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

Tid: Torsdagen den 25 oktober 2001 kl 08.45-12.45.
Examinator och jour: Serik Sagitov, tel. 772-5351, rum MC 1420.
Hjälpmedel: kalkylator, egen formelsamling (4 sidor på 2 blad A4) samt utdelade tabeller.

———————————————

There are six questions with the total marks available being 30 (= 4+6+5+5+4+6). Attempt as many questions, or parts of the questions, as you can. Preliminary grading system:
grade "3" for 12 to 16 marks,
grade "4" for 17 to 21 marks,
grade "5" for 22 to 30 marks.

———————————————

1.(5 marks) Suggest a probability distribution for each of the following. Explain your choice.

a. The number of accidents in a large factory during one 8-hour shift.

b. The number of spades in a bridge hand (ie in a random selection of 13 cards from a pack of 52 cards).

c. The number of the coin-die experiments untill the first '6' is obtained. Compute the mean of this random number.

d. The number of beetles that are killed when a random sample of 40 beetles is subjected to a specified dose of insecticide.

2.(6 marks) Some individuals are carriers of the bacterium Streptococcus pyogenes. To investigate whether there is a relationship between carrier status and tonsil size in schoolchidren, 1398 children were examined and classified according to their carrier status and tonsil size.

	Carrier status		
Tonsil size	Carrier	Non-carrier	Total
Normal	19	497	516
Large	29	560	589
Very large	24	269	293
Total	72	1326	1398

a. State the null and alternative hypotheses.

b. State the form of the test statistic and specify the values of the test statistic that would lead to rejection of the null hypothesis, at 5% level of significance.

c. What would you conclude from the test?

d. Find a point estimate of, and a 95% confidence interval for, the proportion of schoolchidren with normal sized tonsils who are carriers.

3.(4 marks) A loaded die was rolled 6 times with the outcome 324166.

a. Using this data write down the likelihood function of the probability p_6 of observing '6' with one roll of the die.

b. Find the MLE of p_6 and verify that it indeed maximizes the likelihood function obtained previously.

4.(5 marks) An automobile insurance company classifies each driver as a good risk (A_1), a midium risk (A_2), or a poor risk (A_3). Of those currently insured, 30% are good risks, 50% are medium risks, and 20% are poor risks. In any given year, the probability that a driver will have at least one citation is 0.1 for a good risk, 0.3 for a medium risk, and 0.5 for a poor risk.

a. Why do the last three probabilities not sum up to 1: 0.1+0.3+0.5=0.9?

b. If a randomly selected driver insured by this company has at least one citation during the next year, what is the probability that the driver was actually a good risk? A medium risk?

c. Suppose that each insuree has a three-year policy. If accidents are independent from year to year what is the probability that a good risk driver will have at least one citation during a three-year period? Compute similar probabilities for a medium risk, and a poor risk drivers.

d. If a randomly selected driver reports to sitations during the three years, what is the probability that the driver was actually a good risk?

5.(4 marks) Each year a farmer has his ewes (female sheep) impregnated using artificial insemination. The number of lambs born by an individual ewe (X), is known to have the following distribution:

x	0	1	2	3
P(X=x)	0.1	0.3	0.4	0.2

a. Find the mean and standard deviation of X.

b. The farmer has a flock of 100 ewes. If T denotes the total number of lambs born to these 100 ewes, and assuming independence of the numbers of lambs born to different ewes, use a normal approximation to find P($T \geq 160$).

6.(6 marks) Beta distribution $\mathrm{B}(a,b)$ has the following pdf:

$$f(p)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}p^{a-1}(1-p)^{b-1},\ 0<p<1.$$

a. What does it mean that $\mathrm{B}(a,b)$ is a conjugate prior distribution for the binomial distribution $\mathrm{Bin}(n,p)$?

b. Why $\mathrm{B}(1,1)$ might be called a noninformative prior?

c. Because of genetic differences, some people can taste phenyl thiocarbamide and others cannot. Three students from a psychology class were sent out to get data. Their results were as follows:
student 1: 109 persons tested, 23% could not taste;
student 2: 43 persons tested, 21% could not taste;
student 3: 71 persons tested, 28% could not taste.
What is the MLE of the fraction of people who cannot taste this chemical?

d. Suggest a Bayesian estimate of the fraction of people who cannot taste this chemical? Explain your choice of the prior.

Good luck!

ANSWERS

1a. The number of accidents in a large factory during one 8-hour shift can be modelled by a Poisson Pois(λ) distribution. Here λ is the average number of accidents per 8-hour shift. Assumptions:
large number of workers n,
small probability of an accident p for a worker during one 8-hour shift, λ=np,
independence of accidents for different workers.

2b. The number of spades in a bridge hand (ie in a random selection of 13 cards from a pack of 52 cards) can be modelled by a hypergeometric distribution Hg(52,13,0.25).

