Solutions for the exam Matematisk statistik och diskret matematik (MVE050/MVE051/MSG810). Statistik ör fysiker (MSG820). 24 April 2014.

1. (3p) Anna and Beth are going to the party. Anna chooses the color of a dress to wear by tossing a fair coin. “Heads” mean Anna wears black, “tails” mean Anna wears white. Beth chooses the color by tossing a different fair coin. “Heads” mean Beth wears the same color as Anna, “tails” mean Beth wears the opposite color. Introduce the two events $A = \{\text{Anna wears black}\}$ and $B = \{\text{Beth wears black}\}$.

   a) Find the probabilities $\mathbf{P}(A), \mathbf{P}(B), \mathbf{P}(B|A)$. Are the events A and B independent?
   b) Now assume that Anna's coin was fair, but Beth's coin was biased, so that the probability of getting “heads” was $p = 0.6$ for Beth. Find the probabilities $\mathbf{P}(A), \mathbf{P}(B), \mathbf{P}(B|A)$. Are the events A and B independent?

   Solution:

   a) Let $C = \{\text{Anna's coin shows “heads”}\}$, $D = \{\text{Beth's coin shows “heads”}\}$. Note that $A = C$, so
   $$\mathbf{P}(A) = \mathbf{P}(C) = 1/2$$
   since Anna's coin is fair. Then, $B = (C \cap D) \cup (C^c \cap D^c)$, so, since $C \cap D$ and $C^c \cap D^c$ are mutually exclusive,
   $$\mathbf{P}(B) = \mathbf{P}(C \cap D) + \mathbf{P}(C^c \cap D^c) = \mathbf{P}(C)\mathbf{P}(D) + \mathbf{P}(C^c)\mathbf{P}(D^c) = 0.5 \cdot 0.5 + 0.5 \cdot 0.5 = 0.5$$
   Next, $\mathbf{P}(B|A) = \mathbf{P}(D|C) = \mathbf{P}(D)$, since C and D are independent, so
   $$\mathbf{P}(B|A) = 0.5 = \mathbf{P}(B),$$
   and thus A, B are independent.
   b) In the same notation, with the same reasoning,
   $$\mathbf{P}(A) = \mathbf{P}(C) = 1/2$$
   Next, $\mathbf{P}(D) = 0.6$, so
   $$\mathbf{P}(B) = \mathbf{P}(C \cap D) + \mathbf{P}(C^c \cap D^c) = \mathbf{P}(C)\mathbf{P}(D) + \mathbf{P}(C^c)\mathbf{P}(D^c) = 0.5 \cdot 0.6 + 0.5 \cdot 0.4 = 0.5$$
   , and lastly,
   $$\mathbf{P}(B|A) = \mathbf{P}(D|C) = \mathbf{P}(D) = 0.6 \neq 0.5 = \mathbf{P}(B),$$
   so A, B are not independent.

2. (3p) A waiting time X for the bus 16 is distributed uniformly on the interval $(0, 10)$, with a pdf given by
   $$f_X(x) = \begin{cases} 0.1, & 0 < x < 10, \\ 0, & \text{otherwise.} \end{cases}$$

   a) Find $\mathbf{P}(X = 1.5)$.
   b) Find $\mathbf{P}(4 < X \leq 8)$.

c) Find $\mathbf{E}[(X+2)/10]$.

Solution:

a) X is a continuous random variable, so $\mathbf{P}(X = 1.5) = 0$.
b)
$$\mathbf{P}(4 < X \leq 8) = \int_4^8 f_X(x)\,dx = \int_4^8 0.1\,dx = 0.1 \cdot (8-4) = 0.4$$
c)
$$\mathbf{E}[(X+2)/10] = 0.1(\mathbf{E}[X]+2) = 0.1(5+2) = 0.7$$

3. (4p) Rating is a proportion of people supporting the mayor. A newspaper claims that the mayor's rating increased over the last three months by 2% to become 53%. A study involved a survey among $n = 200$ people, with 106 of them supporting the mayor.

a) Build the null and the alternative hypothesis. Find the p-value of the study findings.
b) How many people does the study have to involve for the p-value of the same findings to be less than 0.05?

Solution:

a) Let $p_0 = 0.51$ denote the old rating value, and p denote the new one. The appropriate hypotheses would be

$$H_0 : p = p_0$$
$$H_1 : p > p_0$$

The p-value is found as a probability for the test statistic $\hat{p}$ to be larger than an observed value under the null hypothesis:

$$p\text{-val} = \mathbf{P}(\hat{p} > 0.53 | H_0) = \mathbf{P}(\frac{\hat{p} - p_0)}{\sqrt{p_0(1-p_0)/n} > \frac{0.53-p_0)}{\sqrt{p_0(1-p_0)/n}}} | H_0) = \mathbf{P}(Z > \frac{0.53 - p_0)}{\sqrt{p_0(1-p_0)/200}}),$$

where $Z \sim \mathrm{N}(0,1)$. The latter continues with

$$\mathbf{P}(Z > 0.56) = 0.2877$$

so the p-value is 0.29.
b) For the p-value to be less than 0.05 the following equation must hold:

$$\frac{0.53 - p_0)}{\sqrt{p_0(1-p_0)/n}} > \zeta_{0.95}$$

Solving it for n we obtain

$$n > (\zeta_{0.95})^2 \frac{p_0(1-p_0)}{(0.53-p_0)^2} = (1.645)^2 \frac{0.51 \cdot 0.49}{(0.02)^2} = 1690.5$$

so having more than 1690 respondents is necessary to have p-value smaller than 0.05 for these findings.

