MIT Open Course 18.443. *Statistics for Applications*, Fall 2003

Solution of Homework Problems

Yan Zeng

Version 1.0.1, last revised on 2008-04-03.

Abstract

This is a solution manual for homework problems of MIT OpenCourse: *18.443. Statistics for Applications, Fall 2003*, by Dmitry Panchenko.[1].

Contents

[1] The lecture notes is available at http://ocw.mit.edu/OcwWeb/Mathematics/18-443Fall2003/CourseHome/index.htm

1 Problem Set 1: up to Lecture 2

▶ 1. Prove that

$$\lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-1}=\frac{\lambda^k}{k!}e^{-\lambda}$$

If you can't find easy proof, you can use Stirling's formula:

$$\lim_{n\to\infty}\left(\frac{n}{e}\right)^n\sqrt{2\pi n}/n!=1.$$

Proof. Using the fact that $\lim_{\alpha\to 0}(1+\alpha)^{1/\alpha}=e$, we have

$$\begin{aligned}
\lim_{n\to\infty}\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-1} &= \lim_{n\to\infty}\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\\
&= \lim_{n\to\infty}\frac{\lambda^k\left(1-\frac{\lambda}{n}\right)^n}{k!}\sqrt{\frac{n}{n-k}}\left(\frac{n}{n-k}\right)^{n-k}\frac{1}{e^k}\frac{1}{\left(1-\frac{\lambda}{n}\right)^k}\\
&= \frac{\lambda^k e^{-\lambda}}{k!}\cdot 1\cdot\lim_{n\to\infty}\left(1+\frac{k}{n-k}\right)^{n-k}\cdot\frac{1}{e^k}\cdot 1\\
&= \frac{\lambda^k e^{-\lambda}}{k!}.
\end{aligned}$$

□

▶ 2. Compute EX, EX^2 and $Var(X)$ for $X\sim N(\alpha,\sigma^2)$, $B(p)$, $E(\alpha)$, $\Pi(\lambda)$, $U(0,\theta)$.

Solution. For $X\sim N(\alpha,\sigma^2)$, by the calculation done in the notes ($E\left(\frac{X-\alpha}{\sigma}\right)=0$ and $E\left(\frac{X-\alpha}{\sigma}\right)^2=1$), we have $EX=\alpha$, $Var(X)=\sigma^2$, and $EX^2=(EX)^2+Var(X)=\alpha^2+\sigma^2$.

For $X\sim B(p)$, $EX=1\cdot P(X=1)=p$, $EX^2=1^2\cdot P(X=1)=p$, and $Var(X)=EX^2-(EX)^2=p-p^2$.

For $X\sim E(\alpha)$,

$$EX=\int_0^\infty x\alpha e^{-\alpha x}dx\overset{y=\alpha x}{=}\frac{1}{\alpha}\int_0^\infty ye^{-y}dy=\frac{1}{\alpha}\Gamma(2)=\frac{1}{\alpha},$$

$EX^2=\int_0^\infty x^2\alpha e^{-\alpha x}dx=\frac{1}{\alpha^2}\int_0^\infty y^2e^{-y}dy=\frac{1}{\alpha^2}\Gamma(3)=\frac{2}{\alpha^2}$, and $Var(X)=EX^2-(EX)^2=\frac{1}{\alpha^2}$.

For $X\sim\Pi(\lambda)$, $EX=\sum_{k=1}^\infty\frac{\lambda^k}{(k-1)!}e^{-\lambda}=\lambda\sum_{k=0}^\infty\frac{\lambda^k}{k!}e^{-\lambda}=\lambda$,

$$EX^2=\sum_{k=1}^\infty\frac{k\lambda^k}{(k-1)!}e^{-\lambda}=\sum_{k=1}^\infty\frac{(k-1)\lambda^k}{(k-1)!}e^{-\lambda}+\sum_{k=1}^\infty\frac{\lambda^k}{(k-1)!}e^{-\lambda}=\lambda^2\sum_{k=0}^\infty\frac{\lambda^k}{k!}e^{-\lambda}+EX=\lambda^2+\lambda,$$

and $Var(X)=EX^2-(EX)^2=\lambda+\lambda^2-\lambda^2=\lambda$.

For $X\sim U(0,\theta)$, $EX=\int_0^\theta\frac{x}{\theta}dx=\frac{\theta}{2}$, $EX^2=\int_0^\theta\frac{x^2}{\theta}dx=\frac{\theta^2}{3}$, and $Var(X)=\frac{\theta^2}{3}-\frac{\theta^2}{4}=\frac{\theta^2}{12}$. □

2 Problem Set 2: up to Lecture 4

▶ 1. Use first and second moments in the method of moments to find an estimate of θ in the uniform distribution $U[0,\theta]$ with p.d.f.

$$f(x|\theta)=\begin{cases}\frac{1}{\theta} & \text{when } 0\le x\le\theta\\ 0 & \text{otherwise,}\end{cases}$$

where $0<\theta<\infty$. Compare these two estimates by computing their asymptotic variances.

Solution. $m_1(\theta) = E_\theta X = \frac{\theta}{2}$ and $m_2(\theta) = E_\theta X^2 = \frac{\theta^2}{3}$. So the estimate obtained by first moment is

$$\widehat{\theta}_1 = m_1^{-1}(\bar{X}) = 2\bar{X},$$

and the estimate obtained by second moment is

$$\widehat{\theta}_2 = m_2^{-1}(\bar{X^2}) = \sqrt{3\bar{X^2}}.$$

The asymptotic variance corresponding to $\widehat{\theta}_1$ is (using Exercise 2 of Problem Set 1)

$$\frac{Var_{\theta_0}(X)}{(m_1'(\theta_0))^2} = \frac{\frac{\theta_0^2}{12}}{\left(\frac{1}{2}\right)^2} = \frac{\theta_0^2}{2}.$$

To find the asymptotic variance corresponding to $\widehat{\theta}_2$, we need to calculate $E_{\theta_0}X^4$. The moment generation function of X is

$$E_{\theta_0}e^{tX} = \int_0^{\theta_0} \frac{1}{\theta_0} e^{tx}dx = \frac{e^{t\theta_0}-1}{t\theta_0} = \sum_{k=0}^{\infty} \frac{(t\theta_0)^k}{(k+1)!}.$$

Since $E_{\theta_0}e^{tX} = \sum_{k=0}^{\infty} \frac{t^k}{k!} E_{\theta_0}X^k$, we have $E_{\theta_0}X^k = \frac{\theta_0^k}{k+1}$. So $E_{\theta_0}X^4 = \frac{\theta_0^4}{5}$ and

$$\frac{Var_{\theta_0}(X^2)}{(m_2'(\theta_0))^2} = \frac{\frac{\theta_0^4}{5} - \frac{\theta_0^2}{3}}{\frac{4}{9}\theta_0^2} = \frac{9}{20}\theta_0^2 - \frac{3}{4}.$$

So the first estimator is better if and only if $\theta_0 > \sqrt{\frac{45}{7}}$. $\square$

► 2. Find MLE of the parameter λ for Poisson distribution $\Pi(\lambda)$.

Solution. The likelihood function is

$$\varphi(\lambda) = \prod_{i=1}^n p(X_i|\lambda) = \frac{\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i!} e^{-\lambda}.$$

So

$$\frac{d}{d\lambda}\ln\varphi(\lambda) = \frac{d}{d\lambda}\left(-\lambda + \ln\lambda\sum_{i=1}^n X_i - \ln\prod_{i=1}^n X_i!\right) = -1 + \frac{\sum_{i=1}^n X_i}{\lambda}.$$

When $\lambda < \sum_{i=1}^n X_i$, $\ln\varphi(x)$ is increasing; when $\lambda > \sum_{i=1}^n X_i$, $\ln\varphi(x)$ is decreasing. So $\ln\varphi(\lambda)$ achieves its maximum at the critical point $\widehat{\lambda} = \sum_{i=1}^n X_i$. So the MLE of λ for $\Pi(\lambda)$ is $\widehat{\lambda} = \sum_{i=1}^n X_i$. $\square$

► 3. Find MLE of the parameter α for exponential distribution $E(\alpha)$.