The corresponding random experiment is sampling of n=13 cards from N=52 cards without replacement. The initial proportion of spades is p=0.25.

c. The number of the coin-die experiments untill the first '6' is obtained can be modelled by a geometric distribution Geom(p). Here p is the probability of observing a '6' in a coin-die experiment. We use geometric distribution when model the number of independent trials untill the first success.

Due to LTP $p=0.5\cdot\frac{1}{6}+0.5\cdot(1-\frac{5}{6}\cdot\frac{5}{6})=0.24$. The mean of this random number is $1/p$=4.24.

d. The number of beetles that are killed when a random sample of 40 beetles is subjected to a specified dose of insecticide can be modelled by a binomial distribution Bin$(40, p)$. We have n=40 independent trials with unknown probability p of a beetle to be killed by the insecticide.

2a. Two characteristics of a child: tonsil size and carrier status. H_0: "two factors are independent" versus H_1: "two factors are dependent".

2b. χ^2-test of independence. Test statistic $X^2=\sum_{i=1}^{3}\sum_{j=1}^{2}\frac{(n_{ij}-n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}}$. Reject H_0 at 5% level if $X^2 > 5.99$.

2c. Observed $X^2 = 7.9$. According to the χ^2-distribution table with df=2 the P-value of the test is between 1.0% and 2.5%. Conclusion: reject H_0 at 2.5% level.

2d. Point estimate $\hat{p} = \frac{19}{1398} = 0.0136$. A 95% CI is 0.0136 ± 0.0061 or (0.75%,1.97%).

3a. Statistical model: the number of '6' in 6 rolls X∈Bin(6,p_6). Since$X$=2 the likelihood function is lik(p_6)=P(X=2)=15$p_6^2(1-p_6)^4$.

3b. The MLE p_6 is $\hat{p}_6$=$\frac{2}{6}$=0.333. To verify that it maximizes the likelihood function lik(p_6)=15$p_6^2(1-p_6)^4$ show that the derivative of the log-likelihood l(p_6)=log(15)+2log(p_6)+4log($1-p_6$) is zero at p_6=$\frac{1}{3}$.

Otherwise, one can just compare lik(p_6) for various values of p_6

p_6	0.100	0.200	0.300	0.333	0.400	0.600	0.800
lik(p_6)	0.098	0.246	0.324	0.329	0.253	0.138	0.015

4a. Random experiment: pick up a driver at random. Random events of interest: A_1=the driver is a good risk, A_2=the driver is a medium risk, A_3=the driver is a poor risk,

B=the driver has at least one citation during one year,

C=the driver has at least one citation during a three-year period.

Events (A_1, A_2, A_3) form a partition of the sample space and their probabilities P(A_1)=0.3, P(A_2)=0.5, P(A_3)=0.2 sum up to 1. However, the conditional probabilities P($B|A_1$)=0.1, P($B|A_2$)=0.3, P($B|A_3$)=0.5 are not suppossed to sum up to 1.

4b. LTP: P(B)=0.1·0.3+0.3·0.5+0.5·0.2=0.28. Bayes rule gives the following posterior probabilities

P($A_1|B$)=$\frac{0.1\cdot 0.3}{0.28}$=0.11, P($A_2|B$)=$\frac{0.3\cdot 0.5}{0.28}$=0.54, P($A_3|B$)=$\frac{0.5\cdot 0.2}{0.28}$=0.36.

4c. Since $\overline{C}$=$\overline{B_1}\cap\overline{B_2}\cap\overline{B_3}$, due to independence P($\overline{C}|A_1$)=0.9^3=0.73 and P($C|A_1$)=0.27. Similarly P($C|A_2$)=1-0.7^3=0.66 and P($C|A_3$)=1-0.5^3=0.88.

LTP: P(C)=0.27·0.3+0.66·0.5+0.88·0.2=0.59. Bayes rule gives the following posterior probabilities

P($A_1|C$)=$\frac{0.27\cdot 0.3}{0.59}$=0.14, P($A_2|C$)=$\frac{0.3\cdot 0.5}{0.28}$=0.56, P($A_3|C$)=$\frac{0.5\cdot 0.2}{0.28}$=0.30.

5a. E(X)=1.7, E(X^2)=3.7, Var(X)=0.81, σ=0.9.

5b. Use a normal approximation with the mean 1.7·100=170 and the standard deviation 0.9·10=9 to find

P($T\geq 160$)≈1-Φ($\frac{160-170}{9}$)=1-Φ(−1.11)=Φ(1.11)=0.87.

6b. B(1,1) is a uniform distribution over the unit interval (0,1). All possible values of the parameter $p\in(0,1)$ are equally acceptable prior the experiment. It reflects our luck of prior information concerning the parameter.

6c. The observed counts are 54 can not taste and 169 can. MLE: $\hat{p}=0.244$.

6d. Having no prior information about p we can use B(1,1) as a prior distribution. We add pseudocounts (1,1) to the observed counts (54,169) and obtain a Bayesian estimate of the fraction of people who cannot taste this chemical:

$\hat{p} = 0.244$.