STAT 305 – Introduction to Statistical Inference **Winter 2004**

Final Exam — Practice

Time allowed : 150 minutes.

Authorized material :

- One letter-size cheat sheet (2-sided).
- One scientific calculator without wireless communication feature.

Instructions :

- The exam has 8 pages including this one.
- Answer all 7 questions; the total number of points is 100.
- This exam is worth 35% of the term. You must pass this exam to pass the course.
- Write legibly; give complete solutions. Marks may be taken away for unclear solutions.
- You can use the back side of the sheets as drafts. If you use it for writing answers, indicate it clearly.
- Please use the tables at the end of the textbook if needed. If this were the true exam, tables would be provided.

Last Name : Solutions.

First Name : ____________________

Student Number : ____________________

Signature : ____________________

Question 1

[10 pts] The Laplace distribution has the following density function,

$$f(x) = \frac{\lambda}{2} e^{-\lambda|x-\theta|}$$

which depends on the parameters $\lambda > 0$ and θ. The expectation and variance of that distribution are as follow:

$$\mathrm{E}(X) = \theta\,, \qquad \mathrm{var}(X) = \frac{2}{\lambda^2}.$$

Find estimators of λ and θ using the method of moments.

Let's find $E(X^2)$.

$$E(X^2) = var(X) + \{E(X)\}^2 = \frac{2}{\lambda^2} + \theta^2$$

The method of moments consists of solving the equations:

$$\begin{cases} \theta = \overline{X} \\ \frac{2}{\lambda^2} + \theta^2 = \overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2 \end{cases}$$

Replacing the second equation in the first, we get:

$$\frac{2}{\lambda^2} + (\overline{X})^2 = \overline{X^2}$$

$$\lambda^2 = \frac{2}{\overline{X^2} - (\overline{X})^2}$$

$$\lambda = \sqrt{\frac{2}{\overline{X^2} - (\overline{X})^2}}$$

Therefore, we get

$$\boxed{\begin{aligned} \tilde{\theta} &= \overline{X} \\ \tilde{\lambda} &= \sqrt{\frac{2}{\overline{X^2} - (\overline{X})^2}} \end{aligned}}$$

Question 2

A Bayesian analysis is performed on a sample of n observations from a Poisson distribution with parameter λ. The prior knowledge on λ is modeled by a χ^2 distribution with k degrees of freedom.

a) [10 pts] Find the posterior distribution of λ.

b) [5 pts] Give a $100(1-\alpha)\%$ credible interval for λ. Express your answer in terms of quantiles you could find in one of the tables at the end of the book. *[Hint: Consider the distribution of $(2n+1)\lambda$.]*

a) $X_1, \dots, X_n \sim$ Poisson (λ)

$\pi(\lambda) \sim \chi^2_k = \Gamma(k/2, 1/2)$ is the prior distribution for λ.

The posterior distribution is the "updated" distribution of λ.

$$f(\lambda \mid X_1, \dots, X_n) \propto f(X_1, \dots, X_n \mid \lambda) \cdot \pi(\lambda)$$

$$= \left\{ \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{X_i}}{X_i!} \right\} \cdot \frac{1}{2^{k/2}\,\Gamma(k/2)} \lambda^{k/2-1} e^{-\lambda/2}$$

$$\propto e^{-\lambda n} \lambda^{\Sigma X_i} \lambda^{k/2-1} e^{-\lambda/2}$$

$$= \lambda^{\Sigma X_i + k/2 - 1} e^{-(n+\frac{1}{2})\lambda} \sim \Gamma\left(\Sigma X_i + k/2,\ n + 1/2\right).$$

A gamma distribution with parameters $\left(\sum_{i=1}^{n} X_i + \frac{k}{2}\right)$ and $\left(n + \frac{1}{2}\right)$.

since the posterior distribution is a function of λ.

b) Let's consider

$$(2n+1)\lambda \sim \Gamma\left(\frac{2\sum_{i=1}^{n} X_i + k}{2},\ \frac{2n+1}{2}\Big/(2n+1)\right) \sim \Gamma\left(\frac{2\sum_{i=1}^{n} X_i + k}{2},\ \frac{1}{2}\right) \sim \chi^2_{2\sum_{i=1}^{n} X_i + k}.$$

We used that result many times; it was proved in MT1, Q2.b.

