Tentamentsskrivning i **Mathematisk statistik TMA321, 3p.**

Tid: Tisdagen den 16 augusti, 2005 kl 08.30-12.30

Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736 907 613, rum MC 1421.

Hjälpmedel: valfri rknare, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.

There are five questions with the total number of marks 30. Attempt as many questions, or parts of the questions, as you can. Preliminary grading system:

grade "3" for 12 to 17 marks,

grade "4" for 18 to 23 marks,

grade "5" for 24 and more marks.

1. (6 marks) The sample space for the dice experiment is a set of 36 equally likely outcomes

	1	2	3	4	5	6
1	•	•	•	•	•	•
2	•	•	•	•	•	•
3	•	•	•	•	•	•
4	•	•	•	•	•	•
5	•	•	•	•	•	•
6	•	•	•	•	•	•

Denote by X_1 the number shown by the first die and by X_2 the number shown by the second die. Considered as two functions defined on the sample space the random variables X_1 and X_2 take the following values

1	1	1	1	1	1
2	2	2*	2	2	2
3	3	3	3	3	3
4	4	4	4	4	4
5	5	5	5	5	5
6	6	6	6	6	6

and

1	2	3	4	5	6
1	2	3*	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6
1	2	3	4	5	6

These tables can be used for computing the joint distribution of X_1 and X_2. For example, the joint probability $P(X_1 = 2, X_2 = 3) = 1/36$ since there is only one outcome (marked by a star) when both $X_1 = 2$ and $X_2 = 3$. It follows that X_1 and X_2 are independent random variables because for any pair (k, l) the joint probability $P(X_1 = k, X_2 = l) = P(X_1 = k)P(X_2 = l)$ is the product of the marginal probabilities.

a. Give similar tables for another pair of random variables: $Y = X_1 + X_2$ and $Z = X_1 - X_2$.

b. Draw two barplots depicting the marginal distributions of Y and Z. Find their means and variances.

c. Find the joint distribution of Y and Z. Are these random variables independent?

d. Find the the covariance between Y and Z. What does it say about the relationship between Y and Z?

2. (6 marks) Poker Hand Values. The cards used in poker have two qualities. Suit (obviously spades, hearts, clubs and diamonds) and Rank (two through Ace). With this in mind, here is how you create poker hands:

- Straight Flush - all cards of same suit. Rank in sequence as in a regular straight (see below).

- Four of a Kind - four cards with the same rank.

- Full House - three cards of one rank, and two cards of a second rank.

- Flush - all cards of same suit.

- Straight - all cards with ranks in sequence (ex. 4-5-6-7-8). You can have an ace either high (A-K-Q-J-10) or low (5-4-3-2-1). However, a straight may not 'wraparound'. (Such as K-A-2-3-4, which is not a straight).

and so on. Suppose you have picked five cards from a standard deck of 52 cards at random. Then your probabilities for the mentioned poker hands are

$$P(Straight.flush) = \frac{4 \times 10}{2598960} = 0.000015$$

$$P(Four.of.a.Kind) = \frac{13 \times 48}{\binom{52}{5}} = 0.00024$$

$$P(Full.House) = \frac{13 \times 4 \times 12 \times 6}{\binom{52}{5}} = 0.0014$$

$$P(Flush) = \frac{4 \times \binom{13}{5}}{\binom{52}{5}} = 0.0020$$

$$P(Straight) = \frac{10 \times 4^5}{2598960} = 0.0039$$

a. Give a detailed explanation of these probability computations.

b. Compute the conditional probability $P(Flush|Straight)$.

c. Draw a rough Venn diagram depicting three events: Flush, Straight and Straight Flush. How do you compute the probability that at least one of the three events occurs.

d. Now suppose you and your friend are dealt five cards each from the same standard deck. Let your hand be four kings and 4-spades. What is the (conditional) probability that your friend has four aces?

