STATISTICS 330
Mathematical Statistics

Supplementary Lecture Notes

Cyntha A. Struthers
Dept. of Statistics and Actuarial Science
University of Waterloo Waterloo, Ontario, Canada

Winter 2012

Contents

Chapter 1

PREVIEW

The follow examples will illustrate the ideas and concepts we will study in STAT 330.

1.1 Example

The following table gives the number of fumbles in a game made by 110 Division A football teams during one weekend:

No. of Fumbles: x	0	1	2	3	4	5	6	7	≥ 8	Total
Obs. Frequency: f_x	8	24	27	20	17	10	3	1	0	110

It is believed that a Poisson model will fit these data well. Why might this be a reasonable assumption? *(PROBABILITY MODELS)*

If we let the random variable $X =$ *number of fumbles in a game* and assume that the Poisson model is reasonable then the probability function (p.f.) of X is given by

$$P(X = x) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, \ldots$$

where μ is a parameter of the model which represents the mean number of fumbles in a game. *(RANDOM VARIABLES, PROBABILITY FUNCTIONS, EXPECTATION, MODEL PARAMETERS)* Since μ is unknown we might estimate it using the sample mean

$$\bar{x} = \frac{8(0) + 24(1) + \cdots + 1(7)}{110} = \frac{281}{110} \approx 2.55.$$

(POINT ESTIMATION) The estimate $\hat{\mu} = \bar{x}$ is the maximum likelihood (M.L.) estimate of μ. It is the value of μ which maximizes the likelihood

function. *(MAXIMUM LIKELIHOOD ESTIMATION)* The likelihood function is the probability of the observed data as a function of the unknown parameter(s) in the model. The M.L. estimate is thus the value of μ which maximizes the probability of the observed data.

In this example the likelihood function is given by

$$\begin{aligned} L(\mu) &= P(\text{observing 0 fumbles 8 times, } \ldots, \geq 8 \text{ fumbles 0 times;}\mu), \qquad \mu > 0 \\ &= \frac{110!}{8!24!\cdots 1!0!}\left(\frac{\mu^0 e^{-\mu}}{0!}\right)^8\left(\frac{\mu^1 e^{-\mu}}{1!}\right)^{24}\cdots\left(\frac{\mu^7 e^{-\mu}}{7!}\right)^1\left(\sum_{x=8}^{\infty}\frac{\mu^x e^{-\mu}}{x!}\right)^0 \\ &= c\mu^{8(0)+24(1)+\cdots+1(7)}e^{-(8+24+\cdots+1)} \\ &= c\mu^{-281}e^{-110\mu}, \qquad \mu > 0 \end{aligned}$$

where

$$c = \frac{110!}{8!24!\cdots 1!0!}\left(\frac{1}{0!}\right)^8\left(\frac{1}{1!}\right)^{24}\cdots\left(\frac{1}{7!}\right)^1.$$

The M.L. estimate of μ can be found by solving $\frac{dL}{d\mu} = 0$ (or equivalently $\frac{d\log L}{d\mu} = 0$) and verifying that it corresponds to a maximum.

If we want an interval of values for μ which are reasonable given the data then we could construct a confidence interval (C.I.) for μ. *(INTERVAL ESTIMATION)* To construct C.I.'s we need to find the sampling distribution of the estimator. In this example we would need to find the distribution of the estimator

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

where $X_i =$ *number of fumbles in game* i, $i = 1, \ldots, n$. *(FUNCTIONS OF RANDOM VARIBABLES: cumulative distribution function (c.d.f.) technique, one-to-one transformations, moment generating function (m.g.f.) technique)* Since $X_i \sim \text{POI}(\mu)$ with $E(X_i) = \mu$ and $Var(X_i) = \mu$ the distribution of $\bar{X}$ for large n is approximately $\text{N}(\mu, \mu/n)$ by the Central Limit Theorem. *(LIMITING DISTRIBUTIONS)*

Suppose a football expert claims that ten years ago the mean number of fumbles was 3. Then we would like to test the hypothesis $H : \mu = 3$. *(TESTS OF HYPOTHESIS)* A test of hypothesis uses a test statistic to measure the evidence based on the observed data against the hypothesis. A test statistic with good properties for testing $H : \mu = \mu_0$ is the likelihood ratio statistic, $-2\log\left[L\left(\mu_0\right)/L\left(\hat{\mu}\right)\right]$. *(LIKELIHOOD RATIO STATISTIC)* The limiting distribution of the likelihood ratio statistic is $\chi^2\left(1\right)$ if the hypothesis $H : \mu = \mu_0$ is true.

1.2 Example

The following are relief times in hours for 20 patients receiving a pain killer:

$$1.1,\ 1.4,\ 1.3,\ 1.7,\ 1.9,\ 1.8,\ 1.6,\ 2.2,\ 1.7,\ 2.7,$$
$$4.1,\ 1.8,\ 1.5,\ 1.2,\ 1.4,\ 3.0,\ 1.7,\ 2.3,\ 1.6,\ 2.0$$

It is believed that the Weibull distribution with probability density function (p.d.f.)

$$f(x) = \frac{\beta}{\theta^\beta} x^{\beta-1} e^{-(x/\theta)^\beta}, \qquad x > 0, \quad \theta > 0, \quad \beta > 0$$

will provide a good fit to the data. *(CONTINUOUS MODELS, PROBABILITY DENSITY FUNCTIONS)* The (approximate) likelihood function in this case is assuming independent observation is

$$L(\theta, \beta) = \prod_{i=1}^{20} \frac{\beta}{\theta^\beta} x_i^{\beta-1} e^{-(x_i/\theta)^\beta}, \qquad \theta > 0, \quad \beta > 0$$

where x_i is the observed relief time for the ith patient. *(MULTIPARAMETER LIKELIHOODS)* The M.L. estimates $\hat{\theta}$ and $\hat{\beta}$ are found by simultaneously solving

$$\frac{\partial \log L}{\partial \theta} = 0, \qquad \frac{\partial \log L}{\partial \beta} = 0.$$

Since an explicit solution to these equations cannot be obtained, a numerical solution must be found using an iterative method. *(NEWTON'S METHOD)* Also, since the M.L. estimators cannot be given explicitly, approximate C.I.'s and tests of hypothesis must be based on the asymptotic distributions of the M.L. estimators. *(LIMITING OR ASYMPTOTIC DISTRIBUTIONS OF M.L. ESTIMATORS)*

Chapter 2

RANDOM VARIABLES

2.1 Introduction

2.1.1 Definition

Suppose S is a sample space for a random experiment. Let $\mathcal{B} = \{A_1, A_2, \dots\}$ be a suitable class of subsets of S. A probability set function is a function P with domain $\mathcal{B}$ that satisfies:

(1) $P(A) \geq 0$ for all $A \in \mathcal{B}$.

(2) $P(S) = 1$.

(3) If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise mutually exclusive events, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i).$$

2.1.2 Exercise

If P is a probability set function and A and B are any sets in $\mathcal{B}$ then prove the following:

(a) $P\left(\bar{A}\right) = 1 - P(A)$
(b) $P(\emptyset) = 0$
(c) $P(A) \leq 1$
(d) $P\left(A \cap \bar{B}\right) = P(A) - P(A \cap B)$
(e) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
(f) If $A \subseteq B$ then $P(A) \leq P(B)$

2.1.3 Definition

Suppose S is a sample space for a random experiment. Suppose A and B are subsets of S, that is, A and B are events defined on S. The conditional probability of event A given event B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \text{provided } P(B) > 0.$$

2.1.4 Definition

Suppose S is a sample space for a random experiment. Suppose A and B are events defined on S. A and B are independent events if

$$P(A \cap B) = P(A)\, P(B).$$

2.1.5 Definition

A *random variable* X is a function from a sample space S to the real numbers $\Re$, that is,

$$X : S \to \Re$$

such that $P(X \leq x)$ is defined for all $x \in \Re$.

