Tentamentsskrivning i **TMS101: Basics of mathematical statistics, 3.5p**

Tid: Torsdagen den 21 oktober, 2004 kl 08.30-12.30.

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 four questions with the total number of marks 28. Attempt as many questions, or parts of the questions, as you can. Preliminary grading system (involving max 2 bonus marks):

grade "3" for 12 to 16 marks,

grade "4" for 17 to 21 marks,

grade "5" for 22 marks and more.

---

1. (7 marks) A polymorphic locus of a genom has three alleles A_1, A_2, A_3 with population frequencies $p_1 = 0.80$, $p_2 = 0.19$, $p_3 = 0.01$. A random sample of 100 chromosomes containing the locus was drawn from the population.

a. Compute the probability that allele A_3 is not represented in the sample.

b. Find the probability that the sample is monomorphic.

c. Using Poisson approximation find the probability that there are at least two alleles A_3 in the sample.

d. What is the mean and standard deviation of the number of alleles A_2 in the sample? Using normal approximation find the probability that there are less than 12 alleles A_2 in the sample.

2. (7 marks) The urinary fluoride concentration (measured in ppm) was determined for 11 randomly chosen livestock both at the beginning of and in the middle of their grazing period i a region previously exposed to fluoride pollution:

	1	2	3	4	5	6
Beginning	24.7	46.1	18.5	29.5	26.3	33.9
Middle	12.4	14.1	7.6	9.5	19.7	10.6

	7	8	9	10	11
Beginning	23.1	20.7	18.0	19.3	23.0
Middle	9.1	11.5	13.3	8.3	15.0

a. Estimate the true average difference between two urinary fluoride concentrations. What is the standard error of this estimate?

b. Compute a 95% exact confidence interval for the average difference assuming that the differences are normally distributed. Draw a histogram to see whether the last assumption is reasonable.

c. Does the data suggest that there has been a decrease in the true average urinary fluoride concentration during the period under consideration? Present your answer in terms of two competing hypotheses. What can be said about the corresponding P-value?

3. (7 marks) The CAGE score is a useful screening tool for alcoholism. It asks four questions resulting in the CAGE score - the number of positive answers - ranging from zero to four. A score less than 2 is considered to be a negative screening test for alcoholism, and a score ≥ 2 to be a positive test. Next come the data from a study of a certain population of size 518.

	Alcoholic	Not alcoholic	Total
CAGE score $= 2, 3, 4$	88	15	103
CAGE score $= 0, 1$	29	386	415
Total	117	401	518

a. What are the sensitivity and specificity of the CAGE test?

b. Estimate the prevalence of alcoholism in the population under study. Find the standard error and 95% confidence interval.

c. Find the positive predictive value = probability of disease among patients with a positive test.

d. Which of the above results can be called prior and posterior probabilities? Explain.

4. (7 marks) Scientists have suggested that animals use the earth's magnetic field as a clue to their orientation. An experiment to investigate this theory is conducted by using homing pigeons.

A pair of coils is placed around each pigeon and a magnetic field that reverses the earth's field is applied. This could disorient the bird. Each day for 118 consecutive days a single bird is released. The bird's orientation and the type of day is noted.

	Sunny	Cloudy	Total
Orient home	79	5	84
Get lost	16	18	34
Total	95	23	118

a. State in a parametric form the hypothesis of total disorientation due to the reversed magnetic field. Test this hypothesis at 5% significance level.

b. Suggest a modification of the described experiment to see if the earth's magnetic field plays any role in the bird's orientation.

c. Do the data of the table indicate that the bird's orientation is not independent of the cloud cover? Explain, based on the P-value of the appropriate test. Show how to use the normal probability table for finding the P-value.

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

Good luck!

ANSWERS

1a. 0.366

1b. $2 \cdot 10^{-10}$

1c. Pois(1) distribution: $p_0 + p_1 = 2e^{-1} = 0.736$. $1 - 0.736 = 0.264$.

1d. Mean 19, variance 15.39. Normal approximation $\mathrm{N}(19, 3.92^2)$ for the binomial distribution applied with continuity correction

$$\mathrm{P}(X < 11.5) = \mathrm{P}(\frac{X - 19}{3.92} < -1.91) \approx 0.028$$

2a. $\bar{X} = 13.82, s^2 = 66.79, s = 8.17, s_{\bar{X}} = 2.46$.

2b. Assuming normally distributed differences the 95% CI for the mean difference μ is

$$13.82 \pm 2.228 \cdot 2.46 = 13.82 \pm 5.48$$

The histogram profile is somewhat skewed to the right, still the normality assumption seems reasonable.

2c. H_0: $\mu_1 = \mu_2$, H_1: $\mu_1 \neq \mu_2$ or in terms of $\mu = \mu_1 - \mu_2$

$$H_0 : \mu = 0, H_1 : \mu \neq 0.$$

The one sample t-test statistic is $\frac{13.82}{2.46} = 5.62$. The two-sided P-value for this value according to the t-distribution table with df $= 10$ is smaller than $P_1 = 0.01$. Reject H_0 at 1% significance level and conclude that there was a significant decrease in the average concentration.

3a. Sensitivity measures the ability of the test to recognize a positive case $= 88/117 = 0.75$, while the specificity is the poportion of true negatives $= 386/401 = 0.96$.

3b. The prevalence of alcoholism is found exactly $p = 117/518 = 0.226$ since the data in the table describe the whole population. Thus the standard error is zero and the 95% CI is 0.226 ± 0.

3c. The positive predictive value is $88/103 = 0.85$.

3d. The prevalence of alcoholism is a prior probability of a randomly chosen person to be an alcoholic. The positive predictive value is the posterior probability of a randomly chosen person to be an alcoholic given a positive test result.

4a. Test the simple null hypothesis $H_0 : p_{\text{home}} = p_{\text{lost}} = 0.5$ using the Pearson chi-square test. Expected counts are 59 and 59, so the observed test statistic value is $X^2 = 21.2$. Consult the distribution table of the chi-square distribution with df=2 to see that the critical value at 0.5% level is 10.6. We reject the hypothesis of total disorientation and conclude that even with the reversed magnetic field the pigeons retain some ability to orient home.

4b. Compare orienation of birds with and without the reversed magnetic field. Assign the birds to two groups at random and send the birds pairwise: one with the reversed magnetic field and the other without.

4c. Apply the chi-square test for independence. The observed test statistic value $X^2 = 34$ is larger than the largest critical value in the table for the chi-square distribution with one df 7.88. Reject the null hypothesis of independence at level 0.5%. With df=1 one can use the normal distribution table to estimate the P-value of the test. Take the square root of 34 which is 5.8 and consult the normal distribution table. It gives a one-sided P-value smaller than 0.02%. We conclude that the P-value of the chi-square test is smaller than 0.04%.