3. (6 marks) Ozon levels around Los Angeles have been measured as high as 220 parts per billion (ppb). Concentrations this high can cause the eyes to burn and are a hazard to both plant and animal life. These data were obtained on the ozone level in a forested area near Seattle
    160 176 160 180 167 164
    165 163 162 168 173 179
    170 196 185 163 162 163
    172 162 167 161 169 178 161.
The sum of the observed values is 4226, and the sum of the squared values is 716320.

a. Set eight cells 160-164, 165-169, 170-174, 175-179, 180-184, 185-189, 190-194, 195-199 and draw a histogram. Does it look like an exponential distribution?

b. Assuming an exponential distribution Exp(λ) estimate the parameter λ using the method of moments.

c. Give a justification for the method of moments estimation by referring to the Law of Large Numbers.

d. Apply the result of b) to find a theoretical proportion of observations exceeding 220 ppb. Does the result agree with the data?

4. (6 marks) Galileo's experiment with a ball and an inclined ramp. The balls's trajectory is made horizontal before it falls over the edge, as shown in the picture
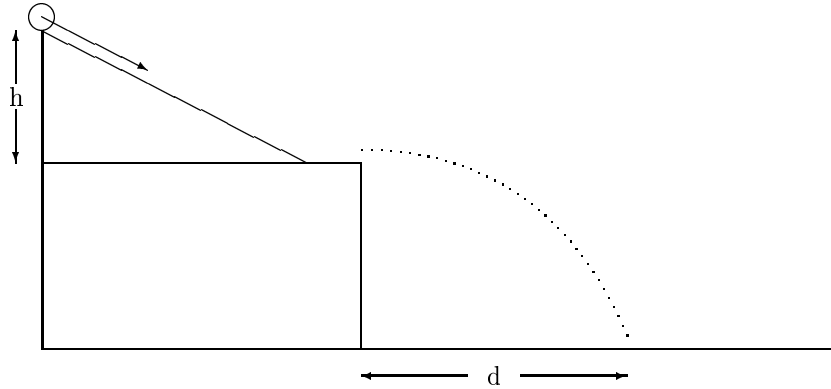
The horizontal distance d from the edge to the point of impact is measured for different values of the initial height h of the ball. Five data points obtained by Galileo in 1608 are shown in the table, where the units are punti (points), one punto is slightly less than 1 mm

h	d
1000	1500
828	1340
800	1328
600	1172
300	800

a. Draw a scatter plot to see if a simple linear regression model is appropriate.

b. Galileo regarded the motion of the ball as the superposition of horizontal and vertical components. At the lower edge of the ramp the verical speed is zero, while the horizontal speed v is proportional to the square root of h due to the energy conservation law. Then $d = vt$ since the horizontal speed is constant during the time t of the free fall.

This argument leads to a more realistic model $d = \beta_0 + \beta_1\sqrt{h} + \epsilon$. Here the noise component ϵ summarizes all the secondary factors (with the primary factor being the starting height h) which influence the distance d. Name some of such secondary factors. How can we theoretically justify the usual assumption of normality for the noise distribution $\epsilon \sim N(0, \sigma^2)$? Is it reasonable to think that the five observations had independent ϵ_i and the same σ^2?

c. Estimate the parameters of the model in b) using the least squares method. Clearly show your calculations.

d. Test the null hypothesis $H_0 : \beta_0 = 0$, which follows from the argument in b).

5. (6 marks) In an experiment on radioactivity, Rutherford and Geiger counted the number of alpha decays occuring in $n = 2608$ time intervals of 7.5 seconds. Assuming that the source consists of a large number of radioactive atoms and that the probability for any one of them to emit an alpha particle in a short interval is small, one would expect the number of decays X to follow a Poisson distribution. Deviations from this hypothesis would indicate that the decays were not independent. One could imagine, for example, that the emission of an alpha particle might cause neighboring atoms to decay, resulting in a clustering of decays.

x	0	1	2	3	4	5	6	7
n_x	57	203	383	525	532	408	273	139
$x \cdot n_x$	0	203	766	1575	2128	2040	1638	973
$x^2 \cdot n_x$	0	203	1532	4725	8512	10200	9828	6811

x	8	9	10	11	12	13	14	> 14	Total
n_x	45	27	10	4	0	1	1	0	2608
$x \cdot n_x$	360	243	100	44	0	13	14	0	10097
$x^2 \cdot n_x$	2880	2187	1000	484	0	169	196	0	48727

a. Find the sample mean $\bar{X}$ and variance s^2 for the data in the table.