Note: '$X \leq x$' is an abbreviation for $\{\omega \in S : X(\omega) \leq x\}$
where $\{\omega \in S : X(\omega) \leq x\} \in \mathcal{B}$.

2.1.6 Definition

The *cumulative distribution function* (c.d.f.) of a random variable X is defined by

$$F(x) = P(X \leq x), \quad x \in \Re.$$

2.1.7 Properties of F

(1) F is a non-decreasing function, that is, $F(x_1) \leq F(x_2)$ for all $x_1 < x_2$

(2) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

(3) F is a right-continuous function, that is, $\lim_{x \to a^+} F(x) = F(a)$

(4) $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a), \ a < b$

(5) $P(X = b) = F(b) - \lim_{a \to b^-} F(a)$.

Note: The definition and properties of the c.d.f. hold for the random variable X regardless of whether S is discrete (finite or countable) or not.

2.2 Discrete Random Variables

2.2.1 Definition

If S is discrete (finite or countable) then X is called a *discrete random variable*. In this case F is a right-continuous step function.

2.2.2 Definition

If X is a discrete random variable then the *probability function* (p.f.) of X is given by

$$f(x) = P(X = x) = F(x) - \lim_{\varepsilon \to 0^+} F(x - \varepsilon), \quad x \in \Re$$

The set $A = \{x : f(x) > 0\}$ is called the support set of X.

2.2.3 Properties of f

(1) $f(x) \geq 0, \ \ x \in \Re$

(2) $\sum\limits_{x \in A} f(x) = 1$

2.2.4 Example

A box contains a red balls and b black balls. Find the p.f. of the random variable X for each of the following:

(a) X = number of red balls in n selections without replacement.

(b) X = number of red balls in n selections with replacement.

(c) X = number of black balls selected before obtaining the first red ball if sampling is done with replacement.

(d) X = number of black balls selected before obtaining the kth red ball if sampling is done with replacement.

2.2.5 Example

If X is a random variable with p.f.

$$f(x) = \frac{\mu^x e^{-\mu}}{x!}, \quad x = 0, 1, \ldots ; \ \ \mu > 0$$

show that

$$\sum_{x=0}^{\infty} f(x) = 1.$$

2.2.6 Exercise

If X is a random variable with p.f.

$$f(x) = \frac{-(1-p)^x}{x \log p}, \quad x = 1, 2, \ldots ; \quad 0 < p < 1$$

show that

$$\sum_{x=1}^{\infty} f(x) = 1.$$

2.3 Continuous Random Variables

2.3.1 Definition

Suppose X is a random variable with c.d.f. F. If F is a continuous function for all $x \in \Re$ and F is differentiable except possibly at countably many points then X is called a *continuous random variable.*

2.3.2 Definition

If X is a continuous random variable with c.d.f. $F(x)$ then the *probability density function* (p.d.f.) of X is $f(x) = F'(x)$ if F is differentiable at x and otherwise we define $f(x) = 0$. The set $A = \{x : f(x) > 0\}$ is called the support set of X.

2.3.3 Properties of f

(1) $f(x) \geq 0$ for all $x \in \Re$

(2) $\int_{-\infty}^{\infty} f(x)dx = \lim_{x \to \infty} F(x) - \lim_{x \to -\infty} F(x) = 1$

(3) $f(x) = \lim_{h \to 0} \frac{F(x+h)-F(x)}{h} = \lim_{h \to 0} \frac{P(x \leq X \leq x+h)}{h}$ if this limit exists

(4) $F(x) = \int_{-\infty}^{x} f(t)dt, \quad x \in \Re$

(5) $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) = \int_{a}^{b} f(x)dx$

(6) $P(X = b) = F(b) - \lim_{a \to b^-} F(a) = F(b) - F(b) = 0$
(since F is continuous).

2.3.4 Example

Suppose X is a random variable with c.d.f.

$$F(x) = \begin{cases} 0 & x \leq a \\ \frac{x-a}{b-a} & a< x < b \\ 1 & x \geq b \end{cases}$$

where $b > a$. Find the p.d.f. of X and sketch both the c.d.f. and the p.d.f. of X.

2.3.5 Example

Consider the function

$$f(x) = \frac{\theta}{x^{\theta+1}}, \quad x \geq 1$$

and 0 otherwise. For what values of θ is this function a p.d.f.?

2.3.6 Gamma Function:

The *gamma function*, denoted by $\Gamma(\alpha)$ for all $\alpha > 0$, is given by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$$

2.3.7 Properties of the Gamma Function

(1) $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1), \ \alpha > 1.$
(2) $\Gamma(n) = (n-1)!, \quad n = 1, 2, \ldots$
(3) $\Gamma(\frac{1}{2}) = \sqrt{\pi}.$

2.3.8 Example

Suppose X is a random variable with p.d.f.

$$f(x) = \frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}}, \qquad x > 0, \quad \alpha, \beta > 0.$$

X is said to have a gamma distribution with parameters α and β and we write $X \backsim \text{GAM}(\alpha, \beta)$. Verify that $\int_{-\infty}^{\infty} f(x)dx = 1$.

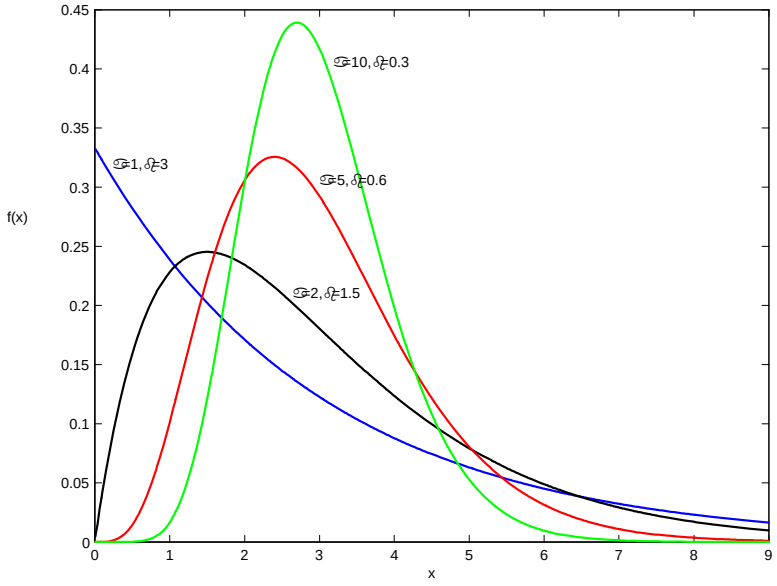


Figure 2.1: GAM(α, β) p.d.f.'s

2.3.9 Exercise

Suppose X is a random variable with p.d.f.

$$f(x) = \frac{\beta}{\theta^\beta} x^{\beta-1} e^{-(x/\theta)^\beta}, \qquad x > 0, \quad \theta, \beta > 0.$$

X is said to have a Weibull distribution with parameters θ and β and we write $X \backsim \text{WEI}(\theta, \beta)$. Verify that $\int_{-\infty}^{\infty} f(x)dx = 1$.

