Tentamentsskrivning i **Basics of Math. Statistics (TMS100), 3p.**

Tid och plats: måndag, oktober 21, 2002 kl 8.45-12.45 i V-huset.
Examinator och jour: Serik Sagitov, tel. 772-5351, rum MC 1421.
Hjälpmedel: kalkylator, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.

There are six 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 and more.
The grading system without bonus marks (applies to years 2000-2001):
grade "3" for 12 to 16 marks,
grade "4" for 17 to 21 marks,
grade "5" for 22 to 30 marks.

---

1. (5 marks) Adult-onset diabetes is known to be highly genetically determined. A study was done comparing frequencies of a particular allele in a sample of such diabetics and a sample of nondiabetics. The data are shown in the following table:

	Diabetic	Normal
Bb or bb	12	4
BB	39	49

a. State two hypotheses of interest. What statistical test can be applied here? Explain your choice of the test.

b. Describe the test statistic and its null distribution. Find the rejection regions at 5% significance level and 1% significance level. Do you accept or reject the null hypothesis?

c. Find the exact P-value of the test. What is the meaning of this probability? How this result agrees with your conclusions in b)?

2. (5 marks) In a raid on a coffee shop, Bayesian trading inspectors take a random sample of 20 packets of coffee, each of nominal weight 125 g. The data they obtain are (weights in grams):

105.3, 113.3, 114.5, 121.2, 122.9, 123.7, 124.0, 124.6, 124.9, 124.9,
124.9, 125.1, 125.5, 125.9, 126.8, 127.7, 128.2, 128.3, 128.5, 130.2,

($\sum x_i = 2470.4$, $\sum x_i^2 = 305829.0$). They model these data as independent values from a Normal $\mathrm{N}(\mu, \sigma^2)$ distribution with known $\sigma^2 = 4$. For μ they assume a prior distribution $N(\mu_0, \sigma_0^2)$ with $\mu_0 = 126$ and $\sigma_0^2 = 1$. The inspectors can impose a fine if their 95% credibility interval falls wholly below the claimed value of $\mu = 125$ g.

a. Find the inspector's posterior distribution. Express the mean of this distibution as a weighted average of μ_0 and the sample mean.

b. Show that the inspector's 95% credibility interval falls wholly below 125 g; they therefore impose a fine on the owners of the coffee shop.

c. Comment briefly as to whether the inspectors are justified in imposing a fine on the basis of this sample.

3. (5 marks) Assume that the joint distribution of three random variables (X_1, X_2, X_3) is multinomial Mn(5;0.5,0.25,0.25).

a. What are the marginal distributions of X_1, X_2, and X_3? What is the correlation coefficient for the random variables X_1 and X_3? Explain the negative sign of this coefficient.

b. Vector (X_1, X_2, X_3) describes the bar heights in a histogram recording the outcome of a certain sampling experiment. Describe the sample size, population distribution, and the sampling procedure.

c. Present in a table format the joint probabilities for (X_1, X_2).

4. (5 marks) In a study of the population size N of racoons in northern Florida a sample of 48 animals was captured. They were injected with a small amount of a radioactive isotope. The "recapture" sample involved a search of the study area for scats (feces = exkrementer). During a certain period of search it was found 71 scats with 31 of them beeing radioactive.

a. Consider the number X of radioactive scats in the recapture sample. Under what assumptions its sampling distribution can be viewed as binomial $X \in \text{Bin}(71, \frac{48}{N})$?

b. Write down the likelihood function of the parameter under estimation N using the data and model given above. Compute the maximum likelihood estimate of N, or explain how you can find it using computer.

c. Find the method of moment estimate of N.

d. Find an approximate 95% confidence interval for N. (Hint: find first a confidence interval for the parameter $p = \frac{48}{N}$.)

5. (5 marks) Three door puzzle. You are on a game show. There are 3 doors. Behind one of the doors, there is a great prize. Behind the other two doors, there is nothing. To win the prize, you must correctly guess which door it is behind and open that door. But wait ... there's more! After you guess, the host opens one of the doors and gives you a chance to change your mind. To analyse the game chances using the probability theory let us introduce two events:
A = the prize is behind the door you have chosen first,
B = the door opened by the host has no prize behind.

a. Assume first that the host knew where the prize was and deliberately opened one without the prize. What is $P(B)$ in this case? Compare $P(A)$ with $P(A|B)$. Is it worth to change your mind and open the other door?

b. Now assume that the host didn't know where the prize was placed and opened a door without the prize by chance. What are $P(B|A)$ and $P(B)$ in this case? Compare the prior $P(A)$ with the posterior probability $P(A|B)$. Shall you change your mind and open the other door knowing that the host opened the door at random?

6. (5 marks) The following figures show the result of an experiment with ten patients on the effect of two supposedly sopoforic drugs, A and B, in producing sleep.