Solution. The likelihood function is

$$\varphi(\alpha) = \prod_{i=1}^n p(X_i|\alpha)1_{[0,\infty)}(X_i) = \alpha^n e^{-\alpha\sum_{i=1}^n X_i} 1_{[0,\infty)}(Y),$$

where $Y = \min\{X_1, \cdots, X_n\}$. $\varphi(\alpha)$ is either zero or positive. When $\varphi(\alpha)$ is positive, we have

$$\frac{d}{d\alpha}\ln\varphi(\alpha) = \frac{d}{d\alpha}(n\ln\alpha - \alpha\sum_{i=1}^n X_i) = \frac{n}{\alpha} - \sum_{i=1}^n X_i$$

and

$$\frac{d^2}{d\alpha^2}\ln\varphi(\alpha) = -\frac{n}{\alpha^2} < 0.$$

So $\ln\varphi(\alpha)$ achieves its maximum at the critical point $\alpha = \frac{n}{\sum_{i=1}^n X_i} = \frac{1}{\bar{X}}$. So the MLE of α for $E(\alpha)$ is $\widehat{\alpha} = 1/\bar{X}$. $\square$

► 4. Consider a parametric family of distributions with the p.d.f. given by

$$f(x|\theta) = \begin{cases} e^{\theta - x} & \text{when } x \geq \theta \\ 0 & \text{when } x < \theta, \end{cases}$$

and where $-\infty < \theta < \infty$. Find MLE of θ for this family.

Solution. The likelihood function is

$$\varphi(\theta) = e^{n\theta - \sum_{i=1}^n X_i} 1_{[\theta,\infty)(Y)} = \begin{cases} e^{n\theta - \sum_{i=1}^n X_i} & \text{if } \theta \leq Y \\ 0 & \text{if } \theta > Y \end{cases},$$

where $Y = \min\{X_1, \cdots, X_n\}$. Since $e^{n\theta - \sum_{i=1}^n X_i}$ is an increasing function of θ, $\varphi(\theta)$ is maximized at $\widehat{\theta} = Y$. So, the MLE of θ for this family is $\widehat{\theta} = \min\{X_1, \cdots, X_n\}$. □

► 5. Consider a parametric family of distributions with the p.d.f. given by

$$f(x|\theta) = \frac{1}{2} e^{-|x-\theta|} \text{ for } -\infty < x < \infty,$$

where $-\infty < \theta < \infty$. Find MLE of θ for this family. (Hint: see Theorem 4.5.1 in the book [1].)

Solution. The likelihood function is

$$\varphi(\theta) = \frac{1}{2^n} e^{-\sum_{i=1}^n |X_i - \theta|}.$$

To remove the absolute value, we use Theorem 4.5.1 in the book [1]:

Theorem 4.5.1 *Let m be a median of the distribution of X, and let d be any other number. Then*

$$E(|X - m|) \leq E(|X - d|).$$

Furthermore, there will be quality in the relation if and only if d is also a median of the distribution of X.

For any given sequence $x_1, \cdots, x_n$, Theorem 4.5.1 implies

$$\frac{\sum_{i=1}^n |x_i - m|}{n} \leq \frac{\sum_{i=1}^n |x_i - d|}{n},$$

where m is a median of $\{x_1, \cdots, x_n\}$ and d is any other number. The equality holds if and only if d is also a median of $\{x_1, \cdots, x_n\}$. This shows $\varphi(\theta)$ is maximized when θ is a median of the observed values of X_1, $\cdots$, X_n. So, the MLE of θ is any median of X_1, $\cdots$, X_n (sample median).

□

3 Problem Set 3: up to Lecture 7

► 1. Compute Fisher information $I(\lambda)$ for a random variable with Poisson distribution $\Pi(\lambda)$.

Solution. $f(x|\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}$, so $l(x|\lambda) = -\lambda + x \log \lambda - \log x!$ and $l''(x|\lambda) = -\frac{x}{\lambda^2}$. This gives Fisher information $I(\lambda) = -E_\lambda l''(X|\lambda) = \frac{E_\lambda X}{\lambda^2} = \frac{1}{\lambda}$. □

► 2. Compute Fisher information $I(\alpha)$ for a random variable with normal distribution $N(\alpha, \sigma^2)$, assuming that σ^2 is a known constant (this means that α is the only parameter of the distribution).

Solution. $l(x|\alpha) = \log \frac{e^{-\frac{(x-\alpha)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} = -\frac{(x-\alpha)^2}{2\sigma^2} - \log(\sqrt{2\pi}\sigma)$. So $l''(x|\alpha) = -\frac{1}{\sigma^2}$, which gives Fisher information $I(\alpha) = \frac{1}{\sigma^2}$. □

► 3. Show that Bernoulli distribution $B(p)$ is an exponential-type distribution. Using this fact, find an efficient estimate of p.

Proof. The Bernoulli distribution can be written as

$$f(x|p) = p^x(1-p)^{1-x} = e^{x \log \frac{p}{1-p}}(1-p) = a(p)b(x)e^{c(p)d(x)},$$

where $a(p) = 1-p$, $b(x) = 1$, $c(p) = \log \frac{p}{1-p}$ and $d(x) = x$. Let $S = \frac{1}{n}\sum_{i=1}^n d(X_i) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$. We have $E_p S = p$. So $\bar{X}$ is an efficient estimate of p. □

► 4. Show that normal distribution $N(\alpha, \sigma^2)$ with given σ^2 (this means you can assume that σ^2 is a known constant, and α is the only parameter of the distribution) is an exponential-type distribution. Using this fact, find an efficient estimate of α.

Proof. $f(x|\alpha) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{x^2}{2\sigma^2}}e^{-\frac{\alpha^2}{2\sigma^2}}e^{\frac{\alpha x}{\sigma^2}}$. So we can express $f(x|\alpha)$ in the standard form of exponential-type distribution, with $a(\alpha) = e^{-\frac{\alpha^2}{2\sigma^2}}$, $b(x) = \frac{e^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}$, $c(\alpha) = \frac{\alpha}{\sigma^2}$, and $d(x) = x$. Let $S = \frac{1}{n}\sum_{i=1}^n d(X_i) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$. Then $E_\alpha S = \alpha$. So $\bar{X}$ is an efficient estimate of α. □

4 Problem Set 4: up to Lecture 10

► 1. ([1], §5.10, No. 9) A manufacturer believes that defective products are produced with unknown probability p, which will be modeled as having a Beta distribution. The manufacturer thinks that p should be around 0.05, but if the first 10 observed products were all defective, the mean of p would rise to 0.9. Find the Beta distribution that has these properties.

Solution. The problem can be formally formulated as the following: There is an i.i.d. sequence of 10 Bernoulli random variables $X_1, \cdots, X_{10}$. The parameter p of their common distribution is believed to follow a prior distribution $\xi(p)$, which is a Beta function $B(\alpha, \beta)$ with $E[p] = \frac{\alpha}{\alpha+\beta} = 0.05$. The expectation of the posterior satisfies $E[p|X_1 = 1, \cdots, X_{10} = 1] = 0.9$. *Determine α and β.*

To solve this problem, we note the posterior distribution of p satisfies

$$\begin{aligned}
& \xi(x|X_1 = x_1, \cdots, X_{10} = x_{10}) \\
= & P(p = x|X_1 = x_1, \cdots, X_{10} = x_{10}) \\
= & \frac{P(X_1 = x_1, \cdots, X_{10} = x_{10}|p = x)P(p = x)}{\int P(X_1 = x_1, \cdots, X_{10} = x_{10}|p = y)P(p = y)dy} \\
= & Cx^{\sum_{i=1}^{10} x_i}(1-x)^{\sum_{i=1}^{10}(1-x_i)}x^{\alpha-1}(1-x)^{\beta-1} \\
= & Cx^{\sum_{i=1}^{10} x_i + \alpha - 1}(1-x)^{\beta + \sum_{i=1}^{10}(1-x_i)-1}£¬
\end{aligned}$$

where C is a constant so that $\xi(p = x|X_1 = x_1, \cdots, X_{10} = x_{10})$ is a probability density function in x. So it's easy to see

$$C = \frac{\Gamma(10+\alpha+\beta)}{\Gamma(\sum_{i=1}^{10} x_i + \alpha)\Gamma(\sum_{i=1}^{10}(1-x_i)+\beta)},$$

and

$$\begin{aligned}
E[p|X_1 = 1, \cdots, X_{10} = 1] &= \int_0^1 xP(p = x|X_1 = \cdots = X_{10} = 1)dx \\
&= \int_0^1 x\frac{\Gamma(10+\alpha+\beta)}{\Gamma(\alpha)\Gamma(10+\beta)}x^{\alpha-1}(1-x)^{10+\beta-1}dx \\
&= \frac{\alpha+10}{\alpha+\beta+10}.
\end{aligned}$$

Solving the equations

$$\begin{cases} \frac{\alpha}{\alpha+\beta} = 0.05 \\ \frac{\alpha+10}{\alpha+\beta+10} = 0.9, \end{cases}$$

we conclude $\alpha = \frac{1}{17}$ and $\beta = \frac{19}{17}$. □

► 2. ([1], §6.2, No. 2) Suppose the proportion p of defective items in a large manufactured lot is known to be either 0.1 or 0.2, and the prior p.f. of p is as follows:

$$\xi(0.1) = 0.7 \text{ and } \xi(0.2) = 0.3.$$

Suppose also that when eight items are selected at random from the lot, it is found that exactly two of them are defective. Determine the posterior p.f. of p.