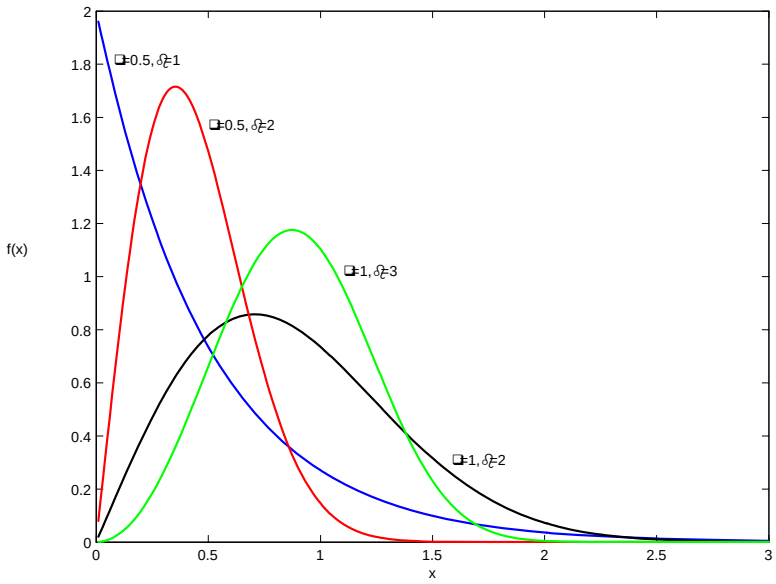


Figure 2.2: WEI(θ, β) p.d.f.'s

2.4 Location and Scale Parameters

In Chapter 6 we look at methods for constructing confidence intervals for an unknown parameter θ. If the parameter θ is either a scale parameter or a location parameter then a confidence interval is easier to construct.

2.4.1 Definition

Suppose X is a continuous random variable with p.d.f. $f(x;\theta)$ where θ is a parameter of the distribution.

Let $F_0(x) = F(x;\theta=0)$ and $f_0(x) = f(x;\theta=0)$. The parameter θ is called a *location parameter* of the distribution if

$$F(x;\theta) = F_0(x-\theta), \quad \theta \in \Re$$

or equivalently

$$f(x;\theta) = f_0(x-\theta), \quad \theta \in \Re.$$

2.4.2 Definition

Suppose X is a continuous random variable with p.d.f. $f(x;\theta)$ where θ is a parameter of the distribution.

Let $F_1(x) = F(x;\theta=1)$ and $f_1(x) = f(x;\theta=1)$. The parameter θ is called a *scale parameter* of the distribution if

$$F(x;\theta) = F_1\left(\frac{x}{\theta}\right), \quad \theta > 0$$

or equivalently

$$f(x;\theta) = \frac{1}{\theta} f_1\left(\frac{x}{\theta}\right), \quad \theta > 0.$$

2.4.3 Example

(1) If $X \sim \text{EXP}(1, \theta)$ then show that θ is a location parameter of the distribution. See Figure 2.3.

(2) If $X \sim \text{EXP}(\theta)$ then show that θ is a scale parameter of the distribution. See Figure 2.4.

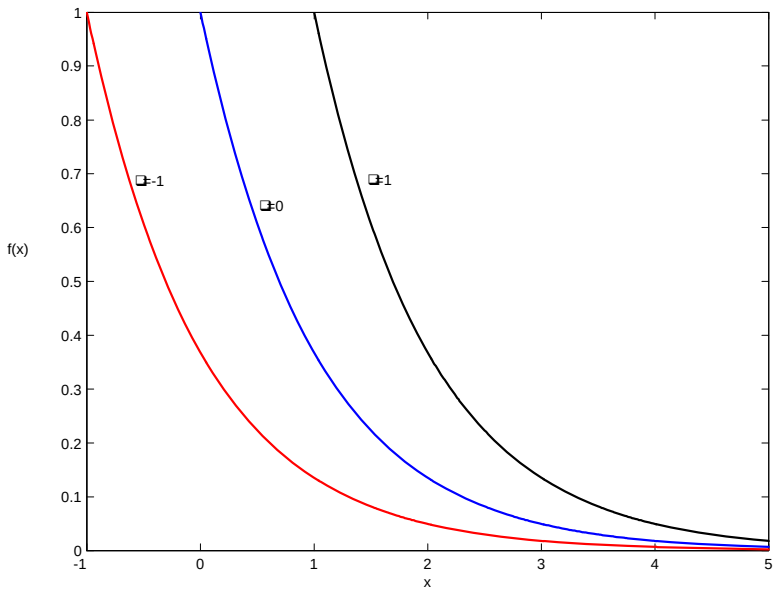


Figure 2.3: $\text{EXP}(1, \theta)$ p.d.f.'s

2.4.4 Exercise

(1) If $X \sim \text{CAU}(1, \theta)$ then show that θ is a location parameter of the distribution. Graph the $\text{CAU}(1, \theta)$ p.d.f. for $\theta = -1$, 0 and 1 on the same graph.

(2) If $X \sim \text{CAU}(\theta, 0)$ then show that θ is a scale parameter of the distribution. Graph the $\text{CAU}(1, \theta)$ p.d.f. for $\theta = 0.5$, 1 and 2 on the same graph.

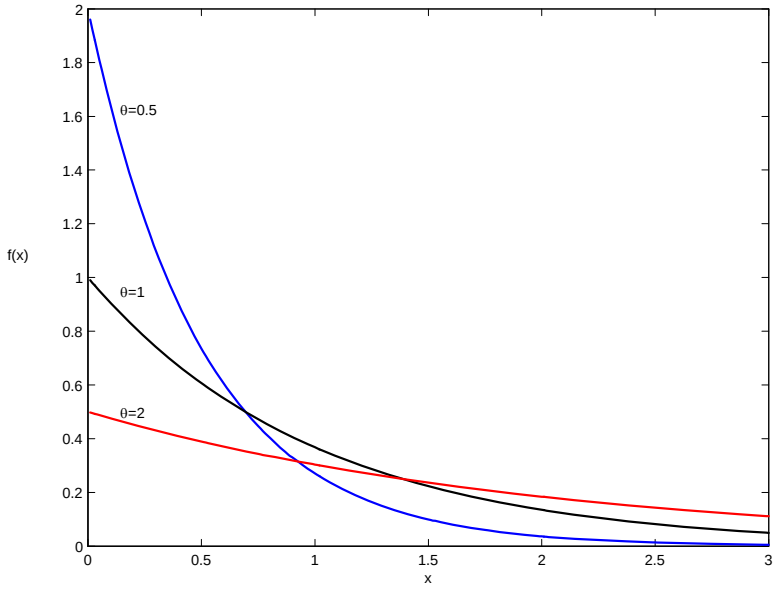


Figure 2.4: EXP(θ) p.d.f.'s

2.5 Functions of a Random Variable

Suppose X is a continuous random variable with p.d.f. f and c.d.f. F and we wish to find the p.d.f. of the random variable $Y = h(X)$ where h is a real-valued function. A useful method is the *c.d.f. technique*. This method involves obtaining an expression for $G(y) = P(Y \leq y)$, the c.d.f. of Y, in terms of F, the c.d.f. of X. The corresponding p.d.f. g of Y is found by differentiating G. Care must be taken to determine the support set of Y.

2.5.1 Example

If $Z \sim \mathrm{N}(0,1)$ find the p.d.f. of $Y = Z^2$. If $X \sim \mathrm{N}(\mu, \sigma^2)$ what is the distribution of $W = \left(\frac{X-\mu}{\sigma}\right)^2$?

2.5.2 Probability Integral Transformation

If X is a continuous random variable with c.d.f. F then
$Y = F(X) \sim \text{UNIF}(0,1)$. $Y = F(X)$ is called the *probability integral transformation.*

2.5.3 Example

Suppose F is a c.d.f. for a continuous random variable. Show that if $U \sim \text{UNIF}(0,1)$ then the random variable $X = F^{-1}(U)$ also has c.d.f. F. Why is this result useful for simulating observations from a continuous distribution?