The $100(1-\alpha)\%$ Credible interval is given by the quantiles $\frac{\alpha}{2}$ and $1-\frac{\alpha}{2}$ of the posterior distribution of λ. These quantiles can be found by using the result above since.

$$\frac{\alpha}{2} = P(\lambda \le \lambda_L) = P(\underbrace{(2n+1)\lambda}_{\sim \chi^2_{2\Sigma X_i + k}} \le (2n+1)\lambda_L). \iff \lambda_L = \frac{\chi^2_{\frac{\alpha}{2},\, 2\sum_{i=1}^{n} X_i + k}}{2n+1}.$$

Similarly, we can find λ_u and we get $1-\alpha = P\left(\lambda \in \left[\frac{\chi^2_{\frac{\alpha}{2}, 2\Sigma X_i + k}}{2n+1}, \frac{\chi^2_{1-\frac{\alpha}{2}, 2\Sigma X_i + k}}{2n+1}\right]\right)$

Question 3

Your group of friends and yourself (total of 30 people) think a coin may be biased and gives head more often than tail. You would like to find statistical evidence to support that hypothesis. To test whether the coin is fair, each of you will flip the coin until a first tail appears and record the number of throws needed.

a) [10 pts] Write down the distribution of your measurements and the hypothesis you want to test. Find the shape of the uniformly most powerful test for these hypothesis. *[Hint: H_1 should be what you want to prove.]*

b) [5 pts] Find the approximate critical value that gives the test above a level of 100(1-α)%?

c) [5 pts] If ~~$n = 30$ and~~ $\alpha = 5\%$, what would be the approximate probability to detect that the actual probability of of having tail when you flip the coin is 45%? How do you call that probability?

a) $X_1, \dots, X_{30} \sim \text{Geo}(p)$ where $p = P$(having tail when the coin is flipped).

$$(*) \begin{cases} H_0: & p \geqslant 1/2 \\ H_1: & p < 1/2. \end{cases}$$

← The alternative hypothesis corresponds to "having head more often than tail". If we can reject Ho, we will "prove" our point.

Let's consider first the simple hypothesis.

$H_0: p = 1/2.$
$H_1: p = p_1 < 1/2.$

The Neyman-Pearson lemma tells us that the most powerful test of level α rejects Ho if.

$$\frac{L(1/2)}{L(p_1)} < c.$$

$$\Leftrightarrow \frac{\prod_{i=1}^{30} (1/2)^{x_i - 1}\, 1/2}{\prod_{i=1}^{30} (1-p_1)^{x_i-1}\, p_1} < c.$$

$$\Leftrightarrow \left(\frac{1/2}{1-p_1}\right)^{\sum_{i=1}^{30} x_i - 30} \left(\frac{1/2}{p_1}\right)^{30} < c$$

$$\Leftrightarrow \left(\frac{1/2}{1-p_1}\right)^{\sum_{i=1}^{30} x_i} < c'$$

A decreasing function of $\sum_{i=1}^{n} x_i$ since $\frac{1/2}{1-p_1} < 1$.

$$\Leftrightarrow \sum_{i=1}^{30} x_i > c''$$

Since the likelihood ratio is monotone, this test will also be the uniformly most powerful test for the hypothesis (*), i.e.

Reject Ho iff $\sum_{i=1}^{30} x_i > c''$ →

Question 3 (continued)

b) By the central limit theorem, $\sum_{i=1}^{30} X_i \approx N\left(30 \cdot \frac{1}{p}, 30 \frac{1-p}{p^2}\right).$

$$\alpha = \max_{p \leq 1/2} P\left(\sum_{i=1}^{30} X_i > c'' \,\middle|\, p\right)$$

$$= P\left(\sum_{i=1}^{30} X_i > c'' \,\middle|\, p = 1/2\right)$$ since the likelihood ratio is monotone.

$$= P\left(\underbrace{\frac{\sum_{i=1}^{30} X_i - 60}{\sqrt{60}}}_{\sim N(0,1).} > \frac{c'' - 60}{\sqrt{60}} \,\middle|\, p = \frac{1}{2}\right)$$

$$= 1 - \Phi\left(\frac{c''-60}{\sqrt{60}}\right)$$

$$\Rightarrow \Phi\left(\frac{c''-60}{\sqrt{60}}\right) = 1 - \alpha \quad \Rightarrow \quad \frac{c''-60}{\sqrt{60}} = z_{1-\alpha}$$

$$\Rightarrow c'' = 60 + \sqrt{60}\, z_{1-\alpha}.$$

Therefore, the UMP test of level α rejects H_0 iff. $\sum_{i=1}^{30} X_i > 60 + \sqrt{60}\, z_{1-\alpha}$.

