be of length <1?

Solution:

a) We assume Stefan's data to be Normal(μ, σ) with some unknown parameters. That means, the quantity $T = \frac{\bar{X}-\mu}{s/\sqrt{n}}$ has t-distribution with $n-1$ degrees of freedom (in our case, $n = 7$). That fact leads us to the formulas for the left and right border of the two-sided 95% CI:

$$L = \bar{X} - t_{6,0.025}\frac{s}{\sqrt{7}}, R = \bar{X} + t_{6,0.025}\frac{s}{\sqrt{7}}$$

Looking up $t_{6,0.025} = 2.447$ in the table and having $\bar{X} = 6.5$ and $s = 1.472$ found from the data, we obtain the confidence interval:

$$[5.1386, 7.8606]$$

b) The length $l = R - L$ of the obtained confidence interval is 2.722. However, we can decrease that by making more observations:

$$l = R - L = 2t_{6,0.025}\frac{s}{\sqrt{n}} = 2 * 2.447\frac{1.472}{\sqrt{n}} = \frac{7.204}{\sqrt{n}},$$

so for that quantity to be less than one (assuming the sample standard deviation stays the same) we need n to be greater than $(7.204)^2 = 51.9$, so $n = 52$ should be enough.
Note: we are overestimating a bit, since we are not taking into account how $t_{n-1,0.025}$ decreases when we increase n.

8. (5p) Tim started working as a ticket controller for a bus company in the beginning of November. Before this job, he had a feeling that around 20% of people under 21 years old usually did not have tickets. During first month of his job, out of $n = 132$ such young people he met, 17 were without tickets. Is there a statistically significant reason to update his old beliefs? Test the hypothesis $H_0 : p = 0.2$ against the appropriate alternative. Decide if you need a two- or one-sided test. What's the p-value for the test? Can you reject H_0 on a significance level $\alpha = 0.01$?

Solution:
We are going to conduct a significance test of a hypothesis $H_0 : p = 0.2$ against the alternative $H_1 : p < 0.2$ to see if the data collected by Tim is significant enough to update his old belief (20% of young people don't buy tickets) to a new one (less than 20% of young people don't buy tickets). The test statistic we are going to use is $Z = \frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$, here $p_0 = 0.2$ is the default value of p and $\hat{p}$ is the point estimate for the proportion, obtained from data. Due to the Central Limit Theorem, under H_0, Z has a distribution very close to Normal$(0, 1)$. Find the value of Z. From the data, $\hat{p} = 17/132 \approx 0.129$, so

$$Z = \frac{0.129 - 0.2}{\sqrt{0.2 * 0.8/132}} \approx -2.04$$

So, the p-value for the test, obtained from the tables for cdf of a Normal(0,1) random variable, is

$$p\text{-value} = 0.0207$$

The obtained p-value is small, but the result is not significant on the level $\alpha = 0.01$, hence we cannot reject H_0 on that significance level.

Note, it is also possible to test H_0 against the alternative $H_1 : p \neq 0.2$, in that case the p-value would double, to become 0.0414.

9. (3p) Barbara is playing roulette in a casino, always putting in the minimal bet of 2 cents. Let us denote by X_i the change in her capital after i-th game. One can assume all of the games to be independent of each other, played by the same rules (thus X_i's are distributed identically to some random variable X). The expected gain after one game is negative: $\mu_X = -0.06$ cents, the standard deviation is $\sigma_X = 2$ cents. Barbara has 2 dollars, her plan is to play $n = 100$ games and then quit, no matter what.

   a) Find the expected value of Barbara's total gain $S_{100} = X_1 + X_2 + \ldots + X_{100}$.
   b) Find an approximate probability for her total gain S_{100} to be positive. (Hint: use the Central Limit Theorem)

   Solution:

   a) The expectation of the total gain:

   $$\mathbf{E}S_{100} = \mathbf{E}[X_1 + X_2 + \ldots + X_{100}] = \mathbf{E}[X_1] + \mathbf{E}[X_2] + \ldots + \mathbf{E}[X_{100}]$$

   $$= -0.06 - 0.06 - \ldots - 0.06 = -100 * 0.06 = -6$$

   b) By CLT, the distribution of $Z = (S_{100} - 100\mu_X)/\sqrt{100}\sigma_X$ can be approximated with Normal$(0, 1)$, so

   $$\mathbf{P}(S_{100} > 0) = \mathbf{P}(S_{100} - 100\mu_X/10\sigma_X > -100\mu_X/10\sigma_X) = \mathbf{P}(Z > -10\mu_X/\sigma_X) = \mathbf{P}(Z > 0.3)$$

   The latter can be found from the tables for Normal distribution to be 0.382