2.5.4 Theorem - One-to-One Transformation of a Random Variable

Suppose X is a continuous random variable with p.d.f. f and support set $A = \{x : f(x) > 0\}$ and $Y = h(X)$ where h is a real-valued function. Let g be the p.d.f. of the random variable Y and let $B = \{y : g(y) > 0\}$. If h is a one-to-one function from A to B and if h' is continuous then

$$g(y) = f(h^{-1}(y)) \cdot \left| \frac{d}{dy} h^{-1}(y) \right|, \quad y \in B.$$

2.5.5 Proof

(1) Suppose h is an increasing function for $x \in A$. Then $h^{-1}(y)$ is also an increasing function and $\frac{d}{dy} h^{-1}(y) > 0$ for $y \in B$.

$$\begin{aligned} \text{Now} \quad G(y) &= P(Y \le y) = P(h(X) \le y) \\ &= P\left(X \le h^{-1}(y)\right) \quad \text{since } h \text{ is a 1-1 increasing function} \\ &= F\left(h^{-1}(y)\right). \end{aligned}$$

$$\begin{aligned} \text{Therefore} \quad g(y) &= \frac{d}{dy} G(y) = \frac{d}{dy} F\left(h^{-1}(y)\right) \\ &= F'\left(h^{-1}(y)\right) \frac{d}{dy} h^{-1}(y) \quad \text{by the Chain Rule} \\ &= f\left(h^{-1}(y)\right) \left| \frac{d}{dy} h^{-1}(y) \right|, \quad y \in B \quad \text{since } \frac{d}{dy} h^{-1}(y) > 0. \end{aligned}$$

(2) Suppose h is a decreasing function for $x \in A$. Then $h^{-1}(y)$ is also a decreasing function and $\frac{d}{dy}h^{-1}(y) < 0$ for $y \in B$.

$$\begin{aligned} \text{Now} \quad G(y) &= P(Y \le y) = P(h(X) \le y) \\ &= P\left(X \ge h^{-1}(y)\right) \quad \text{since } h \text{ is a 1-1 decreasing function} \\ &= 1 - F\left(h^{-1}(y)\right). \end{aligned}$$

$$\begin{aligned} \text{Therefore} \quad g(y) &= \frac{d}{dy}G(y) = \frac{d}{dy}\left[1 - F\left(h^{-1}(y)\right)\right] \\ &= -F'\left(h^{-1}(y)\right)\frac{d}{dy}h^{-1}(y) \quad \text{by the Chain Rule} \\ &= f\left(h^{-1}(y)\right)\left|\frac{d}{dy}h^{-1}(y)\right|, \quad y \in B \quad \text{since } \frac{d}{dy}h^{-1}(y) > 0. \end{aligned}$$

2.5.6 Example

Find the p.d.f. of $Y = \log(X) = \ln(X)$ if X is a continuous random variable with p.d.f.

$$f(x) = \frac{\theta}{x^{\theta+1}}, \quad x \ge 1, \quad \theta > 0.$$

2.5.7 Exercise

If $X \sim \text{EXP}(1)$ then show

$$Y = \theta X^{1/\beta} \sim \text{WEI}(\theta, \beta), \quad \theta, \beta > 0.$$

2.6 Expectation

2.6.1 Definition

If X is a discrete random variable with p.f. $f(x)$ and support set A then the *expectation of* X or the *expected value of* X is defined by

$$E(X) = \sum_{x \in A} x f(x)$$

provided the sum converges absolutely, that is, provided

$$E(|X|) = \sum_{x \in A} |x| \, f(x) dx < \infty.$$

If X is a continuous random variable with p.d.f. $f(x)$ then the *expectation of* X or the *expected value of* X is defined by

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

provided the integral converges absolutely, that is, provided

$$E(|X|) = \int_{-\infty}^{\infty} |x| \, f(x) < \infty.$$

If $E(|X|) = \infty$ then we say that $E(X)$ does not exist.

2.6.2 Example

Find $E(X)$ if $X \sim \text{GEO}(p)$.

2.6.3 Example

Suppose X is a continuous random variable with p.d.f.

$$f(x) = \frac{\theta}{x^{\theta+1}}, \quad x \geq 1, \quad \theta > 0$$

and 0 otherwise. Find $E(X)$. For what values of θ does $E(X)$ exist?

2.6.4 Exercise

Suppose X is a nonnegative continuous random variable with c.d.f. $F(x)$ and $E(X) < \infty$. Show that

$$E(X) = \int_0^\infty [1 - F(x)]\, dx.$$

Hint: Use integration by parts with $u = [1 - F(x)]$.

2.6.5 Theorem

Suppose $h(x)$ is a real-valued function.

If X is a discrete random variable with p.f. $f(x)$ and support set A then

$$E[h(X)] = \sum_{x \in A} h(x) f(x)$$

provided the sum converges absolutely.

If X is a continuous random variable with p.d.f. $f(x)$ then

$$E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) dx$$

provided the integral converges absolutely.

2.6.6 Theorem

Suppose X is a random variable with p.f./p.d.f. $f(x)$, a and b are real constants, and $g(x)$ and $h(x)$ are real-valued functions. Then

$$E[ag(X) + bh(X)] = aE[g(X)] + bE[h(X)].$$

2.6.7 Special Expectations

(1) The *variance* of a random variable:

$$Var(X) = E[(X - \mu)^2] = \sigma^2 \qquad \text{where } \mu = E(X)$$

(2) The *kth moment (about the origin)*:

$$E(X^k)$$

(3) The *kth moment about the mean*:

$$E[(X-\mu)^k]$$

(4) The *kth factorial moment*:

$$E(X^{(k)}) = E[X(X-1)\cdots(X-k+1)].$$

2.6.8 Theorem

$$\begin{aligned} Var(X) &= E(X^2) - \mu^2 \\ &= E[X(X-1)] + \mu - \mu^2, \end{aligned}$$

$$Var(aX+b) = a^2 Var(X).$$

and

$$E(X^2) = \sigma^2 + \mu^2.$$

2.6.9 Example

If $X \sim \text{BIN}(n,p)$ then show

$$E(X^{(k)}) = n^{(k)}p^k, \quad k = 1,2,\ldots$$

and thus find $Var(X)$.

2.6.10 Exercise

Show the following:
(1) If $X \sim \text{POI}(\theta)$ then $E(X^{(k)}) = \theta^k, \;\; k = 1,2,\ldots$
(2) If $X \sim \text{NB}(k,p)$ then $E(X^{(j)}) = (-k)^{(j)}\left(\frac{p-1}{p}\right)^j, \;\; j = 1,2,\ldots$
(3) If $X \sim \text{GAM}(\alpha,\beta)$ then $E(X^p) = \beta^p \Gamma(\alpha+p)/\Gamma(\alpha), \;\; p > -\alpha.$
In each case find $Var(X)$.

2.7 Inequalities

In Chapter 5 we consider limiting distributions of a sequence of random variables. The following inequalities which involve the moments of a distribution are useful for proving limit theorems.

2.7.1 Markov's Inequality

$$P(|X| \geq c) \leq \frac{E(|X|^k)}{c^k}, \quad \text{for all } k,\ c > 0.$$

2.7.2 Proof of Markov's Inequality

Suppose X is a continuous random variable with p.d.f. $f(x)$. (The proof of the discrete case follows by replacing integrals with sums.) Let

$$A = \left\{ x : \left|\frac{x}{c}\right|^k \geq 1 \right\} = \{x : |x| \geq c\} \text{ since } c > 0.$$

Then

$$\begin{aligned}
\frac{E\left(|X|^k\right)}{c^k} &= E\left(\left|\frac{X}{c}\right|^k\right) = \int_{-\infty}^{\infty} \left|\frac{x}{c}\right|^k f(x)\,dx \\
&= \int_A \left|\frac{x}{c}\right|^k f(x)\,dx + \int_{\bar{A}} \left|\frac{x}{c}\right|^k f(x)\,dx \\
&\geq \int_A \left|\frac{x}{c}\right|^k f(x)\,dx \qquad \text{since} \int_{\bar{A}} \left|\frac{x}{c}\right|^k f(x)\,dx \geq 0 \\
&\geq \int_A f(x)\,dx \qquad \text{since } \left|\frac{x}{c}\right|^k \geq 1 \text{ for } x \in A \\
&= P(|X| \geq c) \qquad \text{as required.}
\end{aligned}$$

2.7.3 Chebyshev's Inequality

Suppose X is a random variable with finite mean μ and finite variance σ^2. Then for any $k > 0$

$$P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}.$$

2.7.4 Exercise

Use Markov's Inequality to prove Chebyshev's Inequality.

