Tentamentsskrivning i **Basics of mathematical statistics (TMS100), 3p.**

Tid: Onsdagen den 22 oktober, 2003 kl 08.45-12.45.

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

Hjälpmedel: kalkylator, 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 involving max 3 bonus marks:

grade "3" for 14 to 18 marks,

grade "4" for 19 to 23 marks,

grade "5" for 24 to 33 marks.

---

1. (6 marks) The following are the weaning weights (in pounds) of lambs in a large flock.

68 79 93 67 73 81 82 81 85 78
72 69 64 82 77 59 68 54 71 57
88 97 69 60 92 62 64 64 90 60

a. Estimate the mean and standard deviation of weight of weaning. What do these two numbers say about the weaning weight of lambs?

b. Draw a histogram to see the shape of the weaning weight distribution. Try different grouping intervals of equal length. Does the distribution look normal?

c. What is the standard error of the population mean estimate? Find two two-sided 90% confidence intervals for the average weight using two different formulas. Explain the difference between these two intervals.

d. How many additional observations would you need to make the 90% CI twice as short.

2. (6 marks) Suppose we have two small DNA sequences from two different species

species 1: gga gac tgt aga cag cta atg cta ta
species 2: gaa cgc cct agc cac gag ccc tta tc

a. We wish to gauge whether the two sequences show significant similarity to assess whether they have a remote common ancestor. Assume that the sequences were each generated at random, with the 4 letters a, g, c and t having equal probabilities of occurring at any position. What is the distribution of the

number of positions X where two sequences agree? Justify your answer.

b. What is the observed value x of X? Find $\mathrm{P}(X = x)$, and $\mathrm{P}(X \leq x)$ under the above mentioned model of two unrelated sequences. (Hint: to compute the second probability apply the normal approximation.)

c. Present in parametric form both the null hypothesis, stating that two sequences are unrelated, and the alternative hypothesis, stating that two sequences are related. (Hint: as parameter use the proportion of positions where the sequences agree.) Would you reject the null hypothesis for the observed sequences?

3. (6 marks) Think of a population with the following genotype frequencies for a certain gene with two alleles A and a:

Genotype	AA	Aa	aa
Frequency	0.46	0.48	0.06

a. Consider a random experiment of picking up an individual at random from the population and reading his genotype. Use the conditional probability formula to calculate the proportion of AA-individuals among all homozygous individuals. Illustrate with a Venn diagram.

b. Consider a two-step random experiment of, first, picking up an individual at random from the population, and second, choosing at random one of the alleles constituting individual's genotype. Arguing in terms of this experiment compute the allele frequencies of allele A and allele a.

c. Find the probability that in a random sample of 10 individuals 3 will have genotype AA and 5 will have genotype Aa.

d. Two randomly chosen individuals produce an offspring. Find the probability that the offsping is heterozygous.
Hint. Two offspring chromosomes are drawn independently at random from the population with allele frequencies calculated in (b). Heterozygosity means that one of the chromosomes has allele A while the other chromosome has allele a.

4. (6 marks) The paper "Root regeneration and early growth of red oak seedlings: influence of soil tempreture" reports the results of a regression analysis in which the independent variable x was daily degree hours of soil heat and the dependent y was shoot elongation per seedling (cm).

x	300	350	400	400	450	450	480	480
y	5.8	4.5	5.9	6.2	6.0	7.5	6.1	8.6

x	530	530	580	580	620	620	670	700
y	8.9	8.2	14.2	11.9	11.1	11.5	14.5	14.8

$\sum x_i = 8140$, $\sum x_i^2 = 4,340,600$
$\sum y_i = 145.7$, $\sum y_i^2 = 1505.01$, $\sum x_i y_i = 79,574$

a. Construct a scatterplot for the data. What relationship between x and y does it suggest?

b. Assuming that the simple linear reghression is appropriate, obtain the equation of the estimated regression line.

c. What proportion of variation in shoot elongation is explained by variation in soil heat?

d. Does there appear to be a useful linear relationship between the two variables? State and test an appropriate hypotheses.

5. (6 marks) A research project has been focused on the existence of any relationship between date of patient admission for treatment of alcoholism and patient's birthday. Assuming a 365-day year (i.e., excluding leap year), in the absense of any relation, a patient's admission date is equally likely to be any one of 365 possible days.

The investigators established four different admission categories:

1. within 7 days following patient's birthday,
2. between 8 and 30 days, inclusive, from the birthday,
3. between 31 and 90 days, inclusive, from the birthday,
4. more than 90 days from the birthday.

A sample of 200 patients gave observed frequencies of 11, 24, 69, and 96 for categories 1, 2, 3, and 4 respectively.

a. State an appropriate null hypothesis mathematically in a parametric form. Without analysing the data try to give some reasons why this null hypothesis might fail.

b. Test the relevant hypotheses using a significance level of 0.1. What are your conclusions?

c. The chi-square distribution with one degree of freedom is the distribution of a squared N(0,1)-variable. Using this fact explain how the normal distribution table helps to compute the P-value of the chi-square test with df = 1.

Statistical tables supplied:

1. Normal distribution. 2. Chi-square distribution 3. t-distribution.

Good luck!

ANSWERS

1a. $\bar{X} = 73.53$, $s^2 = 137.64$, $s = 11.73$

1b. Use, for example, intervals 51-60, 61-70, 71-80, 81-90, 91-100. Looking at this histogram it is difficult to say if the distribution is normal.

1c. Estimated standard error $s_{\bar{X}} = 2.14$. Approximate 90% CIs for μ: $73.53 \pm 1.645 \cdot 2.14$. Exact 90% CIs for μ: $73.53 \pm 1.70 \cdot 2.14$ (assumes that the data is normally distributed).

1d. 90 additional observations.

2a. X is $\mathrm{Bin}(26, 0.25)$.

2b. $X = 11$, $\mathrm{P}(X = 11) = \binom{26}{11} 0.25^{11} 0.75^{15} = 0.025$,

$$\mathrm{P}(X \leq 11) \approx \Phi(\tfrac{11-6.5}{2.21}) = 0.98.$$

2c. H_0: $p = 0.25$, H_1: $p > 0.25$. The one-sided P-value $P_1 = 0.02$. Reject H_0 at 5% significance level.

3a. $P(AA|AA \cup aa) = \frac{0.46}{0.46+0.06} = 0.885$, $P(aa|AA \cup aa) = 0.115$.

3b. Law of total probability $p_A = 1 \cdot 0.46 + 0.5 \cdot 0.48 + 0 \cdot 0.06 = 0.7$, $p_a = 0.3$.

3c. 0.0225 using Mn(10;0.46,0.48,0.06) distribution.

3d. $2p_A p_a = 0.42$.

4b. $y = 0.02733x - 4.798$

4c. 84%

4d. Model utility test of H_0: $\beta_1 = 0$ against H_1: $\beta_1 \neq 0$. Test statistic $T = \frac{b_1}{s_{b_1}} = 8.37$ is highly significant according to the t_{14} distribution table.

5a. H_0: $p_0 = \ldots = p_{364} = 1/365$, where p_i = probability of patient's admission at i-th day since the birthday.

5b. Apply chi-square test for the simple null hypothesis stated above. The chi-square test statistic $X^2 = 83.4$ is highly significant according to the χ^2_3 distribution table. There exists a relationship between the admission date and patient's birthday: more patients are admitted sooner after their birthday than expected under H_0.