b. Compute the index of dispersion $t = s^2/\bar{X}$. Why would we expect to find t around 1?

c. It can be shown that for large n the distribution of t is approximately normal with mean 1 and variance $2/(n-1)$ given that X has a Poisson distribution. What is the P-value for the null hypothesis that X follows a Poisson distribution? Would you accept or reject the Poisson model for X?

d. Give a point estimate for the alpha decay intensity λ measured as the number of decays per second. What is the standard error of this estimate? Find a 95% confidence interval for λ.

Statistical tables supplied:
  1. Normal distribution. 2. t-distribution.
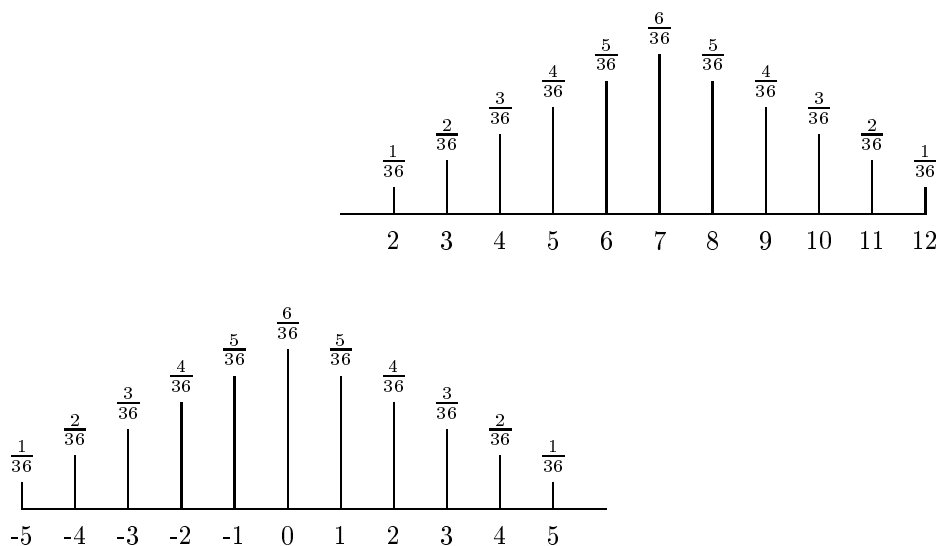
<div align="center">

Good luck!

</div>

ANSWERS

1a. The variables Y and Z take the following values on the sample space:

2	3	4	5	6	7
3	4	5	6	7	8
4	5	6	7	8	9
5	6	7	8	9	10
6	7	8	9	10	11
7	8	9	10	11	12

and

0	-1	-2	-3	-4	-5
1	0	-1	-2	-3	-4
2	1	0	-1	-2	-3
3	2	1	0	-1	-2
4	3	2	1	0	-1
5	4	3	2	1	0

1b. The corresponding probability mass functions

Their means and variances can be computed directly. Otherwise one can first use the discrete uniform distribution $U(6)$ formulas: $E(X_1) = E(X_2) = (1+7)/2 = 3.5$ and $Var(X_1) = Var(X_2) = 35/12 = 2.92$. And then apply properties of the mean and variance $E(Y) = E(X_1) + E(X_2) = 7$, $E(Z) = E(X_1) - E(X_2) = 0$, and because X_1 and X_2 are independent $Var(Y) = Var(Z) = Var(X_1) + Var(X_2) = 5.83$.