2.8 Variance Stabilizing Transformation

In Chapter 6 we look at methods for constructing a confidence interval for an unknown parameter θ. To do this it is often useful to find a transformation $g(X)$ of the observed data X whose variance is approximately constant with respect to θ.

Suppose X is a random variable with finite mean $E(X) = \theta$. Suppose also that X has finite variance $Var(X) = \sigma^2(\theta)$ and standard deviation $\sqrt{Var(X)} = \sigma(\theta)$ depending on θ as well. Let $Y = g(X)$ where g is a differentiable function. By the linear approximation

$$Y = g(X) \approx g(\theta) + g'(\theta)(X - \theta).$$

Therefore

$$E(Y) \approx E[g(\theta) + g'(\theta)(X - \theta)] = g(\theta)$$

since

$$E[g'(\theta)(X - \theta)] = g'(\theta) E[(X - \theta)] = 0.$$

Also

$$Var(Y) \approx Var[g'(\theta)(X - \theta)] = [g'(\theta)]^2 Var(X) = [g'(\theta)\sigma(\theta)]^2.$$

If we want $Var(Y) \approx$ constant with respect to θ then we should choose g such that

$$[g'(\theta)]^2 Var(X) = [g'(\theta)\sigma(\theta)]^2 = \text{ constant}.$$

In other words we need to solve the differential equation

$$\frac{dg}{d\theta} = \frac{k}{\sigma(\theta)}$$

where k is a conveniently chosen constant.

2.8.1 Example

If $X \backsim \text{POI}(\theta)$ then show that the random variable $Y = g(X) = \sqrt{X}$ has approximately constant variance.

2.8.2 Exercise

If $X \backsim \text{EXP}(\theta)$ then show that the random variable $Y = g(X) = \log X$ has approximately constant variance.

2.9 Moment Generating Functions

2.9.1 Definition

If X is a random variable then $M(t) = E(e^{tX})$ is called the *moment generating function (m.g.f.) of* X if this expectation exists for all $t \in (-h, h)$ for some $h > 0$.

Important:

When determining the m.g.f. of a random variable the values of t for which the expectation exists must always be stated.

2.9.2 Example

(1) If $X \sim \text{GAM}(\alpha, \beta)$ then find $M(t)$.

(2) If $X \sim \text{NB}(k, p)$ then find $M(t)$.

2.9.3 Exercise

Show the following:

(1) If $X \sim \text{BIN}(n, p)$ then $M(t) = (q + pe^t)^n$, $t \in \Re$.

(2) If $X \sim \text{POI}(\theta)$ then $M(t) = e^{\mu\left(e^t - 1\right)}$, $t \in \Re$.

2.9.4 Theorem

Suppose the random variable X has m.g.f. $M_X(t)$ defined for $t \in (-h, h)$ for some $h > 0$. Let $Y = aX + b$ where $a, b \in \mathcal{R}$ and $a \neq 0$. Then the m.g.f. of Y is

$$M_Y(t) = e^{bt} M_X(at), \qquad |t| < \frac{h}{|a|}.$$

2.9.5 Example

If $Z \sim \text{N}(0, 1)$ then find $M_Z(t)$, the m.g.f. of Z. Use this to find $M_X(t)$ the m.g.f. of $X \sim \text{N}(\mu, \sigma^2)$.

2.9.6 Exercise

If $X \sim \text{NB}(k, p)$ then find the m.g.f. of $Y = X + k$, $k = 1, 2, \ldots$

2.9.7 Theorem

Suppose the random variable X has m.g.f. $M(t)$ defined for $t \in (-h, h)$ for some $h > 0$. Then $M(0) = 1$ and

$$M^{(k)}(0) = E(X^k), \quad k = 1, 2, \ldots$$

where

$$M^{(k)}(t) = \frac{d^k}{dt^k} M(t)$$

is the kth derivative of $M(t)$.

2.9.8 Important Idea

Suppose $M^{(k)}(t)$, $k = 1, 2, \ldots$ exists for $t \in (-h, h)$ for some $h > 0$, then $M(t)$ has a Maclaurin series given by

$$\sum_{k=0}^{\infty} \frac{M^{(k)}(0)}{k!} t^k$$

where

$$M^{(0)}(0) = M(0) = 1.$$

The coefficient of t^k in this power series is equal to

$$\frac{M^{(k)}(0)}{k!} = \frac{E(X^k)}{k!}.$$

Therefore if we can obtain a Maclaurin series for $M(t)$, for example, by using the Binomial series or the exponential series, then we can find $E(X^k)$ by using

$$E(X^k) = k! \times \text{coefficient of } t^k \text{ in the Maclaurin series for } M(t).$$

2.9.9 Example

If $X \sim \text{GAM}(\alpha, \beta)$ then $M(t) = (1 - \beta t)^{-\alpha}$, $t < 1/\beta$. Find $M^{(k)}(0) = E\left(X^k\right)$ by first evaluating the derivatives directly and secondly by using the Binomial series expansion for $(1 - \beta t)^{-\alpha}$.

In Chapter 4 we look at methods for determining the distributions of functions of random variables. The method of moment generating functions is particularly useful for finding distributions of sums of independent random variables. The following theorem plays an important role in this technique.

2.9.10 Uniqueness Theorem for m.g.f.'s

Suppose the random variable X has m.g.f. $M_X(t)$ and the random variable Y has m.g.f. $M_Y(t)$. Suppose also that $M_X(t) = M_Y(t)$ for all $t \in (-h, h)$ for some $h > 0$. Then X and Y have the same distribution, that is, $P(X \leq s) = F_X(s) = F_Y(s) = P(Y \leq s)$ for all $s \in \Re$.

2.9.11 Example

If $X \sim \text{EXP}(1)$ then find the distribution of $Y = \mu + \beta X$ where $\beta > 0$ and $\mu \in \Re$.

2.9.12 Exercise

If $X \sim \text{GAM}(\alpha, \beta)$, where α is a positive integer and $\beta > 0$, then show

$$\frac{2X}{\beta} \sim \chi^2(2\alpha).$$

2.9.13 Exercise

Suppose the random variable X has m.g.f.

$$M(t) = e^{t^2/2}, \qquad t \in \Re.$$

(a) Find the m.g.f. of $Y = 2X - 1$.

(b) Use the m.g.f. of Y to find $E(Y)$ and $Var(Y)$.

(c) What is the distribution of Y?

2.10 Calculus Review

2.10.1 Geometric Series

$$\sum_{x=0}^{\infty} at^x = a + at + at^2 + \cdots = \frac{a}{1-t}, \qquad |t| < 1.$$

2.10.2 Useful Results

(1)

$$\sum_{x=0}^{\infty} t^x = \frac{1}{1-t}, \qquad |t| < 1$$

(2)

$$\sum_{x=1}^{\infty} xt^{x-1} = \frac{1}{(1-t)^2}, \qquad |t| < 1$$

2.10.3 Binomial Series

(1) For $n \in \mathcal{Z}^+$ (the positive integers)

$$(a+b)^n = \sum_{x=0}^{n} \binom{n}{x} a^x b^{n-x}$$

where

$$\binom{n}{x} = \frac{n!}{x!\,(n-x)!} = \frac{n^{(x)}}{x!}.$$

(2) For $n \in \mathcal{Q}$ (the rational numbers) and $|t| < 1$

$$(1+t)^n = \sum_{x=0}^{\infty} \binom{n}{x} t^x$$

where

$$\binom{n}{x} = \frac{n^{(x)}}{x!} = \frac{n(n-1)\cdots(n-x+1)}{x!}.$$

2.10.4 Important Identities

$$(1) \quad x^{(k)}\binom{n}{x} = n^{(k)}\binom{n-k}{x-k}$$

$$(2) \quad \binom{x+k-1}{x} = \binom{x+k-1}{k-1} = (-1)^x\binom{-k}{x}$$

2.10.5 Exercise

Prove the identities in 2.8.4.

