Tentamentsskrivning i matematisk statistik: **Basics of Math. Statistics, 3p.**

Tid: Tisdagen den 17 oktober 2000 kl 8.45-12.45.

Lärare och Jour: Serik Sagitov, tel. 772-5351, rum MC 1420.

Hjälpmedel: egen formelsamling (fyra A4-blad), utdelade tabeller samt godkänd miniräknare

Grading system (preliminary):

points	0-11	12-16	17-21	22-30
grade	U	3	4	5

1.(5 points) To a good approximation, the restriction sites for a particular enzyme occur along a chromosome according to a Poisson process with rate $\lambda = 10$ restriction sites per kilobase. Imagine that you have picked up at random two DNA fragments produced by the enzyme. Let L_1 and L_2 be the lengths of the fragments and put $M = L_1 + L_2$.

a. Describe the probability distribution of L_1.

b. In what sense the distribution of L_1 is memoryless?

c. Write down the joint density function for the random variables L_1 and L_2.

d. Compute the mean value and standard deviation of M.

2.(5 points) A rare genetic disease is discovered. Although only one in million people carry it, you consider getting screened. You are told that the genetic test is extremely good; it is 100% sensitive (it is always correct if you have the disease) and 99.99% specific (it gives a false positive result only 0.01% of the time).

a. How would you estimate your chances of having the disease prior the genetic test?

b. What is the total probability of getting a positive test result?

c. Find the posterior probability of having the disease given a positive test result.

d. Answer to the questions a-c assuming now that the genetic test has produced a positive result and you think about repeating the test.

3.(5 points) A particular IQ test was taken by a sample of 22 subjects with the sample mean and sample variance computed to be 107 and 241. To see if the population mean is significantly different from 100

a) state a simple null hypothesis and a two-sided alternative hypothesis;

b) apply the one-sample t-test at 5% significance level;

c) compute the p-value of the test;

d) using an appropriate confidence interval find if the population IQ-mean is significantly different from 100 at 1% significance level.

4.(5 points) Which of the following statements are correct? Explain.

a. The width of a 99% confidence interval is twice the width of an 80% confidence interval.

b. If the variance of an unbiased estimate tends to zero as the sample size n tends to infinity, the estimate is consistent.

c. If a chi-square test statistic with 5 degrees of freedom has a value 13, the p-value is less than 2.5% and greater than 1%.

d. The probability that the null hypothesis is correctly rejected is equal to the power of the test.

5.(5 points) The joint frequency function of two r.v. X and Y is given with one ommission:

	x=1	x=2	x=3
y=1	.20	.15	?
y=2	.05	.15	.10
y=3	.00	.10	.25

a. Find the marginal frequency functions of X and Y.

b. Compute $\mathrm{E}(X)$, $\mathrm{E}(X^2)$, $\mathrm{Var}(X)$, and σ_X.

c. Find the conditional expectations $\mathrm{E}(X|Y=1)$, $\mathrm{E}(X|Y=2)$, $\mathrm{E}(X|Y=3)$.

d. Verify the Law of Total Expectation using your answers to a-c.

6.(5 points) Given the DNA sequence

GTG CAC TGG ACT GCT GAG GAG AAG

and using the Dirichlet distribution with parameters $\alpha_A = 3, \alpha_C = 2, \alpha_G = 2, \alpha_T = 3$ as the prior distribution for the four nucleotide frequences (p_A, p_C, p_G, p_T)

a) compute the prior mean values for the four nucleotide frequences;

b) find the posterior joint distribution for (p_A, p_C, p_G, p_T);

c) compute the posterior mean estimates for the four nucleotide frequences;

d) compare the posterior mean estimates with the maximum likelihood estimates for (p_A, p_C, p_G, p_T).

Partial answers and solutions are also welcome. Good luck!

ANSWERS

1a. The length of a DNA fragment is $L_1 \in Exp(10)$ if measured in kilobases or if measured in basepaires $L_1 \in Geom(0.01)$.

1b. $\mathrm{P}(L_1 > x + y | L_1 > x) = \mathrm{P}(L_1 > y)$.