c) If $\alpha = 0.05$, the the test rejects H_0 iff $\sum_{i=1}^{30} X_i > 60 + \sqrt{60}\; 1{,}645 = 72.74$.

$$P(\text{reject } H_0 \mid p = 0.45) = P\left(\sum_{i=1}^{30} X_i > 72.74 \,\middle|\, p = 0.45\right)$$

"detect H_0 is false"

$$= P\left(\frac{\sum_{i=1}^{30} X_i - 30 \cdot \frac{1}{0.45}}{\sqrt{30 \frac{0.55}{0.45^2}}} > \frac{72.74 - 30/0.45}{\sqrt{30 \cdot \frac{0.55}{0.45^2}}} \,\middle|\, p = 0.45\right) = 1 - \Phi(0.67) = 0{,}2514$$

$\sim N(0,1)$

That probability is the power of the test when $p = 0{,}45$.

Question 4

In a casino, you observe a table where the game involves tossing a die. You suspect that the die of the dealer may be tricked. You record the result of each game. For game i, n_i is the number of times the die was tossed and X_i the number of times a 6 appeared. You recorded N games in all.

a) [5 pts] What is the distribution of X_i?

b) [10 pts] Under the model above, the MLE of $p = P(\text{get a } 6)$ is $\hat{p} = \sum_{i=1}^N X_i / \sum_{i=1}^N n_i$. Find the generalized likelihood ratio test of level $\alpha = 5\%$ for $H_0 : p = 1/6$ against the alternative $H_1 : p \neq 1/6$.

a) $X_i \sim \text{Bin}(n_i, p)$, $i = 1, \dots, N$. and $p = P(\text{get a } 6)$.

b) Consider the hypothesis

H_0: $p = 1/6$.

H_1: $p \neq 1/6$.

The GLRT rejects Ho iff.

$$\Lambda = \frac{L(1/6)}{\sup_{p \in [0,1]} L(p)} < c.$$

The max at the denominator is achieved when $p = \hat{p}$, the MLE. We do not need to calculate it here since it is given:

$$\hat{p} = \sum_{i=1}^N x_i \Big/ \sum_{i=1}^N n_i$$

Therefore,

$$\Lambda = \frac{\prod_{i=1}^N \binom{n_i}{x_i} (1/6)^{x_i} (5/6)^{n_i - x_i}}{\prod_{i=1}^N \binom{n_i}{x_i} \hat{p}^{x_i} (1-\hat{p})^{n_i - x_i}}$$

$$= \left(\frac{1/6}{\hat{p}}\right)^{\sum_{i=1}^N x_i} \left(\frac{5/6}{1-\hat{p}}\right)^{\sum_{i=1}^N (n_i - x_i)}$$

Moreover, under Ho, $-2 \log \Lambda \sim \chi^2_1$

Thus,

$$\alpha = P(\text{reject Ho} \mid \text{Ho true})$$
$$= P(\Lambda < c \mid \text{Ho true}).$$
$$= P(-2\log \Lambda > -2 \log c \mid \text{Ho true})$$

iff $-2 \log c = \chi^2_{1-\alpha, 1} = 3.84$ for $\alpha = 0.05$.

$c = 0.1466$.

The GLRT rejects Ho iff.

$$\left(\frac{1/6.}{\hat{p}}\right)^{\sum_{i=1}^N x_i} \left(\frac{5/6}{1-\hat{p}}\right)^{\sum_{i=1}^N (n_i - x_i)} < 0.1466.$$

where $\hat{p} = \sum_{i=1}^N x_i \Big/ \sum_{i=1}^N n_i$

Question 5

Consider the following discrete joint distribution

	$y=0$	1	2	
$x=0$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{2}$
1	$\frac{1}{9}$	$\frac{1}{6}$	$\frac{1}{18}$	$\frac{1}{3}$
2	$\frac{1}{12}$	$\frac{1}{12}$	0	$\frac{1}{6}$
	$\frac{13}{36}$	$\frac{1}{2}$	$\frac{5}{36}$	1

a) [10 pts] Calculate the correlation between X and Y. Note that $\text{var}(X) = 5/9$ and $\text{var}(Y) = 73/162$ (you do not need to compute them).

b) [5 pts] What is the conditional distribution of X given that $Y = 2$? Give its mass function.