1c. The next joint distribution table shows that Y and Z are dependent since we can find at least one pair of values (i, j) with $P(Y = i; Z = j) \neq P(Y = i)P(Z = j)$. In fact the last relation holds for any pair of values (i, j).

|	2	3	4	5	6	7	8	9	10	11	12	Total
-5	0	0	0	0	0	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{36}$
-4	0	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0	$\frac{2}{36}$
-3	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	$\frac{3}{36}$
-2	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	$\frac{4}{36}$
-1	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{5}{36}$
0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	$\frac{6}{36}$
1	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{5}{36}$
2	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	$\frac{4}{36}$
3	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	$\frac{3}{36}$
4	0	0	0	0	$\frac{1}{36}$	0	$\frac{1}{36}$	0	0	0	0	$\frac{2}{36}$
5	0	0	0	0	0	$\frac{1}{36}$	0	0	0	0	0	$\frac{1}{36}$
	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$	1

1d. Since $YZ = X_1^2 - X_2^2$ we have

$$Cov(Y, Z) = E(YZ) - E(Y)E(Z) = E(X_1^2) - E(X_2^2) = 0$$

implying that the correlation coefficient is also equal to zero. This is another example of two random variables that are uncorrelated despite being dependent. The correlation coefficient is a numeric measure of the strength of <u>linear</u> relationship between two random variables.

2a. We can apply the division rule of probability because all $\binom{52}{5} = 2598960$ outcomes of the random experiment are equally likely. We illustrate by counting the number of favorable outcomes for the Full House by applying the multiplication principle of combinatorics. Consider a four step procedure: 1. choose a rank, 2. choose three cards out of four cards in this rank, 3. choose a second rank, 4. choose two out of of four cards in the second rank. It remains to multiply the number of outcomes in each step $13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2}$ and divide it by 2598960 to obtain $P(Full.House) = 0.0014$.

2b. $P(Flush|Straight) = P(Straight.flush)/P(Straight) = 0.000015/0.0039 = 0.0038$.

2c. $P(Flush \cup Straight \cup Straight.flush) = P(Flush) + P(Straight) - P(Straight.flush) = 0.0059$.
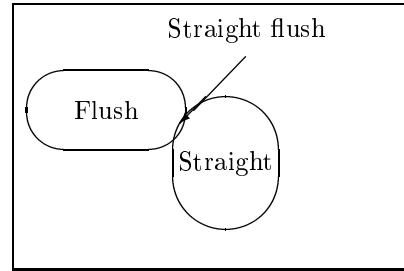
2d. Since the second hand is taken from a deck of 47 cards and there are 43 candidates for the fifth card, we have

$P(Friend's\ hand = four\ aces + something | My\ hand = four\ kings + 4\spadesuit) = \frac{43}{\binom{47}{5}} = 0.000028$.

3a. The histogram does give an impression of an exponential distribution.

3b. If $X \sim \text{Exp}(\lambda)$, then the first moment is $E(X) = 1/\lambda$. Replacing the population mean $E(X)$ with the corresponding sample mean $\bar{X} = 4226/25 = 169$ we find the method of moment estimate $\tilde{\lambda} = 1/\bar{X} = 0.006$.

3c. The sample mean $\bar{X} = (X_1 + \ldots + X_n)/n$ is the average of independent and identically distributed random variables. According to the law of large numbers the empirical first moment $\bar{X}$ converges to the theoretical first moment $\mu = E(X)$ as the sample size tends to infinity. Similarly, the empirical second moment $\bar{X}^2 = (X_1^2 + \ldots + X_n^2)$ may be used as a replacement for the theoretical second moment $E(X^2)$, when there are two parameters to be estimated.

The law of large numbers ensures that a method of moments estimate converges to the true parameter value as the sample size increases (consistency property of a point estimate).

3d. If $X \sim \text{Exp}(\lambda)$, then $P(X > x) = e^{-\lambda x}$. Putting $x = 220$ and $\lambda = 0.006$ yields $P(X > 220) = 0.27$. This implies that every fourth observation should exceed 220, which is not the case.

4a. The relation is close to linear although the slope tends to decrease.