Solution.

$$\begin{aligned}
& \xi(0.1|X_1 = x_1, \cdots, X_n = x_n) \\
= & P(p = 0.1|X_1 = x_1, \cdots, X_n = x_n) \\
= & \frac{P(X_1 = x_1, \cdots, X_n = x_n|p = 0.1)P(p = 0.1)}{P(X_1 = x_1, \cdots, X_n = x_n|p = 0.1)P(p = 0.1) + P(X_1 = x_1, \cdots, X_n = x_n|p = 0.2)P(p = 0.2)} \\
= & \frac{0.1^2 \cdot 0.9^6 \cdot 0.7}{0.1^2 \cdot 0.9^6 \cdot 0.7 + 0.2^2 \cdot 0.8^6 \cdot 0.3} \\
\approx & \ 0.5418.
\end{aligned}$$

Accordingly, $\xi(0.2|X_1 = x_1, \cdots, X_n = x_n) \approx 1 - 0.5418 = 0.4582$. □

► 3. ([1], §6.2, No. 4) Suppose that the prior of some parameter θ is a Gamma distribution for which the mean is 10 and the variance is 5. Determine the prior p.d.f. of θ.

Solution. If $\theta \sim \Gamma(\alpha, \beta)$, by the calculation in Lecture 8, we have $E[\theta] = \frac{\alpha}{\beta^2}$ and $Var(\theta) = \frac{\alpha}{\beta^2}$. Solving the equation

$$\begin{cases} \frac{\alpha}{\beta} = 10 \\ \frac{\alpha}{\beta^2} = 5, \end{cases}$$

we conclude $\alpha = 20$ and $\beta = 2$. □

► 4. ([1], §6.3, No. 12) Suppose that the time in minutes required to serve a customer at a certain facility has exponential distribution $E(\alpha)$ with α unknown and the prior of α is a Gamma distribution with mean 0.2 and the standard deviation 1. If the average time required to serve a sample of 20 customers is 3.8 minutes, what is the posterior distribution of α.

Solution. Denote by $\xi(x)$ the prior p.d.f. of α, which follows a Gamma distribution $\Gamma(a, b)$. By the calculation in Lecture 8 and given condition, we have $E[\alpha] = \frac{a}{b} = 0.2$ and $\sqrt{Var(\alpha)} = \frac{\sqrt{a}}{b} = 1$. This implies $a = \frac{1}{25}$ and $b = \frac{1}{5}$. We note

$$\begin{aligned}
\xi(x|X_1 = x_1, \cdots, X_n = x_n) &= \frac{P(X_1 = x_1, \cdots, X_n = x_n|\alpha = x)P(\alpha = x)}{\int P(X_1 = x_1, \cdots, X_n = x_n|\alpha = y)P(\alpha = y)dy} \\
&= \frac{x^n e^{-x\sum_{i=1}^n x_i} \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}}{Constant} \\
&= \frac{(b + \sum_{i=1}^n x_i)^{n+a}}{\Gamma(n+a)} x^{n+a-1} e^{-(b+\sum_{i=1}^n x_i)x} \\
&= \frac{(b + n\bar{X})^{n+a}}{\Gamma(n+a)} x^{n+a-1} e^{-(b+n\bar{X})x}.
\end{aligned}$$

Plug in $n = 20$, $\bar{X} = 3.8$, $a = \frac{1}{25}$, and $b = \frac{1}{5}$, we can have the posterior of α. □

► 5. ([1], §6.4, No. 2) Suppose that a proportion p of defective items in a large shipment is unknown, and the prior of p is Beta distribution with the parameters $\alpha = 5$, $\beta = 10$. Suppose that 20 items are selected at random from the shipment, and that exactly one of these items is found to be defective. If the squared error loss function is used, what is the Bayes estimate of p?

Solution. By the calculation in Lecture 10, the posterior p.d.f. of p is

$$\xi(x|X_1, \cdots, X_n) = \frac{\Gamma(\alpha+\beta+n)}{\Gamma(\alpha+\sum_{i=1}^n X_i)\Gamma(\beta+n-\sum_{i=1}^n X_i)} x^{(\alpha+\sum_{i=1}^n X_i)-1}(1-x)^{(\beta+n-\sum_{i=1}^n X_i)-1}.$$

If the squared error loss function is used, the Bayes estimate of p is its expectation under the posterior p.d.f. So

$$\widehat{p} = E[p|X_1, \cdots, X_n] = \int_0^1 x \frac{\Gamma(35)x^5(1-x)^{28}}{\Gamma(6)\Gamma(29)} dx = \frac{6}{35}.$$

□

► 6. ([1], §6.4, No. 5) Suppose that the number of defects in a roll of magnetic recording tape has Poisson distribution $\Pi(\lambda)$ with λ unknown, and the prior of λ is a Gamma distribution with parameters $\alpha = 3$, $\beta = 1$. On five randomly selected rolls the numbers of defects were to be 2, 2, 6, 0, 3. If the squared loss function is used, what is the Bayes estimate of λ?

Solution. The posterior p.d.f. of λ is

$$\xi(x|X_1, \cdots, X_n) = \frac{(\beta+1)^{\alpha+\sum_{i=1}^n X_i} x^{\alpha+\sum_{i=1}^n X_i - 1} e^{-(\beta+1)x}}{\Gamma(\alpha+\sum_{i=1}^n X_i)}.$$

So the Bayes estimate is

$$\widehat{\lambda} = E[\lambda|X_1, \cdots, X_n] = \int_0^\infty x \frac{2^{3+13}x^{3+13-1}e^{-2x}}{\Gamma(3+13)} dx = 8.$$

□

5 Problem Set 5: up to Lecture 13

► 1. Find sufficient statistic for Gamma distribution $\Gamma(\alpha, \beta)$, where the value of β is known and the value of α is unknown.

Solution. The joint p.d.f. of a Gamma distribution $\Gamma(\alpha, \beta)$ with known β is

$$f(x_1, \cdots, x_n|\alpha) = \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^n (x_1 \cdots x_n)^{\alpha-1} e^{-\beta \sum_{i=1}^n x_i} 1_{\{x_1 \geq 0, \cdots, x_n \geq 0\}} = u(x_1, \cdots, x_n) v(T(x_1, \cdots, x_n), \alpha),$$

where $u(x_1, \cdots, x_n) = e^{-\beta \sum_{i=1}^n x_i} 1_{\{x_1 \geq 0, \cdots, x_n \geq 0\}}$, $v(T, \alpha) = \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^n T^{\alpha-1}$ and $T(x_1, \cdots, x_n) = \prod_{i=1}^n x_i$. So $T(X_1, \cdots, X_n) = \prod_{i=1}^n X_i$ is a sufficient statistic by Neyman-Fisher factorization criterion. □

► 2. Consider a family of distribution with p.d.f. given by

$$f(x|\theta) = a(\theta)b(x)e^{c(\theta)d(x)}$$

- the so called exponential family of distributions. Given an i.i.d. sample X_1, $\cdots$, X_n with distribution from this family, find sufficient statistic.

Solution.

$$f(x_1, \cdots, x_n|\theta) = a(\theta)^n \prod_{i=1}^n b(x_i) e^{c(\theta)\sum_{i=1}^n d(x_i)} = u(x_1, \cdots, x_n) v(T(x_1, \cdots, x_n), \theta),$$

where $T(x_1, \cdots, x_n) = \sum_{i=1}^n d(x_i)$, $v(t, \theta) = a(\theta)^n e^{tc(\theta)}$, and $u(x_1, \cdots, x_n) = \prod_{i=1}^n b(x_i)$. So a sufficient statistic is $T(X_1, \cdots, X_n) = \sum_{i=1}^n d(X_i)$ by Neyman-Fisher factorization criterion. □

► 3. Find jointly sufficient statistics for Gamma distribution $\Gamma(\alpha, \beta)$, where now both β and α are unknown.