2.10.6 Multinomial Therorem

If n is a positive integer and $a_1, a_2, \ldots, a_k$ are real numbers, then

$$(a_1 + a_2 + \cdots + a_k)^n = \sum\sum\cdots\sum \frac{n!}{x_1!x_2!\cdots x_k!} a_1^{x_1} a_2^{x_2} \cdots a_k^{x_k}$$

where the summation extends over all non-negative integers $x_1, x_2, \ldots, x_k$ with $x_1 + x_2 + \cdots + x_k = n$.

2.10.7 Hypergeometric Identity

$$\sum_{x=0}^{\infty} \binom{a}{x}\binom{b}{n-x} = \binom{a+b}{n}$$

2.10.8 Exponential Series

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \cdots = \sum_{x=0}^{\infty} \frac{x^n}{n!}, \qquad x \in \Re.$$

2.10.9 Logarithmic Series

$$\ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \cdots, \qquad -1 < x \leq 1$$

2.10.10 First Fundamental Theorem of Calculus (FTCI)

If f is continuous on $[a, b]$ then the function g defined by

$$g(x) = \int_a^x f(t)\, dt, \qquad a \leq x \leq b$$

is continuous on $[a, b]$, differentiable on (a, b) and $g'(x) = f(x)$.

2.10.11 FTCI and the Chain Rule

Suppose we want the derivative with respect to x of $G(x)$ where

$$G(x) = \int_a^{h(x)} f(t)\,dt, \qquad a \le x \le b$$

and $h(x)$ is a differentiable function on $[a,b]$. If we define

$$g(u) = \int_a^u f(t)\,dt$$

then $G(x) = g(h(x))$. Then by the Chain Rule

$$\begin{aligned} G'(x) &= g'(h(x)) \cdot h'(x) \\ &= f(h(x)) \cdot h'(x) \qquad a < x < b. \end{aligned}$$

2.10.12 Exercise

Find $G'(x)$ if

$$G(x) = \int_{h_1(x)}^{h_2(x)} f(t)\,dt, \qquad a \le x \le b$$

Hint:

$$G(x) = \int_{h_1(x)}^{c} f(t)\,dt + \int_c^{h_2(x)} f(t)\,dt, \quad a < c < b$$

2.10.13 Improper Integrals

(a) If $\int_a^b f(x)\,dx$ exists for every number $b \ge a$ then

$$\int_a^\infty f(x)\,dx = \lim_{b\to\infty} \int_a^b f(x)\,dx$$

provided this limit exists. If the limit exists we say the improper integral converges otherwise we say the improper integral diverges.

(b) If $\int_a^b f(x)\,dx$ exists for every number $a \leq b$ then

$$\int_{-\infty}^{b} f(x)\,dx = \lim_{a\to-\infty} \int_a^b f(x)\,dx$$

provided this limit exists.

(c) If both $\int_a^{\infty} f(x)\,dx$ and $\int_{-\infty}^{a} f(x)\,dx$ are convergent then we define

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_{-\infty}^{a} f(x)\,dx + \int_a^{\infty} f(x)\,dx$$

where a is any real number.

2.10.14 Comparison Test for Improper Integrals

Suppose that f and g are continuous functions with $f(x) \geq g(x) \geq 0$ for $x \geq a$.

(a)

$$\text{If } \int_a^{\infty} f(x)\,dx \text{ is convergent then } \int_a^{\infty} g(x)\,dx \text{ is convergent.}$$

(b)

$$\text{If } \int_a^{\infty} g(x)\,dx \text{ is divergent then } \int_a^{\infty} f(x)\,dx \text{ is divergent.}$$

2.10.15 Useful Result for Comparison Test

$$\int_1^{\infty} \frac{1}{x^p}dx \text{ converges if and only if } p > 1.$$

2.10.16 Useful Inequalities

$$\frac{1}{1+y^p} \leq \frac{1}{y^p}, \qquad y \geq 1,\, p > 0$$

$$\frac{1}{1+y^p} \geq \frac{1}{y^p+y^p} = \frac{1}{2y^p}, \qquad y \geq 1,\, p > 0$$

Chapter 3

Joint Distributions

3.1 Joint and Marginal CDF's

3.1.1 Definition

Suppose X and Y are random variables defined on a sample space S. The *joint c.d.f. of X and Y* is given by

$$F(x,y) = P(X \leq x, Y \leq y), \quad (x,y) \in \Re^2.$$

3.1.2 Properties of F

(1) F is non-decreasing in x for fixed y

(2) F is non-decreasing in y for fixed x

(3) $\lim_{x \to -\infty} F(x,y) = 0$ and $\lim_{y \to -\infty} F(x,y) = 0$

(4) $\lim_{(x,y) \to (-\infty,-\infty)} F(x,y) = 0$ and $\lim_{(x,y) \to (\infty,\infty)} F(x,y) = 1$

3.1.3 Definition

The *marginal c.d.f. of X* is given by

$$F_1(x) = P(X \leq x) = \lim_{y \to \infty} F(x,y) = F(x,\infty), \quad x \in \Re.$$

The *marginal c.d.f. of Y* is given by

$$F_2(y) = P(Y \leq y) = \lim_{x \to \infty} F(x,y) = F(\infty,y), \quad y \in \Re.$$

Note:

The definitions and properties of the joint c.d.f. and the marginal c.d.f.'s hold for both (X, Y) discrete random variables and for (X, Y) continuous random variables.

3.2 Joint Discrete Random Variables

3.2.1 Definition

Suppose X and Y are random variables defined on a sample space S. If S is discrete then X and Y are discrete random variables.

The *joint p.f.* of X and Y is given by

$$f(x, y) = P(X = x, Y = y), \quad (x, y) \in \Re^2.$$

The set $A = \{(x, y) : f(x, y) > 0\}$ is called the support set of (X, Y).

3.2.2 Properties of f

(1) $f(x, y) \geq 0$ for $(x, y) \in \Re^2$

(2) $\sum\limits_{(x,y)\in} \sum\limits_{A} f(x, y) = 1$

(3) For any set $R \subset \Re^2$,

$$P\left[(X, Y) \in R\right] = \sum_{(x,y)\in} \sum_{R} f(x, y).$$

3.2.3 Definition

Suppose X and Y are discrete random variables with joint p.f. $f(x, y)$.

The *marginal p.f. of* X is given by

$$f_1(x) = P(X = x) = \sum_{y} f(x, y), \quad x \in \Re$$

and the *marginal p.f. of* Y is given by

$$f_2(y) = P(Y = y) = \sum_{x} f(x, y), \quad y \in \Re.$$

3.2.4 Example

In a fourth year statistics course there are 10 actuarial science students, 9 statistics students and 6 math business students. Five students are selected at random without replacement.

Let X be the number of actuarial students selected and let Y be the number of statistics students selected.

Find

(a) the joint p.f. of X and Y

(b) the marginal p.f. of X

(c) the marginal p.f. of Y

(d) $P(X > Y)$.

3.2.5 Exercise

The Hardy-Weinberg law of genetics states that, under certain conditions, the relative frequencies with which three genotypes AA, Aa and aa occur in the population will be θ^2, $2\theta(1-\theta)$ and $(1-\theta)^2$ respectively where $0 < \theta < 1$. Suppose n members of the population are selected at random.

Let X be the number of AA types selected and let Y be the number of Aa types selected.

Find

(a) the joint p.f. of X and Y

(b) the marginal p.f. of X

(c) the marginal p.f. of Y

(d) $P(X + Y = t)$ for $t = 0, 1, \ldots$.

