Tentamentsskrivning i Statistisk slutledning MVE155/MSG200, 7.5 hp.

Tid: tisdagen den 12 mars, 2013 kl 14.00-18.00
Examinator och jour: Serik Sagitov, tel. 772-5351, mob. 0736 907 613, rum H3026 i MV-huset.
Hjälpmedel: Chalmersgodkänd räknare, **egen** formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.
CTH: för "3" fordras 12 poäng, för "4" - 18 poäng, för "5" - 24 poäng.
GU: för "G" fordras 12 poäng, för "VG" - 20 poäng.

Important! For each problem do your best to
- describe and justify statistical models you apply,
- state clearly hypotheses you test,
- discuss different relevant approaches you have learned in the course.

---

1. (2 points) From Wikipedia: "The American Psychological Association's 1995 report Intelligence: Knowns and Unknowns stated that the correlation between IQ and crime was -0.2. It was -0.19 between IQ scores and number of juvenile offenses in a large Danish sample; with social class controlled, the correlation dropped to -0.17. A correlation of 0.20 means that the explained variance is less than 4%."

Explain the last sentence.

2. (5 points) The Laplace distribution with a positive parameter λ is a two-sided exponential distribution. Its density function is $f(x) = \frac{\lambda}{2} e^{-\lambda|x|}$ for $x \in (-\infty, \infty)$.

a. The variance of this distribution is $2\lambda^{-2}$ and kurtosis is 6. Prove this using the formula $\int_0^\infty x^k e^{-x} dx = k!$ valid for any natural number k.

b. Take $\lambda = \sqrt{2}$. Plot carefully the density $f(x)$ together with the standard normal distribution density.

c. Use the drawn picture to explain the exact meaning of the following citation. "Kurtosis is a measure of the peakedness of the probability distribution of a real-valued random variable, although some sources are insistent that heavy tails, and not peakedness, is what is really being measured by kurtosis".

3. (5 points) The following 16 numbers came from normal random number generator on a computer:

5.33	4.25	3.15	3.70
1.61	6.39	3.12	6.59
3.53	4.74	0.11	1.60
5.49	1.72	4.15	2.28

a. Write down the likelihood function for the mean and variance of the generating normal distribution. (Hint: to avoid tedious calculations on your calculator use the numbers in the next subquestion.)

b. In what sense the sum of the sample values (which is close to 58), and the sum of their squares (which is close to 260) are sufficient statistics in this case?

c. Turning to the log-likelihood function compute the maximum likelihood estimates for the mean and variance. Is the MLE for the variance unbiased?

4. (3 points) Questions concerning hypotheses testing methodology. Try to give detailed answers.

a. Consider a hypothetical study of the effects of birth control pills. In such a case, it would be impossible to assign women to a treatment or a placebo at random. However, a non-randomized study might be conducted by carefully matching control to treatments on such factors as age and medical history.

The two groups might be followed up on for some time, with several variables being recorded for each subject such as blood pressure, psychological measures, and incidences of various problems. After termination of the study, the two groups might be compared on each of these many variables, and it might be found, say, that there was a "significant difference" in the incidence of melanoma.

What is a common problem with such "significant findings"?

b. You analyse cross-classification data summarized in a two by two contingency table. You wanted to apply the chi-square test but it showed that one of the expected counts was below 5. What alternative statistical test you may try applying?

c. Why tests like Wilcoxon, Friedman, and Kruskal-Wallis tests are often called distribution-free tests?

5. (5 points) A public policy polling group is investigating whether people living in the same household tend to make independent political choices. They select 200 homes where exactly three voters live. The residents are asked separately for their opinion ("yes" or "no") on a city charter amendment. The results of the survey are summarized in the table:

Number of saying "yes"	0	1	2	3
Frequency	30	56	73	41

Based on these data can we claim that opinions are formed independently?

6. (5 points) Suppose you have a data of size n for which the linear regression model seems to work well. The key summary statistics are represented by sample means $\bar{x}, \bar{y}$, sample standard deviations s_x, s_y, and a sample correlation coefficient r.

An important use of the linear regression model is forecasting. Assume we are interested in the response to a particular value x of the explanatory variable.

a. The exact $100(1-\alpha)\%$ confidence interval for the mean response value is given by the formula:

$$b_0 + b_1 x \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{\frac{1}{n} + \frac{1}{n-1}\left(\frac{x-\bar{x}}{s_x}\right)^2}.$$

Explain carefully the meaning and role of each of the terms.

b. Another important formula in this context

$$b_0 + b_1 x \pm t_{\alpha/2,n-2} \cdot s \cdot \sqrt{1 + \frac{1}{n} + \frac{1}{n-1}\left(\frac{x-\bar{x}}{s_x}\right)^2}$$

is called the exact $100(1-\alpha)\%$ prediction interval. Explain the difference between these two formulae. Illustrate by a simple example.

c. Comment on the predictor properties depending on the distance from the given value x to the sample mean $\bar{x}$. Illustrate using appropriate plots.

7. (5 points) In an experimental study two volunteer male subjects aged 23 and 25 underwent three treatments to compare a new drug against no drug and placebo. Each volunteer had one treatment per day and the time order of these three treatments was randomized.

a. Comment on the details of the experimental design.

b. Find the exact null distribution for the test statistic of an appropriate non-parametric test.

Statistical tables supplied:
1. Normal distribution table
2. Chi-square distribution table
3. t-distribution table
4. F-distribution table

Partial answers and solutions are also welcome. Good luck!