Solution.

$$\begin{aligned} f(x_1, \cdots, x_n|\alpha, \beta) &= \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^n (x_1 \cdots x_n)^{\alpha-1} e^{-\beta \sum_{i=1}^n x_i} 1_{\{x_1 \geq 0, \cdots, x_n \geq 0\}} \\ &= u(x_1, \cdots, x_n) v(T_1(x_1, \cdots, x_n), T_2(x_1, \cdots, x_n), \alpha, \beta), \end{aligned}$$

where $u(x_1, \cdots, x_n) = 1_{\{x_1 \geq 0, \cdots, x_n \geq 0\}}$, $v(t_1, t_2, \alpha, \beta) = \left[\frac{\beta^\alpha}{\Gamma(\alpha)}\right]^n t_1^{\alpha-1} e^{-\beta t_2}$, and $T_1(x_1, \cdots, x_n) = \prod_{i=1}^n x_i$, $T_2(x_1, \cdots, x_n) = \sum_{i=1}^n x_i$. So a jointly sufficient statistic is $(\prod_{i=1}^n X_i, \sum_{i=1}^n X_i)$ by Neyman-Fisher factorization criterion. □

► 4. Suppose that we have an i.i.d.d sample $X_1, \cdots, X_n$ from exponential distribution $E(\alpha)$ with α unknown. Is the MLE of α a minimal sufficient statistic?

Solution. The joint p.d.f. of exponential distribution $E(\alpha)$ is

$$f(x_1, \cdots, x_n|\alpha) = \alpha^n e^{-\alpha \sum_{i=1}^n x_i} 1_{\{x_1 \geq 0, \cdots, x_n \geq 0\}}.$$

If $\min\{x_1, \cdots, x_n\} < 0$, $f(x_1, \cdots, x_n|\alpha) \equiv 0$; if $\min\{x_1, \cdots, x_n\} \geq 0$, take derivative with respect to α and solve the equation $\frac{\partial}{\partial \alpha} f(x_1, \cdots, x_n|\alpha) = 0$, we get the MLE $\widehat{\alpha}$ of α:

$$\widehat{\alpha} = \frac{n}{\sum_{i=1}^n x_i} = \frac{1}{\bar{x}}.$$

If we let $T(x_1, \cdots, x_n) = \frac{1}{\bar{x}}$, $u(x_1, \cdots, x_n) = 1_{\{x_1 \geq 0, \cdots, x_n \geq 0\}}$, and $v(t, \alpha) = \alpha^n e^{-\frac{\alpha n}{t}}$, then $f(x_1, \cdots, x_n|\alpha)$ can be written as $u(x_1, \cdots, x_n) v(T(x_1, \cdots, x_n), \alpha)$. By Neyman-Fisher factorization criterion, $\widehat{\alpha}$ is a sufficient statistic. Hence, the MLE $\widehat{\alpha} = \frac{1}{\bar{X}}$ of α is a minimal sufficient statistic. □

► 5. Consider an i.i.d. sample $X_1, \cdots, X_n$ from uniform distribution $U[0, \theta]$ and let

$$Y_n = \max\{X_1, \cdots, X_n\}.$$

Find the estimate of unknown θ of the form cY_n that minimizes the squared error loss

$$E_\theta(cY_n - \theta)^2.$$

Solution. It is not clear what the problem means by "the estimate of unknown θ of the form cY_n". But we show some computation anyway.

It is easy to see Y_n has p.d.f. $nt^{n-1} 1_{\{0 \leq t \leq \theta\}}$. So

$$E_\theta(cY_n - \theta)^2 = \int_0^\theta (ct - \theta)^2 n t^{n-1} dt = \left(\frac{n}{n+2} c^2 - \frac{2n}{n+1} c + 1\right) \theta^{n+2}.$$

□

6 Problem Set 6: up to Lecture 19

► 1. [1] page 415, No. 7. In the June 1986 issue of *Consumer Reports*, some data on the calorie content of beef hot dog is given. Here are the numbers of calories in 20 different hot dog brands:

186,	181,	176,	149,	184,	190,	158,	139,	175,	148,
152,	111,	141,	153,	190,	157,	131,	149,	135,	132

Find 90% confidence intervals for the mean μ and the variance σ^2.

Solution. By the calculation of Lecture 17, we first calculate the sample mean and sample second moment: $\bar{X} = 156.85$ and $\bar{X^2} = 25088.95$. So $\bar{X^2} - (\bar{X})^2 = 487.03$. We have to find constant c_1, c_2 and c, such that $P(\chi^2_{19} \le c_1) = P(\chi^2_{19} \ge c_2) = \frac{1-\alpha}{2} = 0.05$ and $P(t_{19} \le -c) = P(t_{19} \ge c) = \frac{1-\alpha}{2} = 0.05$. Looking up the tables on page 774 and page 776 of [1], we get $c_1 = 6.844$, $c_2 = 30.14$ and $c = 1.729$. Therefore the 90% confidence interval for the variance σ^2 is

$$\left[\frac{n(\bar{X^2} - (\bar{X})^2)}{c_2}, \frac{n(\bar{X^2} - (\bar{X})^2)}{c_2}\right] = [323.18, 1423.22]$$

and the 90% confidence interval for the mean μ is

$$\left[\bar{X} - c\sqrt{\frac{1}{n-1}(\bar{X^2} - (\bar{X})^2)}, \bar{X} + c\sqrt{\frac{1}{n-1}(\bar{X^2} - (\bar{X})^2)}\right] = [148.10, 165.60].$$

□

► 2. [1] page 469 No. 2. Consider two p.d.f.'s $f_0(x)$ and $f_1(x)$ which are defined as follows:

$$f_0(x) = \begin{cases} 1 & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$f_1(x) = \begin{cases} 2x & \text{for } 0 \le x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Suppose that a single observation X is taken from a distribution for which the p.d.f. $f(x)$ is either $f_0(x)$ or $f_1(x)$, and the following simple hypotheses are to be tested:

$$H_0 : f(x) = f_0(x),$$
$$H_1 : f(x) = f_1(x).$$

a. Describe a test procedure for which the value of $\alpha(\delta) + 2\beta(\delta)$ is a minimum.
b. Determine the minimum value of $\alpha(\delta) + 2\beta(\delta)$ attained by that procedure.

Solution. We suppose the notation $\alpha(\delta)$ and $\beta(\delta)$ mean: $\alpha(\delta) = P_0(\delta \ne H_0)$ and $\beta(\delta) = P_1(\delta \ne H_1)$. If we let $\xi(0) = \frac{1}{3}$ and $\xi(1) = \frac{2}{3}$, Problem (a) is reduced to finding the Bayes decision ruls that minimizes the Bayes error $\xi(0)P_0(\delta \ne H_0) + \xi(1)P_1(\delta \ne H_1)$. By Lecture 19, the Bayes decision rule is given by (note $\frac{f_0(X)}{f_1(X)} = \frac{1}{2X}$ and $\frac{\xi(1)}{\xi(0)} = 2$)

$$\delta_* = \begin{cases} H_0 : & X < \frac{1}{4} \\ H_1 : & X > \frac{1}{4} \\ H_0 \text{ or } H_1 : & X = \frac{1}{4}. \end{cases}$$

To solve Problem (b), we set δ_* more precisely as

$$\delta_* = \begin{cases} H_0 : & X \le \frac{1}{4} \\ H_1 : & X > \frac{1}{4}. \end{cases}$$

Then

$$\alpha(\delta_*) + 2\beta(\delta_*) = P_0(\delta_* \ne H_0) + 2P_1(\delta_* \ne H_1) = P_0(X > \frac{1}{4}) + 2P_1(X \le \frac{1}{4}) = \int_{\frac{1}{4}}^1 dx + 2\int_0^{\frac{1}{4}} 2x dx = \frac{7}{8}.$$

□

▶ 3. [1] page 469 No. 4. Consider again the conditions of Exercise 2, but suppose now that it is desired to find a test procedure for which $\alpha(\delta) \le 0.1$ and $\beta(\delta)$ is minimum.
a. Describe the procedure.
b. Determine the minimum value of $\beta(\delta)$ attained by the procedure.

Solution. By the Theorem in Lecture 19, we need to find a constant c, such that

$$P_0\left(\frac{f_0(X)}{f_1(X)} < c\right) = 0.1.$$

Note

$$P_0\left(\frac{f_0(X)}{f_1(X)} < c\right) = P_0\left(\frac{1}{2X} < c\right) = P_0(X > \frac{1}{2c}) = 1_{\{c \ge \frac{1}{2}\}}(1 - \frac{1}{2c}) = 0.1.$$

So $c = \frac{5}{9}$. Then the decision rule

$$\delta = \begin{cases} H_0: & \frac{1}{2X} \ge \frac{5}{9} \\ H_1: & \frac{1}{2X} < \frac{5}{9} \end{cases}$$

is the rule for which $\alpha(\delta) \le 0.1$ and $\beta(\delta)$ is minimum. In this case,

$$\beta(\delta) = P_1(\delta \neq H_1) = P_1(X \le \frac{9}{10} = 0.81.$$

□

▶ 4. Consider two hypotheses, null hypothesis H_1 and alternative hypothesis H_2:

H_1: the distribution P is Bernoulli with probability of success $p = 0.2$,
H_2: the distribution P is Bernoulli with probability of success $p = 0.4$.