3.3 Joint Continuous Random Variables

3.3.1 Definition

Suppose that $F(x, y)$ is continuous and that

$$f(x, y) = \frac{\partial^2}{\partial x \partial y} F(x, y)$$

exists and is continuous except possibly along a finite number of curves. Suppose also that

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1.$$

Then X and Y are said to be continuous random variables with joint p.d.f. f. The set $A = \{(x, y) : f(x, y) > 0\}$ is called the support of (X, Y).

Note: We arbitrarily define $f(x, y)$ to be equal to 0 when $\frac{\partial^2}{\partial x \partial y} F(x, y)$ does not exist.

3.3.2 Properties of f

(1) $f(x, y) \geq 0$ for all $(x, y) \in \Re^2$

(2)

$$P\left[(X, Y) \in R\right] = \iint_R f(x, y) dx dy, \qquad R \subset \Re^2$$

3.3.3 Definition

Suppose X and Y are continuous random variables with joint p.d.f. $f(x, y)$. Then the *marginal p.d.f. of* X is given by

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y) dy, \qquad x \in \Re$$

and the *marginal p.d.f. of* Y is given by

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y) dx, \qquad y \in \Re.$$

3.3.4 Example

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x,y) = x + y, \quad 0 \le x \le 1, \quad 0 \le y \le 1$$

and 0 otherwise. Show that

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1.$$

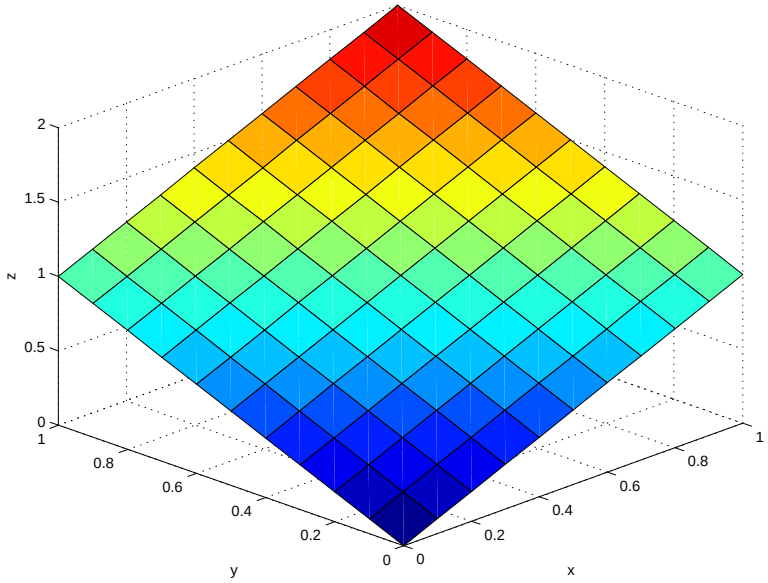


Figure 3.1: Graph of joint p.d.f. for Example 3.3.4

Find
(1) $P\left(X \le \frac{1}{3}, Y \le \frac{1}{2}\right)$
(2) $P(X \le Y)$
(3) $P(X + Y \le \frac{1}{2})$
(4) $P(XY \le \frac{1}{2})$
(5) the marginal p.d.f. of X and the marginal p.d.f. of Y
(6) the joint c.d.f. of X and Y
(7) the marginal c.d.f. of X and the marginal c.d.f. of Y

3.3.5 Exercise

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x,y) = ke^{-x-y}, \quad 0 < x < y < \infty$$

and 0 otherwise. Determine k and sketch $f(x,y)$.

Find
(1) $P\left(X \leq \frac{1}{3}, Y \leq \frac{1}{2}\right)$
(2) $P(X \leq Y)$
(3) $P(X + Y \geq 1)$
(4) the marginal p.d.f. of X and the marginal p.d.f. of Y
(5) the joint c.d.f. of X and Y
(6) the marginal c.d.f. of X and the marginal c.d.f. of Y

3.3.6 Exercise

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x,y) = \frac{k}{(1+x+y)^3}, \quad 0 < x < \infty, \;\; 0 < y < \infty$$

and 0 otherwise. Determine k and sketch $f(x,y)$.

Find
(1) $P\left(X \leq \frac{1}{3}, Y \leq \frac{1}{2}\right)$
(2) $P(X \leq Y)$
(3) $P(X + Y \geq 1)$
(4) the marginal p.d.f. of X and the marginal p.d.f. of Y
(5) the joint c.d.f. of X and Y
(6) the marginal c.d.f. of X and the marginal c.d.f. of Y

3.4 Independent Random Variables

3.4.1 Definition

Two random variables X and Y are called *independent random variables* if,

$$P(X \in A \text{ and } Y \in B) = P(X \in A)P(Y \in B)$$

for all sets A and B of real numbers.

3.4.2 Theorem

Suppose X and Y are random variables with joint c.d.f. $F(x, y)$, joint p.f./p.d.f. $f(x, y)$, marginal c.d.f.'s $F_1(x)$ and $F_2(y)$ respectively, and marginal p.f./p.d.f.'s $f_1(x)$ and $f_2(y)$ respectively. Suppose also that $A_1 = \{x : f_1(x) > 0\}$ is the support set of X, and $A_2 = \{y : f_2(y) > 0\}$ is the support set of Y. Then X and Y are independent random variables if and only if either of the following holds:

$$f(x, y) = f_1(x)f_2(y) \quad \textit{for all } (x, y) \in A_1 \times A_2$$

where $A_1 \times A_2 = \{(x, y) : x \in A_1, y \in A_2\}$

$$F(x, y) = F_1(x)F_2(y) \quad \textit{for all } x \in \Re \text{ and } y \in \Re.$$

3.4.3 Corollary

If X and Y are independent random variables then $h(X)$ and $g(Y)$ are also independent random variables where h and g are real-valued functions.

3.4.4 Example

(1) In Example 3.2.4 are X and Y independent random variables?

(2) In Example 3.3.4 are X and Y independent random variables?

3.4.5 Exercise

In Exercises 3.2.5, 3.3.5 and 3.3.6 are X and Y independent random variables?

3.4.6 Factorization Theorem for Independence

Suppose X and Y are random variables with joint p.f./p.d.f. $f(x, y)$, and marginal p.f./p.d.f.'s $f_1(x)$ and $f_2(y)$ respectively. Suppose also that $A = \{(x, y) : f(x, y) > 0\}$ is the support set of (X, Y), $A_1 = \{x : f_1(x) > 0\}$ is the support set of X, and $A_2 = \{y : f_2(y) > 0\}$ is the support set of Y. Then X and Y are independent random variables if and only if $A = A_1 \times A_2$ and there exist non-negative functions $g(x)$ and $h(y)$ such that

$$f(x, y) = g(x)h(y)$$

for all $(x, y) \in A_1 \times A_2$.

Notes:

(1) If the Factorization Theorem for Independence holds then f_1 will be proportional to g and f_2 will be proportional to h.

(2) The above definitions and theorems can easily be extended to the random vector $(X_1, X_2, \ldots, X_n)$.

3.4.7 Example

Suppose X and Y are discrete random variables with joint p.f.

$$f(x, y) = \frac{\theta^{x+y} e^{-2\theta}}{x!y!}, \quad x = 0, 1 \ldots, \quad y = 0, 1, \ldots.$$

Are X and Y independent random variables? Find the marginal p.f. of X and the marginal p.f. of Y.

3.4.8 Example

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x, y) = \frac{3}{2} y(1 - x^2), \quad -1 \leq x \leq 1, \quad 0 \leq y \leq 1$$

and 0 otherwise. Are X and Y independent random variables? Find the marginal p.d.f. of X and the marginal p.d.f. of Y.

3.4.9 Example

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x, y) = \frac{2}{\pi}, \quad 0 \leq x \leq \sqrt{1 - y^2}, \quad -1 < y < 1$$

and 0 otherwise. Are X and Y independent random variables? Find the marginal p.d.f. of X and the marginal p.d.f. of Y.