NUMERICAL ANSWERS

1. Coefficient of determination is the squared sample correlation $r^2 = (0.2)^2 = 0.04$.

2a. Since the mean is 0, the variance is computed as

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)dx = \lambda \int_0^{\infty} x^2 e^{-\lambda x} dx = 2\lambda^{-2}.$$

The kurtosis is the scaled fourth moment

$$\beta_2 = \sigma^{-4} \int_{-\infty}^{\infty} x^4 f(x)dx = \frac{\lambda^5}{4} \int_0^{\infty} x^2 e^{-\lambda x} dx = 6.$$

2b. The Laplace curve is symmetric. Its shape is formed by two exponentially declining curves: one for positive x and the other for the negative x.

2c. For $\lambda = \sqrt{2}$ the mean is 0, the skewness is 0, and the kurtosis is 6. Compared to the normal curve with the same mean but smaller kurtosis (=3), the Laplace distribution has heavier tails. Moreover, since the variances are equal, the two curves should cross 4 times. This implies that the Laplace curve must also have higher peakedness.

3a. Given $\sum_{i=1}^n x_i = 58$ and $\sum_{i=1}^n x_i^2 = 260$, the likelihood function is

$$L(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{2\sigma^2}} = \frac{1}{(2\pi)^8\sigma^{16}} e^{-\frac{260-116\mu+16\mu^2}{2\sigma^2}}.$$

3b. It is sufficient to know $\sum_{i=1}^n x_i$ and $\sum_{i=1}^n x_i^2$ to compute the likelihood function.

3c. The MLE for the mean is $\bar{x} = 3.63$ and the MLE for the variance $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^n x_i^2 - \bar{x}^2 = 3.11$. These are computed by taking the derivative of the log-likelihood

$$l(\mu, \sigma^2) := \ln L(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - n\ln\sigma - \frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{2\sigma^2}$$

and solving a pair of equations

$$\begin{aligned} \frac{-2\sum_{i=1}^n x_i + 2n\mu}{2\sigma^2} &= 0, \\ -\frac{n}{\sigma} + \frac{\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2}{\sigma^3} &= 0. \end{aligned}$$

Since

$$E(n^{-1}\sum_{i=1}^n X_i^2 - \bar{X}^2) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2,$$

$\hat{\sigma}^2$ is a biased estimate of σ^2.

4a. Multiple testing.

4b. Exact Fisher's test.

4c. Nonparametric tests do not assume a particular form of the population distribution like normal distribution.

5. The null hypothesis is that everybody votes independently. Let p be the population proportion for 'yes'. Then the number of 'yes' for three voters in a household has the binomial distribution

model $X \sim \text{Bin}(3, p)$ with an unspecified parameter p. So the null hypothesis can be expressed in the following form

$$H_0 : p_0 = (1-p)^3,\ p_1 = 3p(1-p)^2,\ p_2 = 3p^2(1-p),\ p_3 = p^3.$$

The MLE of p is the sample mean $\hat{p} = 0.5417$. We use the Pearson chi-square test with expected counts

$$E_0 = n(1-\hat{p})^3 = 19,\ E_1 = 3n\hat{p}(1-\hat{p})^2 = 68,\ E_2 = 3n\hat{p}^2(1-\hat{p}) = 81,\ E_3 = 3n\hat{p}^3 = 32.$$

The observed chi-square test statistic is $X^2 = 11.8$ which has a P-value less than 0.5% according to the approximate null distribution χ^2_{df} with $\text{df} = 4 - 1 - 1 = 2$.

Reject the null hypothesis of independent voting.

7b. Friedman's test for $I = 3$ treatments and $J = 2$ blocks. The test statistic

$$Q = \frac{12J}{I(I+1)} \sum_{i=1}^{I} (\bar{R}_{i\cdot} - \frac{I+1}{2})^2$$

is obtained from the ranks given by two subjects (R_{ij}) to the three treatments. Under the null distribution all 36 possible rank combinations

$$(R_{ij}) = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 3 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 3 \end{pmatrix}, \ldots, \begin{pmatrix} 3 & 1 \\ 2 & 2 \\ 1 & 3 \end{pmatrix}, \begin{pmatrix} 3 & 1 \\ 2 & 3 \\ 1 & 2 \end{pmatrix}, \begin{pmatrix} 3 & 3 \\ 2 & 2 \\ 1 & 1 \end{pmatrix}$$

are equally likely. The corresponding vector of rank averages $(\bar{R}_{1\cdot}, \bar{R}_{2\cdot}, \bar{R}_{3\cdot})$ takes 5 values (up to permutations)

$$A_1 = (1,2,3), A_2 = (1,2.5,2.5), A_3 = (1.5,1.5,3), A_4 = (1.5,2,2.5), A_5 = (2,2,2)$$

according to the following table

	$1,2,3$	$1,3,2$	$2,1,3$	$2,3,1$	$3,1,2$	$3,2,1$
$1,2,3$	A_1	A_2	A_3	A_4	A_4	A_5
$1,3,2$	A_2	A_1	A_4	A_3	A_5	A_4
$2,1,3$	A_3	A_4	A_1	A_5	A_2	A_4
$2,3,1$	A_4	A_3	A_5	A_1	A_4	A_2
$3,1,2$	A_4	A_5	A_2	A_4	A_1	A_3
$3,2,1$	A_5	A_4	A_4	A_2	A_3	A_1

Next we have

$(\bar{R}_{1\cdot}, \bar{R}_{2\cdot}, \bar{R}_{3\cdot}) =$	A_1	A_2	A_3	A_4	A_5
$\sum_{i=1}^{3} (\bar{R}_{i\cdot} - 2)^2 =$	2	1.5	1.5	0.5	0
Probability =	1/6	1/6	1/6	1/3	1/6

Thus the null distribution of Q is the following one

$$P(Q=0) = 1/6, \quad P(Q=1) = 1/3, \quad P(Q=2) = 1/3, \quad P(Q=3) = 1/6.$$