Given two observations X_1, X_2 construct the most powerful test of size $\alpha_1 = 0.05$. Compute the power of this test.

Solution. The joint p.d.f. of a Bernoulli random variable is $f(x_1, \cdots, x_n|p) = p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)}$. So $f_1(X_1, X_2) = 0.2^{X_1+X_2}0.8^{2-(X_1+X_2)}$, $f_2(X_1, X_2) = 0.4^{X_1+X_2}0.6^{2-(X_1+X_2)}$ and

$$\frac{f_1(X_1, X_2)}{f_2(X_1, X_2)} = 0.5^{X_1+X_2}\left(\frac{4}{3}\right)^{2-(X_1+X_2)}.$$

By the theorem in Lecture 19, we need to find c such that $P_1\left(\frac{f_1(X_1,X_2)}{f_2(X_1,X_2)} < c\right) = \alpha_1 = 0.05$. Note

$$P_1\left(\frac{f_1(X_1, X_2)}{f_2(X_1, X_2)} < c\right) = 1_{\{\frac{16}{9} < c\}}0.64 + 1_{\{\frac{1}{4} < c\}}0.04 + 1_{\{\frac{2}{3} < c\}}0.32.$$

So to achieve the error control 0.05, we must randomize the test (see Lecture 20): we need to find c and p such that

$$P_1\left(\frac{f_1(X_1, X_2)}{f_2(X_1, X_2)} < c\right) + (1-p)P_1\left(\frac{f_1(X_1, X_2)}{f_2(X_1, X_2)} = c\right) = 0.05.$$

The above equation implies that $P_1\left(\frac{f_1(X_1,X_2)}{f_2(X_1,X_2)} < c\right) \le 0.05$. So we must have $c \in (\frac{1}{4}, \frac{2}{3}]$. To make $P_1\left(\frac{f_1(X_1,X_2)}{f_2(X_1,X_2)} = c\right)$ non-zero, we must choose $c = \frac{2}{3}$. This leads to the equation

$$0.04 + (1-p)P_1(X_1 = 0, X_2 = 1 \text{ or } X_1 = 1, X_2 = 0) = 0.05,$$

which gives us $p = \frac{31}{32}$. Therefore the most powerful test of size $\alpha_1 = 0.05$ is

$$\delta = \begin{cases} H_1: & 0.5^{X_1+X_2}\left(\frac{4}{3}\right)^{2-(X_1+X_2)} > \frac{2}{3} \\ H_2: & 0.5^{X_1+X_2}\left(\frac{4}{3}\right)^{2-(X_1+X_2)} < \frac{2}{3} \\ H_1 \text{ or } H_2: & 0.5^{X_1+X_2}\left(\frac{4}{3}\right)^{2-(X_1+X_2)} = \frac{2}{3}, \end{cases}$$

where in the last case of equality we break the tie at random by choosing H_1 with probability $p = \frac{31}{32}$ and choosing H_2 with probability $1 - p = \frac{1}{32}$. In this case, the power of the test is

$$\begin{aligned} P_2(\delta = H_2) &= P_2\left(\frac{f_1(X_1, X_2)}{f_2(X_1, X_2)} < c\right) + (1-p)P_2\left(\frac{f_1(X_1, X_2)}{f_2(X_1, X_2)} = c\right) \\ &= P_2(X_1 = X_2 = 1) + \frac{1}{32}P_2(X_1 = 0, X_2 = 1 \text{ or } X_1 = 1, X_2 = 0) \\ &= 0.16 + \frac{1}{32}0.48 \\ &= 0.175. \end{aligned}$$

□

► 5. Suppose that we have ten observations $X_1, \cdots, X_{10}$ from a normal distribution for which the mean μ is unknown and the variance σ^2 is 4. Given two simple hypotheses

$$H_1: \mu = 1,\ H_2: \mu = -1$$

find the most powerful test δ with type 1 error $\alpha_1(\delta) = 0.05$. Compute the power of this test.

Solution. We apply the theorem in Lecture 19. First, we note

$$\frac{f_1(X_1, \cdots, X_{10})}{f_2(X_1, \cdots, X_{10})} = \exp\left\{-\frac{\sum_{i=1}^{10}[(X_i - \mu_1)^2 - (X_i - \mu_2)^2]}{2\sigma^2}\right\} = e^{\frac{\sum_{i=1}^{10} X_i}{2}}.$$

This implies

$$\begin{aligned} P_1\left(\frac{f_1(X_1, \cdots, X_{10}}{f_2(X_1, \cdots, X_{10})} < c\right) &= P_1(\sum_{i=1}^{10} X_i < 2\log c) \\ &= P_1\left(\frac{\sum_{i=1}^{10}(X_i - 1)}{\sqrt{10}\sigma} < \frac{2\log c - 10}{\sqrt{10}\sigma}\right) \\ &= \Phi\left(\frac{2\log c - 10}{2\sqrt{10}}\right). \end{aligned}$$

Solving $\Phi\left(\frac{2\log c - 10}{2\sqrt{10}}\right) = 0.05$ gives us $c = 0.818$. So the most powerful test δ with type 1 error 0.05 is given by

$$\delta = \begin{cases} H_1: & e^{\frac{\sum_{i=1}^{10} X_i}{2}} \geq 0.818 \\ H_2: & e^{\frac{\sum_{i=1}^{10} X_i}{2}} < 0.818. \end{cases}$$

Consequently, the power of the test is given by

$$P_2(\delta = H_2) = P_2\left(e^{\frac{\sum_{i=1}^{10} X_i}{2}} < c\right) = \Phi\left(\frac{2\log c + 10}{2\sqrt{10}}\right) = 0.94.$$

□

7 Problem Set 7: up to Lecture 22

In all problems take level of significance $\alpha = 0.05$.

► 1. [1] page 541, No. 4. According to a simple genetic principle, if both the mother and the father of a child have genotype Aa, then there is probability 1/4 that the child will have genotype AA, probability 1/2 that she will have genotype Aa, and probability 1/4 that she will have genotype aa. In a random sample of 24 children having both parents with genotype Aa, it is found that 10 have genotype AA, 10 have genotype Aa, and four have genotype aa. Investigate whether the simple genetic principle is correct by carrying out a χ^2 test of goodness-of-fit.

Solution. We define $p_1 = P(AA)$, $p_2 = P(Aa)$, and $p_3 = P(aa)$. Then the hypotheses under testing is

$$\begin{cases} H_1: & p_1 = p_1^0, p_2 = p_2^0, p_3 = p_3^0 \\ H_2: & \text{otherwise}, \end{cases}$$

where $p_1^0 = \frac{1}{2}, p_2^0 = p_3^0 = \frac{1}{4}$. Let $v_1 = \sum_{i=1}^n 1_{\{X_i=AA\}}$, $v_2 = \sum_{i=1}^n 1_{\{X_i=Aa\}}$, and $v_3 = \sum_{i=1}^n 1_{\{X_i=aa\}}$. Then

$$T = \sum_{k=1}^{3} \frac{(v_k - np_k^0)^2}{np_k^0} = \frac{11}{3} \approx 3.667.$$

The χ^2 test of goodness-of-fit is

$$\delta = \begin{cases} H_1: & T \leq c \\ H_2: & T > c, \end{cases}$$

where c is such that $0.05 = \alpha = \chi^2_{3-1}(c, \infty)$. Looking up table gives $c = 5.991$. So the simple genetic principle is accepted at level of significance 0.05. The following Matlab code facilitates the calculation of T:

```
function T = pearson_test(counts, probabilities)

%PEARSON_TEST  gives the statistic of chi-square test of goodness-of-fit for simple
%              hypotheses, based on Pearson's Theorem. See [1] Lecture 23.
%
%              [1] Dmitry Panchenko: 18.443. Statistics for Applications,
%              Fall 2003, MIT OpenCourseWare.
%
%              Yan Zeng, 07/29/2008.

numData = sum(counts);

T = sum( (counts - numData * probabilities).^2 ./ probabilities ) /
numData;

end
```

□

► 2. [1] page 542, No. 9(a). The 50 values in Table 9.3 are intended to be a random sample from a standard normal distribution.

-1.28	-1.22	-0.45	-0.35	0.72
-0.32	-0.80	-1.66	1.39	0.38
-1.38	-1.26	0.49	-0.14	-0.85
2.33	-0.34	-1.96	-0.64	-1.32
-1.14	0.64	3.44	-1.67	0.85
0.41	-0.01	0.67	-1.13	-0.41
-0.49	0.36	-1.24	-0.04	-0.11
1.05	0.04	0.76	0.61	-2.04
0.35	2.82	-0.46	-0.63	-1.61
0.64	0.56	-0.11	0.13	-1.81

a. Carry out a χ^2 test of goodness-of-fit by dividing the real line into five intervals, each of which has probability 0.2 under the standard normal distribution.