4. (4p) In the game of Blackjack, the aim is to get the score as close to 21 as possible, but not higher. Aces are worth 11 points, Jacks, Queens and Kings are worth 10, and the number cards are worth the same amount of points as the card's value. If a player has exactly 21 point in his hand at some point, he wins immediately. If a player has more than 21 points in his hand at some point, he loses immediately. Assume that a game is played with one standard 52-card deck.

   a) First, player is dealt two cards. What is the probability for those two cards to worth exactly 21 point in total?
   b) What is the probability for the two cards to be worth more than 21 point?
   c) Now assume that instead of using a standard 52-cards deck, the cards were taken from ten 52-cards decks shuffled together. What is the probability for the first two cards to be worth more than 21 points together now?

   Solution:

   a) To get 21 points from two cards, one of them has to be an Ace, and the other has to be one of the cards that cost 10 points: a Ten, a Jack, a Queen, or a King. There is four Aces in the deck, and $4 \cdot 4 = 16$ cards worth 10 points. Hence there is $4 \cdot 16 = 64$ combinations giving 21 points, out of a total of $\binom{52}{2} = 52 \cdot 51/2 = 1326$ possible combinations of two cards, yielding the probability $64/1326 = 0.048$ of scoring 21 from two cards.
   b) The only way to get more than 21 point is to get two Aces. The number of combinations of two Aces is $\binom{4}{2} = 4 \cdot 3/2 = 6$, so the probability is $6/1326 = 0.0045$
   c) Again, the only way to get more than 21 point is by having two Aces. The total number of Aces in ten decks shuffled together is 40, so there is $\binom{40}{2} = 40 \cdot 39/2 = 780$ combinations out of $\binom{520}{2} = 519 \cdot 520/2 = 134940$ possible combinations of two cards. The probability to get more than 21 point with two cards then becomes $780/134940 = 0.0058$.

5. (4p) In a lottery, 2 million tickets are sold, each costs 10 SEK. Afterwards, 100000 tickets win 15 SEK, 2000 tickets win 500 SEK, 1000 tickets win 1000 SEK, 100 tickets win 20000 SEK, 10 win 100000 SEK, and the jackpot awarded to one person is two million SEK. Let X be the total gain of one lottery ticket: the winning minus the cost.

   a) What is the probability of winning at least something, if you only have one ticket, i.e. what is $\mathbf{P}(X > 0)$?
   b) What is the expected yield from one lottery ticket, i.e. what is $\mathbf{E}[X]$? What would the expected yield be if there was no jackpot (and the rest of the numbers were the same)?
   c) Find the variance of the number of winning tickets among $n = 100$ tickets (you can think of that number as Binomially distributed).

   Solution:

   a) The amount of winning tickets is $100000 + 2000 + 1000 + 100 + 10 + 1 = 103111$, out of the total amount of 2000000 tickets, yielding the probability of a ticket to be winning $103111/2000000 \approx 0.052$.

b) First, note that we can represent X as $X = Y - 10$, where Y is a random variable representing the winnings of one ticket, and 10 is the ticket cost. Let us use the definition of expectation of a random variable to find $\mathbf{E}[Y]$:

$$\mathbf{E}[Y] = \sum_{\text{all possible } x} x\mathbf{P}(Y = x) = 0.05 * 15 + 0.001 * 500 + 0.0005 * 1000 + 0.00005 * 20000$$
$$+ 0.000005 * 100000 + 0.0000005 * 2000000 = 4.25$$

hence $\mathbf{E}[X] = 4.25 - 10 = -5.75$. Without the jackpot,

$$\mathbf{E}[Y] = \sum_{\text{all possible } x} x\mathbf{P}(Y = x) = 0.05 * 15 + 0.001 * 500 + 0.0005 * 1000 + 0.00005 * 20000$$
$$+ 0.000005 * 100000 = 3.25,$$

and $\mathbf{E}[X] = 3.25 - 10 = -6.75$.

c) The variance of a $\text{Bin}(n = 100, p = 0.103)$ random variable is given by $np(1-p) = 100 * 0.052 * 0.947 = 4.92$.