4b. Some of the secondary factors that influence the outcome of the experiment:

1. measurement errors for both h and d,

2. initial speed of the ball,

3. friction,
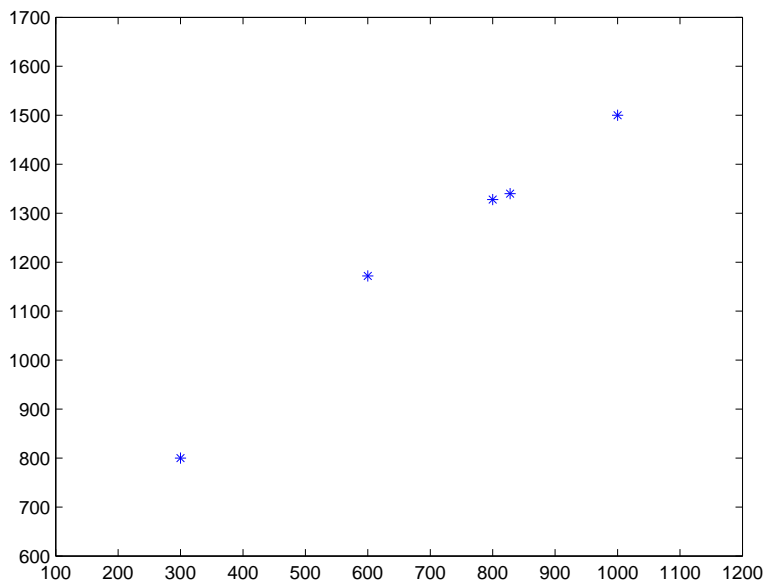
Figure 1: *The scatter plot*

4. the initial position of the ball's center of mass,

5. air motion, and air resistance,

6. the initial angle of the free fall.

The systematic components of these factors, like average friction and air resistance, will be reflected in the intercept parameter β_0. The noise summarizes all the random factors which can be thought to be mutually independent and each of them having relatively small contribution. Then the normality assumption with zero mean can be justified by the Central limit theorem.

The noise values for the five experiments can indeed be viewed as independent with the same variance, since the noise factors have no "memory" from earlier trials and the experiment setting is unchanged.

4c. Explanatory variable $x = \sqrt{h}$ and dependent variable $y = d$. Sample means $\bar{x} = 26.1$, $\bar{y} = 1228$. Sample standard deviations $s_x = 5.525$, $s_y = 265.9$, sample covariance $c_{xy} = \frac{n}{(n-1)}\left(\overline{xy} - \bar{x}\bar{y}\right) = \frac{5}{4}\left(33224 - 32050\right) = 1467.5$, sample correlation $r = \frac{c_{xy}}{s_x s_y} = 0.999$.

Least square estimates: the slope $b_1 = r \cdot \frac{s_y}{s_x} = 48.1$, the intercept $b_0 = \bar{y} - b_1\bar{x} = -26.9$. The fitted regression line $y = 48.1x - 26.9$. Estimated noise

variance $s^2 = \frac{n-1}{n-2} s_y^2 (1 - r^2) = 202.5$, $s = 14.2$.

4d. The standard error of b_0 is $s_{b_0} = \frac{s}{s_x \sqrt{n-1}} \sqrt{\frac{1}{n} \sum x_i^2} = 34.2$. The test statistic $T = b_0/s_{b_0} = -0.787$ is not significant according to the t-distribution table with df=3. The two-sided P-value of the test is larger than 40% since 0.787 is smaller than 0.978. Can not reject the null hypothesis.

5a. Sample mean $\bar{X} = 10097/2608 = 3.87$ and variance $s^2 = \frac{n}{n-1}(48727/2608 - 3.87^2) = 3.69$ so that $s = 1.92$.

5b. The index of dispersion $t = 0.9543$ is expected to be close to 1 since the Poisson distribution has mean and variance equal to each other.

5c. The ratio $\frac{t-1}{\sqrt{2/(n-1)}} = -1.65$ has the standard normal distribution under the null hypothesis. The two-sided P-value of the null hypothesis is $P(|Z| \geq 1.65) = 2(1 - 0.9505) = 0.099$. There are about 10% chances to observe this kind of deviation given the null hypothesis is true. Can not reject the Poisson model.

5d. Point estimate of the decay intensity: $\hat{\lambda} = \bar{X}/7.5 = 0.52$. Its standard error is $s_{\hat{\lambda}} = s_{\bar{X}}/7.5 = \frac{s/\sqrt{2608}}{7.5} = 0.005$. Since the sample mean is approximately normally distributed, a 95% confidence interval for λ can be computed as $\hat{\lambda} \pm 1.96 s_{\hat{\lambda}} = 0.52 \pm 0.01$.