Solution. For $x_1 = -0.842$, $x_2 = -0.253$, $x_3 = 0.253$, and $x_4 = 0.842$, we have $\Phi(x_k) = 0.2k$ $(k = 1, 2, 3, 4)$. Define $B_1 = (-\infty, -0.842]$, $B_2 = (-0.842, -0.253]$, $B_3 = (-0.253, -.253]$, $B_4 = (0.253, 0.842]$, and $B_5 = (0.842, \infty)$, then for $p_k = P(X \in B_k)$ $(k = 1, 2, 3, 4, 5)$, the hypotheses under testing is

$$\begin{cases} H_1: & p_1 = p_2 = p_3 = p_4 = p_5 = 0.2 \\ H_2: & \text{otherwise.} \end{cases}$$

Define $v_k = \sum_{i=1}^n 1_{\{X_i \in B_k\}}$ $(k = 1, 2, 3, 4, 5)$. Then it is easy to count $v_1 = 15$, $v_2 = 10$, $v_3 = 7$, $v_4 = 12$, and $v_5 = 6$. [2] Using the Matlab program in the previous problem, we have

$$T = \sum_{k=1}^{5} \frac{(v_k - 50 \cdot 0.2)^2}{50 \cdot 0.2} = 5.4.$$

The χ^2 test of goodness-of-fit is

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c, \end{cases}$$

where c is such that $0.05 = \alpha = \chi^2_{5-1}(c, \infty)$. Looking up table gives $c = 9.488$. So the assumption that the samples are from a standard normal random variable is accepted at level of significance 0.05. □

► 3. [1] page 549, No. 3 (Read example at the bottom of page 543.) Consider a genetics problem in which each individual in a certain population must have one of six genotypes, and it is desired to test the null hypothesis H_0 that the probabilities of the six genotypes can be represented in the form specified in Eq. (9.2.2).
a. Supose that in a random sample of n individuals, the observed numbers of individuals having the six genotypes are $N_1, \cdots, N_6$. Find the M.L.E.'s of θ_1 and θ_2 when the null hypotheses H_0 is true.
b. Suppose that in a random sample of 150 individuals, the observed numbers are as follows:

$$N_1 = 2, N_2 = 36, N_3 = 14, N_4 = 36, N_5 = 20, N_6 = 42.$$

Determine the value of Q and the corresponding tail area.

Solution. (a) Under the null hypotheses H_0, the likelihood function is

$$f(\theta_1, \theta_2) = \theta_1^{2N_1}\theta_2^{2N_2}(1-\theta_1-\theta_2)^{2N_3}(2\theta_1\theta_2)^{N_4}[2\theta_1(1-\theta_1-\theta_2)]^{N_5}[2\theta_2(1-\theta_1-\theta_2)]^{N_6}.$$

[2] E.g. in Matlab, set A as the data matrix, and input sum(sum(A<=-0.842)). It gives the number of data points that are no greater than -0.842.

So the log-likelihood function is

$$\begin{aligned}\log f(\theta_1,\theta_2) &= 2N_1\log\theta_1 + 2N_2\log\theta_2 + 2N_3\log(1-\theta_1-\theta_2) + N_4(\log 2+\log\theta_1+\log\theta_2)\\ &\quad +N_5[\log 2+\log\theta_1+\log(1-\theta_1-\theta_2)] + N_6[\log 2+\log\theta_2+\log(1-\theta_1-\theta_2)].\end{aligned}$$

To find the maximum likelihood estimator, we calculate the partial derivatives of $\log f(\theta_1,\theta_2)$ with respect to θ_1 and θ_2:

$$\begin{cases}\frac{\partial}{\partial\theta_1}f(\theta_1,\theta_2) = \frac{2N_1+N_4+N_5}{\theta_1} - \frac{2N_3+N_5+N_6}{1-\theta_1-\theta_2}\\ \frac{\partial}{\partial\theta_2}f(\theta_1,\theta_2) = \frac{2N_2+N_4+N_6}{\theta_2} - \frac{2N_3+N_5+N_6}{1-\theta_1-\theta_2}.\end{cases}$$

Setting the above equations to zero, we can solve for θ_1 and θ_2.

(b) Plugging the values of N_i's into the equations for θ_1 and θ_2, we have $\theta_1 = 0.2$ and $\theta_2 = 0.5$. The corresponding statistic T is equal to (using the Matlab program provided in Problem 1)

$$\begin{aligned} & \frac{(N_1-150\cdot 0.2^2)^2}{150\cdot 0.2^2} + \frac{(N_2-150\cdot 0.5^2)^2}{150\cdot 0.5^2} + \frac{(N_3-150\cdot 0.3^2)^2}{150\cdot 0.3^2}\\ & +\frac{(N_4-150\cdot 0.2)^2}{150\cdot 0.2} + \frac{(N_5-150\cdot 0.12)^2}{150\cdot 0.12} + \frac{(N_6-150\cdot 0.3)^2}{150\cdot 0.3}\\ \approx\ & 4.37.\end{aligned}$$

By the theorem in Lecture 25, $T \to \chi^2_{6-2-1} = \chi^2_3$ as $n\to\infty$. For $c = 7.815$, $\chi^2_3(c,\infty) = 0.05$. So the decision rule that accepts the null hypotheses for $T \le c$ has level of significance 0.05. In our case, since $T = 4.37 < c = 7.815$, the null hypotheses is accepted at level of significance 0.05. □

8 Problem Set 8: up to Lecture 27

► 1. [1] page 549, No. 2. At the fifth hockey game of the season at a certain arena, 200 people were selected at random and asked how many of the previous four games they had attended. The results are given in the following table:

Number of games previously attended	Number of people
0	33
1	67
2	66
3	15
4	19

Test the hypothesis that these 200 observed values can be regarded as a random sample from a binomial distribution; that is, there exists a number θ ($0<\theta<1$) such that the probabilities are as follows:

$$p_0 = (1-\theta)^4, p_1 = 4\theta(1-\theta)^3, p_2 = 6\theta^2(1-\theta)^2, p_3 = 4\theta^3(1-\theta), p_4 = \theta^4.$$

Solution. The hypotheses under testing is

$$\begin{cases}H_1: & p_0 = (1-\theta)^4, p_1 = 4\theta(1-\theta)^3, p_2 = 6\theta^2(1-\theta)^2, p_3 = 4\theta^3(1-\theta), p_4 = \theta^4, \text{for some } \theta\in(0,1)\\ H_2: & \text{otherwise.}\end{cases}$$

Under the binomial distribution assumption, the likelihood function is $f(\theta) = \prod_{i=0}^4 p_i^{N_i}$, where N_i is the number of people who previously attended i games. So the log-likelihood function is

$$\begin{aligned}\log f(\theta) &= 4N_0\log(1-\theta) + N_1[\log 4+\log\theta+3\log(1-\theta)] + N_2[\log 6+2\log\theta+2\log(1-\theta)]\\ &\quad +N_3[\log 4+3\log\theta+\log(1-\theta)] + 4N_4\log\theta.\end{aligned}$$

Take derivative with respect to θ and set the result as zero, we can solve for the MLE of θ: $\widehat{\theta} = 0.4$. Then we have $p_0(\widehat{\theta}) = 0.1296$, $p_1(\widehat{\theta}) = 0.3456$, $p_2(\widehat{\theta}) = 0.3456$, $p_3(\widehat{\theta}) = 0.1536$, and $p_4(\widehat{\theta}) = 0.0256$. So the statistic T for the χ^2 test of goodness-of-fit is equal to

$$\begin{aligned}
& \frac{(33 - 200 \cdot 0.1296)^2}{200 \cdot 0.1296} + \frac{(67 - 200 \cdot 0.3456)^2}{200 \cdot 0.3456} + \frac{(66 - 200 \cdot 0.3456)^2}{200 \cdot 0.3456} \\
& + \frac{(15 - 200 \cdot 0.1536)^2}{200 \cdot 0.1536} + \frac{(19 - 200 \cdot 0.0256)^2}{200 \cdot 0.0256} \\
\approx \; & 47.81.
\end{aligned}$$

By the theorem in Lecture 25, $T \to \chi^2_{5-1-1} = \chi^2_3$ as $n \to \infty$. For any given level of significance α, we can choose c such that $\chi^2_3(c, \infty) = \alpha$, and the decision rule

$$\delta = \begin{cases} H_1: & T \le c \\ H_2: & T > c \end{cases}$$

will have level of significance α. For example, for $c = 7.815$, $\chi^2_3(c, \infty) = 0.05$. Since $T > c$ in this case, we reject the hypotheses at level of significance 0.05. $\square$

► 2. [1] page 554, No. 1. Chase and Dummer (1992) studied the attitudes of school-aged children in Michigan. The children were asked which of the following was most important to them: good grades, athletic ability, or popularity. Additional information about each child was also collected, and the following table shows the results for 478 children classified by sex and their response to the survey questions:

	Good grades	Athletic ability	Popularity
Boys	117	60	50
Girls	130	30	91

Test the null hypothesis that a child's answer to the survey question is independent of his or her sex.