Patient	Drug A	Drug B	Difference (B – A)
1	+0.7	+1.9	+1.2
2	–1.6	+0.8	+2.4
3	–0.2	+1.1	+1.3
4	–1.2	+0.1	+1.3
5	–0.1	–0.1	0.0
6	+3.4	+4.4	+1.0
7	+3.7	+5.5	+1.8
8	+0.8	+1.6	+0.8
9	0.0	+4.6	+4.6
10	+2.0	+3.4	+1.4
Mean	+0.75	+2.33	+1.58

a. Assuming that the difference is normally distributed find a 99% confidence interval for the mean difference μ. State clearly what property of the normal distribution do you use here.

b. Apply the principle of duality between CI and hypotheses testing for the purpose of testing $H_0 : \mu = 0$ against a two-sided alternative.

c. Discuss the sampling design of the experiment. Does it take care of the placebo effect? Explain.

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

Good luck!

ANSWERS

1a. H_0: same frequencies for diabetic and normal people or $p_{11} = p_{12}$ and $p_{21} = p_{22}$. H_1: different frequencies for diabetic and normal people or $p_{11} \neq p_{12}$. Apply the chi-square test of homogeneity because the data are collected as two samples, and H_0 states that two population distributions are equal.

1b. Observed test statistic $X^2 = 5.1$. Approximate null distribution is χ^2 with df=1. Critical values: at 5% significance level $\chi_1^2(0.05) = 3.84$, at 1% significance level $\chi_1^2(0.05) = 6.63$. Reject H_0 at 5% significance level and accept H_0 at 1% significance level.

1c. P-value 2(1-Φ(2.26))=0.024. It is the probability of the observed deviation from the results predicted by H_0 assuming that the null hypothesis is true.

2a. $\bar{X} = 123.52$, $\mu_1 = 126 \cdot \frac{4}{4+20\cdot 1} + 123.52 \cdot \frac{20\cdot 1}{4+20\cdot 1} = 123.93$, $\sigma_1^2 = \frac{4\cdot 1}{4+20\cdot 1} = 0.167$. Posterior distribution: $\mu \in \mathrm{N}(123.93; 0.167)$.

2b. (123.60, 124.26).

2c. $s = 6.01$ contradicts $\sigma = 2$ assumption. Three outliers (105.3, 113.3, 114.5) contradict the normality assumption.

3a. $X_1 \in \mathrm{Bin}(5; 0.5)$, $X_2 \in \mathrm{Bin}(5; 0.25)$, $X_3 \in \mathrm{Bin}(5; 0.25)$, $\rho_{13} = -0.58$.

3b. Population distribution: $p_1 = 0.5, p_2 = p_3 = 0.25$. Sampling procedure: $n = 5$ independent observations.

3c.

	$X_2 = 0$	$X_2 = 1$	$X_2 = 2$	$X_2 = 3$	$X_2 = 4$	$X_2 = 5$
$X_1 = 0$	0.001	0.005	0.010	0.010	0.005	0.001
$X_1 = 1$	0.010	0.039	0.059	0.039	0.010	0.000
$X_1 = 2$	0.039	0.117	0.117	0.039	0.000	0.000
$X_1 = 3$	0.078	0.156	0.078	0.000	0.000	0.000
$X_1 = 4$	0.078	0.078	0.000	0.000	0.000	0.000
$X_1 = 5$	0.031	0.000	0.000	0.000	0.000	0.000

4a. Each scat found can belong to any of N animals so that the model of sampling with replacement applies.

4b. $L(N) = \binom{71}{31}(\frac{48}{N})^{31}(1-\frac{48}{N})^{40}$. Matlab solution of finding the MLE:

```
>>[y,i]=max(binopdf(31,71,48./(49:1000)));N=48+i
>>N=110
```

Or solve $\frac{d\log L(N)}{dN} = 0$, that is $-\frac{31}{N} + \frac{40}{1-48/N}\frac{48}{N^2} = 0$.

4c. $\tilde{N} = 109.9$

4d. $\hat{p} = 0.4366$, $s_{\hat{p}} = 0.0593$, 95% CI for p is $0.4366 \pm 0.0593 \cdot 1.96 = [0.3204, 0.5528]$. Therefore a 95% CI for N is [86.8,149.8].

5a. P(B)=1, P(A)=1/3, P($B|A$)=1, $P(A|B) = \frac{1 \cdot 1/3}{1} = 1/3$. Since $P(A|B) < P(\bar{A}|B)$ it is worth to open the other door.

5b. P(A)=1/3, P($B|A$)=1, P($B|\bar{A}$)=1/2, $P(B) = 1 \cdot 1/3 + 1/2 \cdot 2/3 = 2/3$. Therefore $P(A|B) = \frac{1 \cdot 1/3}{2/3} = 1/2$. Since $P(A|B) = P(\bar{A}|B)$ the prize is behind one the two closed doors with equal probabilities.

6a. $\bar{X} = 1.58$, $s = 1.23$, $s_{\bar{X}} = 0.39$, $t_{0.005}(9) = 3.25$, $s_{\bar{X}}t_{0.005}(9) = 1.27$, $1.58 \pm 3.25 \cdot 0.39 = 1.58 \pm 1.27$.

6b. Reject the null hypothesis since the CI (1.58 ± 1.27) does not cover $\mu_o = 0$.

6c. Yes, since possible placebo effect cancels out when the difference is taken.