a) Let's calculate the moments we need. Remember that $E(g(X)) = \sum_x g(x) P(X=x)$.

$$E(X) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{3} + 2 \cdot \frac{1}{6} = \frac{2}{3}.$$

$$E(Y) = 0 \cdot \frac{13}{36} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{5}{36} = \frac{7}{9}$$

$$E(XY) = \sum_{x,y} xy \cdot P(X=x, Y=y)$$

$$= (\text{zeros}) + 1 \cdot 1 \cdot \frac{1}{6} + 1 \cdot 2 \cdot \frac{1}{18} + 2 \cdot 1 \cdot \frac{1}{12}. = \frac{4}{9}$$

$$\text{cov}(X,Y) = E(XY) - E(X)E(Y) = \frac{4}{9} - \frac{2}{3} \cdot \frac{7}{9} = \frac{-2}{27}.$$

$$\text{cor}(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{\text{var}(X)\,\text{var}(Y)}} = \frac{-2/27}{\sqrt{5/9 \cdot 73/162}} = \frac{-2\sqrt{2}}{\sqrt{365}} = -0.148.$$

b) $$P(X=x \mid Y=2) = \frac{P(X=x, Y=2)}{P(Y=2)},$$

For instance, for $x = 0$.

$$P(X=0 \mid Y=2) = \frac{1/12}{5/36} = 3/5$$

x	0	1	2
$P(X=x \mid Y=2)$	$3/5$	$2/5$	0

Question 6

Answer the following two questions by a short paragraph.

a) [5 pts] Describe one possible use of bootstrap.

b) [5 pts] Explain why randomization is important in a controlled trial. Use an example to illustrate your explanation.

a) Bootstrap can be used to determine the variance of an estimator.

Bootstrap can be used to perform a test comparing two populations.

b) Randomization is important in a controlled trial to ensure groups are comparable. If the groups are not randomized, than the success / failure of a treatment could be due to the choice of the groups rather than the effect of the treatment itself.

For example, a sailman made a test on a pill for sea sickness. the results were amazing, none of those taking the pill were sick, but many not taking it were sea sick. However, the pill had been given to the crew (which we may expect does not tend to be sea sick). and the passengers were used as the control group... The effect cannot be associated to the pill itself (it may, but we cannot see it from that design).

Question 7

You studied the lifetime of electronic components, but since your study can not last forever, some of the components were still in working order when you had to stop your data collection. Among the n components you studied, m failed at time $X_i, i = 1, \dots, m$ (in hours) and the $n - m$ others did last for the 10000 hours that the study lasted, so $X_{m+1}, \dots, X_n$ are all greater than 10000, but could not be observed. The exponential distribution is a good model for the lifetime of these components. You would like to estimate the parameter λ of that exponential.

a) [5 pts] Find the survival function of of X, $S(x) = P(X > x)$, where X follows an exponential with mean $1/\lambda$.

b) [10 pts] Find the maximum likelihood estimate of λ from the sample above. Because you did not observe the exact value of each variable, the likelihood is

$$L(\lambda) = \left\{\prod_{i=1}^{m} f(Y_i)\right\}\left\{\prod_{i=m+1}^{n} S(Y_i)\right\}$$

where $Y_i = \min(X_i, 10000)$ are the values you actually observed.

a) $S(x) = P(X > x) = \int_x^\infty \lambda e^{-\lambda t}\, dt = e^{-\lambda x}.$

b) The likelihood is slightly different from what we are used to see, but the idea remains the same: find the value of λ that maximizes it!

$$L(\lambda) = \left\{\prod_{i=1}^{m} \lambda e^{-\lambda y_i}\right\}\left\{\prod_{i=m+1}^{n} e^{-\lambda y_i}\right\}.$$

$$= \lambda^m e^{-\lambda \sum_{i=1}^{n} y_i}$$

$$\ell(\lambda) = m \log \lambda - \lambda \sum_{i=1}^{n} y_i$$

$$\ell'(\lambda) = \frac{m}{\lambda} - \sum_{i=1}^{n} y_i = 0 \quad \text{iff } \lambda = \frac{m}{\sum_{i=1}^{n} y_i}$$

$$\ell''(\lambda) = \frac{-m}{\lambda^2} < 0 \Rightarrow \text{it is a max.}$$

Therefore,

$$\boxed{\hat{\lambda} = \frac{m}{\sum_{i=1}^{n} y_i}}$$