Solution. $N_{11} = 117$, $N_{12} = 60$, $N_{13} = 50$, $N_{21} = 130$, $N_{22} = 30$, $N_{23} = 91$, $N_{1+} = 227$, $N_{2+} = 251$, $N_{+1} = 247$, $N_{+2} = 90$, and $N_{+3} = 141$. By the calculation in Lecture 26, the χ^2 statistic is

$$\begin{aligned}
T &= [(117 - 227 \cdot 247/478)^2/(227 \cdot 247) + (60 - 227 \cdot 90/478)^2/(227 \cdot 90) \\
&\quad + (50 - 227 \cdot 141/478)^2/(227 \cdot 141) + (130 - 251 \cdot 247/478)^2/(251 \cdot 247) \\
&\quad + (30 - 251 \cdot 90/478)^2/(251 \cdot 90) + (91 - 251 \cdot 141/478)^2/(251 \cdot 141)] \times 478 \\
&\approx 21.46.
\end{aligned}$$

For any given level of significance α, if we choose the decision rule

$$\delta = \begin{cases} \text{independent}: & T \le c \\ \text{dependent}: & T > c, \end{cases}$$

where c is such that $\chi^2_{(2-1)(3-1)}(c, \infty) = \alpha$, the decision rule δ will have level of significance α. For example, for $c = 5.991$, $\chi^2_2(c, \infty) = 0.05$. Since $T > 5.991$, we reject the independence hypotheses at level of significance 0.05. The following Matlab program facilitates the calculation of T:

```
function T = independence_test(A)

%INDEPENDENCE_TEST    gives the statistic of chi-square test of
%                     independence, which is based on Pearson's Theorem.
%                     See [1] Lecture 26.
%
%                     [1] Dmitry Panchenko: 18.443. Statistics for Applications,
%                     Fall 2003, MIT OpenCourseWare.
```

```
%
%                    Yan Zeng, 07/29/2008.


rowSum = sum(A,2);
columnSum = sum(A);
numData = sum(sum(A));

T = 0;
[numRow, numColumn] = size(A);
for i = 1:numRow
    for j= 1:numColumn
       T = T + (A(i,j)- rowSum(i)*columnSum(j)/numData)^2/(rowSum(i)*columnSum(j)/numData);
    end
end

end
```

□

► 3. [1] page 555, No. 5. Suppose that 300 persons are selected at random from a large population, and each person in the sample is classified according to blood type, O, A, B, or AB, also according to Rh, positive or negative. The observed numbers are given in the following table.

	O	A	B	AB
Rh positive	82	89	54	19
Rh negative	13	27	7	9

Test the hypothesis that the two classifications of blood types are independent.

Solution. $N_{11} = 82$, $N_{12} = 89$, $N_{13} = 54$, $N_{14} = 19$, $N_{21} = 13$, $N_{22} = 27$, $N_{23} = 7$, $N_{24} = 9$, $N_{1+} = 244$, $N_{2+} = 56$, $N_{+1} = 95$, $N_{+2} = 116$, $N_{+3} = 61$, and $N_{+4} = 28$. By the calculation in Lecture 26, the χ^2 statistic T is equal to (using the Matlab program provided in Problem 2)

$$\begin{aligned}
& [(82-244\cdot 95/300)^2/(244\cdot 95)+(89-244\cdot 116/300)^2/(244\cdot 116)+(54-244\cdot 61/300)^2/(244\cdot 61)\\
& +(19-244\cdot 28/300)^2/(244\cdot 28)+(13-56\cdot 95/300)^2/(56\cdot 95)+(27-56\cdot 116/300)^2/(56\cdot 116)\\
& +(7-56\cdot 61/300)^2/(56\cdot 61)+(9-56\cdot 28/300)^2/(56\cdot 28)]\times 300\\
\approx\ & 8.6.
\end{aligned}$$

For any given level of significance α, if we choose the decision rule

$$\delta = \begin{cases} \text{independent}: & T \le c \\ \text{dependent}: & T > c, \end{cases}$$

where c is such that $\chi^2_{(2-1)(4-1)}(c, \infty) = \alpha$, the decision rule δ will have level of significance α. For example, for $c = 7.815$, $\chi^2_3(c, \infty) = 0.05$. Since $T > 7.815$, we reject the independence hypotheses at level of significance 0.05. □

► 4. [1] page 561, No. 1. The survey of Chase and Dummer (1992) discussed in Exercise 1 of Section 9.3 ([1] page 554, No. 1) was actually collected by sampling from three sub-populations according to the locations of the schools: rural, suburban, and urban. The following table shows the response to the survey question classified by school location:

	Good grades	Athletic ability	Popularity
Rural	57	42	50
Suburban	87	22	42
Urban	103	26	49

Test the null hypothesis that the distribution of responses is the same in all three types of school location.

Solution. As argued in Lecture 27, we can continue to use the test of goodness-of-fit for independence. $N_{11} = 57$, $N_{12} = 42$, $N_{13} = 50$, $N_{21} = 87$, $N_{22} = 22$, $N_{23} = 42$, $N_{31} = 103$, $N_{32} = 26$, $N_{33} = 49$, $N_{1+} = 149$, $N_{2+} = 151$, $N_{3+} = 178$, $N_{+1} = 247$, $N_{+2} = 90$, and $N_{+3} = 141$. By the calculation in Lecture 26 and using the Matlab program provided in Problem 2, the statistic T can be easily obtained as 18.83.

For any given level of significance α, if we choose the decision rule

$$\delta = \begin{cases} \text{homogeneous}: & T \le c \\ \text{nonhomogeneous}: & T > c, \end{cases}$$

where c is such that $\chi^2_{(3-1)(3-1)}(c, \infty) = \alpha$, the decision rule δ will have level of significance α. For example, for $c = 9.488$, $\chi^2_4(c, \infty) = 0.05$. Since $T > 9.488$, we reject the homogeneity hypotheses at level of significance 0.05. $\square$

9 Problem Set 9: up to Lecture 31

▶ 1. [1] page 574, No. 4. Use the Kolmogorov-Smirnov test to test the hypothesis that the 25 values in the following table form a random sample from a uniform distribution on the interval $[0, 1]$.

0.42	0.06	0.88	0.40	0.90
0.38	0.78	0.71	0.57	0.66
0.48	0.35	0.16	0.22	0.08
0.11	0.29	0.79	0.75	0.82
0.30	0.23	0.01	0.41	0.09

Solution. The hypotheses under testing is

$$\begin{cases} H_1: & F(x) = F_0(x) = x \\ H_2: & \text{otherwise.} \end{cases}$$

According to the calculation in Lecture 28, the decision rule

$$\delta = \begin{cases} H_1: & D_n \le 1.35 \\ H_2: & D_n > 1.35 \end{cases}$$

will have level of significance 0.05. Here $D_n = \sqrt{n}\sup_{0\le x\le 1}|F_n(x) - x|$ with $n = 25$. To find D_n, we use the following Matlab program (the input A should be a vector):

```
function [maxValue, xMax] = ks_test1(A)

x = 0:0.01:1;
y = zeros(1, numel(x));

% Obtain empirical distribution based on data vector A
for i = 1:numel(x)
    y(i) = mean(A<=x(i));
end

plot(x,x,'b',x,y,'r');
[maxValue, maxIndex] = max(abs(y-x));

xMax = x(maxIndex);

end
```

At $x = 0.42$, $|F_n(x) - F(x)|$ achieves its maximum 0.18. So $D_n = \sqrt{25} \times 0.18 = 0.9$ and we accept H_1 at level of significance 0.05. $\square$

▶ 2. [1] page 574, No. 5. Use the Kolmogorov-Smirnov test to test the hypothesis that the 25 values given in Exercise 4 form a random sample from a continuous distribution for which the p.d.f. $f(x)$ is as follows:

$$f(x) = \begin{cases} \frac{3}{2} & \text{for } 0 < x \leq \frac{1}{2} \\ \frac{1}{2} & \text{for } \frac{1}{2} < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Solution. The distribution function corresponding to $f(x)$ is

$$F_0(x) = \begin{cases} 0 & x = 0 \\ \frac{3}{2}x & 0 < x \leq \frac{1}{2} \\ \frac{x+1}{2} & \frac{1}{2} < x < 1 \\ 1 & x = 1. \end{cases}$$