6. (4p) Find a generating function $A(x)$ of the sequence $\{a_n\}_{n\geq 0}$ defined by the following recursion:

$$\begin{cases} a_0 = 0, \\ a_1 = 1, \\ a_n = 2a_{n-1} + 3a_{n-2} - 2, \quad n = 2, 3, \ldots \end{cases}$$

Solution:
Write out the recursion for $n = 2, 3, \ldots$ and multiply with the corresponding power of x:

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

Sum up:

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

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

$$A(x) - a_1x - a_0 = 2x(A(x) - a_0) + 3x^2A(x) - 2x^2\frac{1}{1-x}$$

Plug in the values for a_0, a_1:

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

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

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

so

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

7. (3p) A car driver wants to build a confidence interval for his car's average weekly fuel consumption. He gathers the following data:

   52.5 54.0 47.5 51.0 57.0 52.0 51.0 55.0 51.0 49.5 51.5 [liters/week]

   Build a 99% confidence interval for the mean weekly fuel consumption. What kind of statistical assumptions do you need to make?

   Solution:
   Assume his data is an i.i.d. sample from a Normal distribution with mean μ and variance σ^2. We can then build a 99% confidence interval (L, R) for μ with the unknown σ^2 using the formulas

   $$L = \bar{X} - t_{0.99,n-1}\frac{s}{\sqrt{n}}, \quad R = \bar{X} + t_{0.99,n-1}\frac{s}{\sqrt{n}}$$

   We find $\bar{X} = 52.0$, $s^2 = 6.8$, $s = 2.6$ $t_{0.995,10} = 3.169$, $\sqrt{11} = 3.31$, so the confidence interval is given by:

   $$L = 52.0 - 3.169\frac{2.6}{3.31} = 49.5, \quad R = 52.0 + 3.169\frac{2.6}{3.31} = 54.5$$

8. (5p) Erik is going to a different city for the first time in his life. Each day, from 9 AM to 9 PM his mom calls him according to a Poisson process of intensity 0.5 calls/hour. If he is not answering to more than one call in a row, mom is getting worried. Normally Erik always picks up, but now he wants to go to the movies, where he will turn the phone off. There is two options: watch the whole movie of 2 hours length, or watch the two short ones, one hour each, with a 40 minutes break in between during which he will turn the phone on (but won't see any missed calls).

   a) Is there a difference in terms of chances of worrying mom? In which scenario is Erik more likely to miss two or more calls in a row, and why?
   b) What is the probability of missing two or more calls in a row if Erik decides to go for a two hours movie?
   c) What is the probability of missing two or more calls in a row if Erik instead goes for two shorter movies? Assume that Erik can answer the phone normally during the break.

   Solution:

   a) Even though the total length of a time interval is the same in both cases, in the second scenario there is a break in between, during which Erik can possibly receive a call which is going to break the streak, so it seems like the probability should be smaller in the second case.

b) Denote by X the amount of calls Erik's mom makes during the movie. By definition of Poisson process, we know that $X \sim \text{Pois}(\lambda t)$, where $\lambda = 0.5$, $t = 1$, so $X \sim \text{Pois}(1)$. The probability of Erik missing two or more calls is then given by:

$$\mathbf{P}(X \geq 2) = 1 - \mathbf{P}(X \leq 1) = 1 - \mathbf{P}(X = 0) - \mathbf{P}(X = 1) = 1 - e^{-1} - e^{-1} \approx 0.264$$

c) Denote by X_1, X_2 the amount of calls Erik's mom makes during the first and the second movie, and by Y the amount of calls during the break. By definition of Poisson process, we know that X_1, X_2 and Y are independent. Moreover, $X_1 \sim \text{Pois}(0.5 \cdot 1)$, $X_2 \sim \text{Pois}(0.5 \cdot 1)$ and $Y \sim \text{Pois}(0.5 \cdot 0.66)$. Notice that Erik can only miss two calls in a row if and only if either he receives two calls in a row during one of the movies, or he receives exactly one call during each movie, and no calls in between — and those two events are mutually exclusive. We can write:

$$\mathbf{P}(\text{miss more than 1 call in a row})$$

$$= \mathbf{P}(X_1 \geq 2 \text{ or } X_2 \geq 2) + \mathbf{P}(X_1 = 1, X_2 = 1, Y = 0).$$

Now, because X_1, X_2, Y are independent, Poisson-distributed with respective parameters, we can find

$$\mathbf{P}(X_1 \geq 2 \text{ or } X_2 \geq 2) = 1 - \mathbf{P}(X_1 \leq 1 \text{ and } X_2 \leq 1)$$

$$= 1 - \mathbf{P}(X_1 \leq 1)\mathbf{P}(X_2 \leq 1) = 1 - (e^{-0.5} + 0.5e^{-0.5})^2,$$

and

$$\mathbf{P}(X_1 = 1, X_2 = 1, Y = 0) = \mathbf{P}(X_1 = 1)\mathbf{P}(X_2 = 1)\mathbf{P}(Y = 0) = 0.5e^{-0.5} \cdot 0.5e^{-0.5} \cdot e^{-0.33},$$

hence

$$\mathbf{P}(\text{miss more than 1 call in a row})$$

$$= 1 - (e^{-0.5} + 0.5e^{-0.5})^2 + 0.5e^{-0.5} \cdot 0.5e^{-0.5} \cdot e^{-0.33} \approx 0.238,$$

which is indeed slightly lower than the probability of missing two calls in the first scenario.