1c. Due to independence between L_1 and L_2 when measured in kb

$$f_{L_1 L_2}(x, y) = f_{L_1}(x) f_{L_2}(y) = 100 \cdot \exp(-10(x + y)).$$

1d. According to the formulas for the exponential distribution

$$E(M) = E(L_1) + E(L_2) = \frac{1}{10} + \frac{1}{10} = 0.2 \ kb,$$

and due to independence

$$Var(M) = Var(L_1) + Var(L_2) = \frac{1}{100} + \frac{1}{100} = 0.02 \ kb^2, \ \sigma_M = 0.14 \ kb.$$

2a. Prior the genetic test the probability of B={you have the disease} might be estimated as $\mathrm{P}(B)$=0.000001.

2b. We know that the conditional probabilities of A={positive test result} are $\mathrm{P}(A|B)$=1 and $\mathrm{P}(A|\bar{B})$=0.0001. Use the LTP to see that $\mathrm{P}(A)$=1·0.000001+ $0.0001 \cdot 0.999999$=0.000101.

2c. According to Bayes' formula: $\mathrm{P}(B|A)=\frac{1 \cdot 0.000001}{0.000101}$=0.0099.

2d. Prior the second test the probability of B is $\mathrm{P}^*(B)=\mathrm{P}(B|A)$=0.01 and according to the LTP the probability of C={second positive test result} is $\mathrm{P}^*(C)=1 \cdot 0.01 + 0.0001 \cdot 0.99$=0.0101. Thus the posterior probability of having the disease after two positive results is $\mathrm{P}^*(B|C)=\frac{1 \cdot 0.01}{0.0101}$=0.99.

3a. $H_0 : \ \mu = 100$ vs. $H_A : \ \mu \neq 100$

3b. The observed T-score: $\frac{107-100}{\sqrt{241/22}}$=2.115 is greater than the critical level $t_{21}(0.025)$=2.08. Therefore H_0 is rejected at 5% significance level.

3c. According to Table 4 the p-value of the t-test lies between 2% and 5% beeing closer to the latter.

3d. Assuming that the IQ-score is normally distributed we calculate the exact 99% confidence interval of the mean as

$$107 \pm t_{21}(0.005) \cdot \sqrt{241/22} = 107 \pm 9.37.$$

Since the interval covers the value 100 we accept H_0 at 1% significance level.

4a. Correct with the confidence interval formula $\bar{X} \pm z(\frac{\alpha}{2}) \cdot \frac{s}{\sqrt{n}}$ in mind.

4b. Correct.

4c. Correct, because due to Table 3 $\chi^2_5(0.025)$=12.83 and $\chi^2_5(0.01)$=15.09.

4d. Correct.

5a. P($X = 1$)=0.25, P($X = 2$)=0.4, P($X = 3$)=0.35;
P($Y = 1$)=0.35, P($Y = 2$)=0.3, P($Y = 3$)=0.35.
5b. E(X)=2.1, E(X^2)=5, Var(X)=0.59, σ_X=0.77.,
5c. E($X|Y = 1$)=1.43, E($X|Y = 2$)=2.17, E($X|Y = 3$)=2.71.
5d. E(X)=1.43×0.35+2.17×0.3+2.71×0.35=2.1.

6a. E(p_A)=$\frac{3}{10}$, E(p_C)=$\frac{2}{10}$, E(p_G)=$\frac{2}{10}$, E(p_T)=$\frac{3}{10}$.
6b. Dirichlet(9,6,12,7).
6c. $\hat{p}_A$=$\frac{9}{34}$=0.265, $\hat{p}_C$=$\frac{6}{34}$=0.177, $\hat{p}_G$=$\frac{12}{34}$=0.353, $\hat{p}_T$=$\frac{7}{34}$=0.206.
6d. MLE $\hat{p}_A$=$\frac{6}{24}$=0.25, $\hat{p}_C$=$\frac{4}{24}$=0.167, $\hat{p}_G$=$\frac{10}{24}$=0.417, $\hat{p}_T$=$\frac{4}{24}$=0.167 do not use prior information available.