The hypotheses under testing is

$$\begin{cases} H_1: & F(x) = F_0(x) \\ H_2: & \text{otherwise.} \end{cases}$$

According to the calculation in Lecture 28, the decision rule

$$\delta = \begin{cases} H_1: & D_n \leq 1.35 \\ H_2: & D_n > 1.35 \end{cases}$$

will have level of significance 0.05. We use the following Matlab program to find D_n:

```
function [maxValue, xMax] = ks_test2(A)

x = 0:0.01:1;
y = zeros(1, numel(x));

% Obtain empirical distribution based on data vector A
for i = 1:numel(x)
    y(i) = mean(A<=x(i));
end

% Obtain the true distribution
fx = zeros(1, numel(x));
for i = 1:numel(x)
    if x(i) <= 0.5
        fx(i) = 1.5*x(i);
    else
        fx(i) = (x(i)+1)/2;
    end
end

plot(x,fx,'b',x,y,'r');
[maxValue, maxIndex] = max(abs(y-fx));

xMax = x(maxIndex);

end
```

At $x = 0.66$, $|F_n(x) - F(x)|$ achieves its maximum 0.15. So $D_n = \sqrt{25} \times 0.15 = 0.75$ and we accept H_1 at level of significance 0.05. $\square$

► 3. [1] page 637, No. 6. Consider again the conditions of Exercise 4 and 5. Suppose that the prior probability is 1/2 that the 25 values given in the table for Exercise 4 were obtained from a uniform distribution on the interval $[0, 1]$, and 1/2 that they were obtained from a distribution for which the p.d.f. is as given in Exercise 5. Find the posterior probability that they were obtained from a uniform distribution.

Solution. We let θ be a Binomial random variable so that $\theta = 1$ means the 25 values given in the table for Problem 1 (Exercise 4 in the textbook [1]) were obtained from a uniform distribution on the interval $[0, 1]$, and $\theta = 0$ means these values were obtained from a distribution whose p.d.f. is as given in Problem 2 (Exercise 5 in the textbook [1]). Then the a priori distribution of θ is $P(\theta = 0) = P(\theta = 1) = 0.5$. The posterior probability that these values were obtained from a uniform distribution is given by

$$
\begin{aligned}
& P(\theta = 1 | X_1 = x_1, \cdots, X_{25} = x_{25}) \\
= \; & \frac{P(X_1 = x_1, \cdots, X_{25} = x_{25} | \theta = 1) P(\theta = 1)}{P(X_1 = x_1, \cdots, X_{25} = x_{25} | \theta = 1) P(\theta = 1) + P(X_1 = x_1, \cdots, X_{25} = x_{25} | \theta = 0) P(\theta = 0)} \\
= \; & \frac{P(X_1 = x_1, \cdots, X_{25} = x_{25} | \theta = 1)}{P(X_1 = x_1, \cdots, X_{25} = x_{25} | \theta = 1) + P(X_1 = x_1, \cdots, X_{25} = x_{25} | \theta = 0)} \\
= \; & \frac{1}{1 + 1.5 \sum_{i=1}^{25} 1_{\{0 < x_i \le 0.5\}} + 0.5 \sum_{i=1}^{25} 1_{\{0.5 < x_i < 1\}}} \\
= \; & \frac{1}{1 + 28.5} \\
= \; & 0.034.
\end{aligned}
$$

$\square$

► 4. [1] page 637. For the data in Table 10.9 (see below) find the 90% confidence intervals for β_0, β_1, σ^2 and construct 90% prediction interval for y at $x = 1.5$. Test the hypothesis that $\beta_1 \ge 0$ at the level of significance 0.1.

i	x_i	y_i	i	x_i	y_i
1	0.3	0.4	6	1.0	0.8
2	1.4	0.9	7	2.0	0.7
3	1.0	0.4	8	−1.0	−0.4
4	−0.3	−0.3	9	−0.7	−0.2
5	−0.2	0.3	10	0.7	0.7

Solution. We base our calculations on the content in Lecture 31. First, we find the MLE $\widehat{\beta}_0$, $\widehat{\beta}_1$ and $\widehat{\sigma}^2$:

$$\widehat{\beta}_1 = \frac{\bar{XY} - \bar{X}\bar{Y}}{\bar{X^2} - (\bar{X})^2} = \frac{0.504 - 0.1386}{1.016 - 0.1764} = 0.4352, \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{X} = 0.1472,$$

and

$$\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i)^2 = 0.0451.$$

We set $\alpha = 0.1$ and choose α_1 and α_2 such that $\chi^2_{10-2}(0, c_1) = \frac{\alpha}{2} = 0.05$ and $\chi^2_{10-2}(c_2, \infty) = \frac{\alpha}{2} = 0.05$. Then $c_1 = 2.7326$, $c_2 = 15.5073$, and the 90% confidence interval for σ^2 is $[\frac{n\widehat{\sigma}^2}{c_2}, \frac{n\widehat{\sigma}^2}{c_1}] = [0.0291, 0.1650]$.

To find the confidence interval for β_1, we need to find c such that $t_{10-2}(-c, c) = 1 - \alpha = 0.9$. So $c = 1.8595$ and the 90% confidence interval for β_1 is

$$\left[\widehat{\beta}_1 - c\sqrt{\frac{\widehat{\sigma}^2}{(n-2)(\bar{X^2} - (\bar{X})^2)}}, \widehat{\beta}_1 + c\sqrt{\frac{\widehat{\sigma}^2}{(n-2)(\bar{X^2} - (\bar{X})^2)}} \right] = [0.2828, 0.5876].$$

Consequently, the 90% confidence interval for β_0 is

$$\left[\widehat{\beta}_0 - c\sqrt{\frac{\widehat{\sigma}^2}{n-2}\left(1+\frac{(\bar{X})^2}{\bar{X}^2-(\bar{X})^2}\right)}, \widehat{\beta}_0 + c\sqrt{\frac{\widehat{\sigma}^2}{n-2}\left(1+\frac{(\bar{X})^2}{\bar{X}^2-(\bar{X})^2}\right)}\right] = [-0.0064, 0.3008].$$

At $x = 1.5$, the prediction $\widehat{y}$ of y is $\widehat{\beta}_0 + \widehat{\beta}_1 \cdot 1.5 = 0.8$. So the 90% prediction interval for y at $x = 1.5$ is

$$\left[\widehat{Y} - c\sqrt{\frac{\widehat{\sigma}^2}{n-2}\left(10+1+\frac{(\bar{X}-X)^2}{\bar{X}^2-(\bar{X})^2}\right)}, \widehat{Y} + c\sqrt{\frac{\widehat{\sigma}^2}{n-2}\left(10+1+\frac{(\bar{X}-X)^2}{\bar{X}^2-(\bar{X})^2}\right)}\right] = [0.3086, 1.2914].$$

Finally, to construct a decision rule for the hypotheses testing:

$$\begin{cases} H_0: & \beta_1 \geq 0 \\ H_1: & \beta_1 < 0, \end{cases}$$

we apply the results in Section 10.3 of [1]. Namely, we have shown $\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{n(\bar{X}^2-(\bar{X})^2)}\right)$. So

$$\sqrt{n(\bar{X}^2-(\bar{X})^2)}\left(\frac{\widehat{\beta}_1-\beta_1}{\sigma}\right)$$

is a standard normal random variable. We have also shown $\frac{n\widehat{\sigma}^2}{\sigma^2}$ has χ^2_{n-2} distribution with $n-2$ degree of freedom and is independent of $\widehat{\beta}_1$. So

$$T = \frac{\sqrt{n(\bar{X}^2-(\bar{X})^2)}\left(\frac{\widehat{\beta}_1-\beta_1}{\sigma}\right)}{\sqrt{\frac{1}{n-2}\frac{n\widehat{\sigma}^2}{\sigma^2}}} = \sqrt{(n-2)(\bar{X}^2-(\bar{X})^2)}\left(\frac{\widehat{\beta}_1-\beta_1}{\widehat{\sigma}}\right)$$

has t distribution with $n-2$ degree of freedom. So if we let $\beta_1 = 0$ in the expression of T, T becomes a statistic. And it makes sense to consider the following decision rule

$$\delta = \begin{cases} H_0: & T > c \\ H_1: & T \leq c, \end{cases}$$

where c is such that $t_8(-\infty, c) = 0.1$. Looking up table gives us $c = -1.3968$. In our case, $T = 5.3111 > c$. So we accept H_0 at level of significance 0.1. □

References

[1] Morris H. DeGroot and Mark J. Schervish. *Probability and statistics*, 3rd Edition. Pearson Addison Wesley, 2001. 4, 5, 6, 7, 8, 9, 10, 12, 13, 14, 15, 16, 17, 18, 19, 20