3.5 Conditional Distributions

3.5.1 Definition

Suppose X and Y are random variables with joint p.f./p.d.f. $f(x, y)$, and marginal p.f./p.d.f.'s $f_1(x)$ and $f_2(y)$ respectively. Suppose also that $A = \{(x, y) : f(x, y) > 0\}$.

The *conditional p.f./p.d.f. of X given $Y = y$* is given by

$$f_1(x|y) = \frac{f(x, y)}{f_2(y)}$$

for $(x, y) \in A$ provided $f_2(y) \neq 0$.

The *conditional p.f./p.d.f. of Y given $X = x$* is given by

$$f_2(y|x) = \frac{f(x, y)}{f_1(x)}$$

for $(x, y) \in A$ provided $f_1(x) \neq 0$.

Notes:

(1) If X and Y are discrete random variables then

$$f_1(x|y) = P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x, y)}{f_2(y)}$$

and

$$\sum_x f_1(x|y) = \sum_x \frac{f(x, y)}{f_2(y)} = \frac{1}{f_2(y)} \sum_x f(x, y) = \frac{f_2(y)}{f_2(y)} = 1.$$

Similarly for $f_2(y|x)$.

(2) If X and Y are continuous random variables

$$\int_{-\infty}^{\infty} f_1(x|y)dx = \int_{-\infty}^{\infty} \frac{f(x, y)}{f_2(y)}dx = \frac{1}{f_2(y)} \int_{-\infty}^{\infty} f(x, y)dx = \frac{f_2(y)}{f_2(y)} = 1$$

Similarly for $f_2(y|x)$.

(3) If X is a continuous random variable then $f_1(x) \neq P(X = x)$ and $P(X = x) = 0$ for all x. Therefore to justify the definition of the conditional p.d.f. of Y given $X = x$ when X and Y are continuous random variables we consider $P(Y \leq y|X = x)$ as a limit:

$$\begin{aligned}
& P(Y \leq y|X = x) \\
= & \lim_{h\to 0} P(Y \leq y|x \leq X \leq x+h) \\
= & \lim_{h\to 0} \frac{\int_x^{x+h}\int_{-\infty}^{y} f(u,v)\,dvdu}{\int_x^{x+h} f_1(u)\,du} \\
= & \lim_{h\to 0} \frac{\frac{d}{dh}\int_x^{x+h}\int_{-\infty}^{y} f(u,v)\,dvdu}{\frac{d}{dh}\int_x^{x+h} f_1(u)\,du} \text{ by L'Hospital's Rule} \\
= & \lim_{h\to 0} \frac{\int_{-\infty}^{y} f(x+h,v)\,dv}{f_1(x+h)} \text{ by the Fundamental Theorem of Calculus} \\
= & \frac{\lim_{h\to 0}\int_{-\infty}^{y} f(x+h,v)\,dv}{\lim_{h\to 0} f_1(x+h)} \\
= & \frac{\int_{-\infty}^{y} f(x,v)\,dv}{f_1(x)}
\end{aligned}$$

assuming that the limits exist and that integration and the limit operation can be interchanged. If we differentiate the last term with respect to y using the Fundamental Theorem of Calculus we have

$$\frac{d}{dy}P(Y \leq y|X = x) = \frac{f(x,y)}{f_1(x)}$$

which gives us a justification for using

$$f_2(y|x) = \frac{f(x,y)}{f_1(x)}$$

as the conditional probability density function of Y given $X = x$.

3.5.2 Example

In Example 3.4.9 find the conditional p.d.f. of X given $Y = y$ and the conditional p.d.f. of Y given $X = x$.

3.5.3 Exercise

In Exercise 3.2.5 show that

$$Y|X = x \sim \text{BIN}\left(n - x,\ \frac{2\theta\,(1-\theta)}{1-\theta^2}\right).$$

3.5.4 Exercise

In Example 3.3.4 and Exercises 3.3.5 and 3.3.6 find the conditional p.d.f. of X given $Y = y$ and the conditional p.d.f. of Y given $X = x$. Check that $\int_{-\infty}^{\infty} f_1(x|y)dx = \int_{-\infty}^{\infty} f_2(y|x)dy = 1$.

3.5.5 Product Rule

Suppose X and Y are random variables with joint p.f./p.d.f. $f(x, y)$, marginal p.f./p.d.f.'s $f_1(x)$ and $f_2(y)$ respectively and conditional p.f./p.d.f.'s $f_1(x|y)$ and $f_2(y|x)$. Then

$$f(x, y) = f_1(x|y)f_2(y) = f_2(y|x)f_1(x).$$

3.5.6 Example

Find the marginal p.f. of X if $Y \sim \text{POI}(\mu)$ and $X|Y = y \sim \text{BIN}(y, p)$.

3.5.7 Example

Find the marginal p.f. of X if $Y \sim \text{GAM}(\alpha, \frac{1}{\theta})$ and $X|Y = y \sim \text{WEI}(y^{-1/p}, p)$.

3.5.8 Theorem

Suppose X and Y are random variables with marginal p.f./p.d.f.'s $f_1(x)$ and $f_2(y)$ respectively and conditional p.f./p.d.f.'s $f_1(x|y)$ and $f_2(y|x)$. Let $A_1 = \{x : f_1(x) > 0\}$ and $A_2 = \{y : f_2(y) > 0\}$. Then X and Y are independent random variables if and only if either of the following holds:

$$f_1(x|y) = f_1(x) \quad \text{for all } x \in A_1$$

or

$$f_2(y|x) = f_2(y) \quad \text{for all } y \in A_2.$$

3.6 Joint Expectations

3.6.1 Definition

Suppose $h(x, y)$ is a real-valued function.

If X and Y are discrete random variables with joint p.f. $f(x, y)$ and support A then

$$E[h(X, Y)] = \sum_{(x,y) \in A} \sum h(x, y) f(x, y)$$

provided the joint sum converges absolutely.

If X and Y are continuous random variables then with joint p.d.f. $f(x, y)$ then

$$E[h(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) f(x, y) dx dy$$

provided the joint integral convergs absolutely.

3.6.2 Theorem

Suppose X and Y are random variables with joint p.f./p.d.f. $f(x, y)$, a and b are real constants, and $g(x, y)$ and $h(x, y)$ are real-valued functions. Then

$$E[ag(X, Y) + bh(X, Y)] = aE[g(X, Y)] + bE[h(X, Y)].$$

3.6.3 Corollary

(1)

$$E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$$

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$.

(2) If $X_1, X_2, \ldots, X_n$ are random variables and $a_1, a_2, \ldots, a_n$ are real constants then

$$E\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_i$$

where $\mu_i = E(X_i)$.

3.6.4 Theorem

(1) If X and Y are independent random variables and $g(x)$ and $h(y)$ are real valued functions then

$$E\left[g(X)h(Y)\right] = E[g(X)]E[h(Y)].$$

(2) More generally if $X_1, X_2, \ldots, X_n$ are independent random variables and $h_1, h_2, \ldots, h_n$ are real valued functions then

$$E\left[\prod_{i=1}^{n} h_i(X_i)\right] = \prod_{i=1}^{n} E\left[h_i(X_i)\right].$$

3.6.5 Definition

The *covariance* of random variables X and Y is given by

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)].$$

If $Cov(X,Y) = 0$ then X and Y are called *uncorrelated* random variables.

3.6.6 Theorem

If X and Y are random variables then

$$Cov(X,Y) = E(XY) - \mu_X\mu_Y.$$

If X and Y are independent random variables then $Cov(X,Y) = 0$.

3.6.7 Theorem

(1) Suppose X and Y are random variables and a and b are real constants then

$$\begin{aligned} Var(aX+bY) &= a^2Var(X) + b^2Var(Y) + 2abCov(X,Y) \\ &= a^2\sigma_X^2 + b^2\sigma_Y^2 + 2abCov(X,Y). \end{aligned}$$

(2) Suppose $X_1, X_2, \ldots, X_n$ are random variables with $Var(X_i) = \sigma_i^2$ and $a_1, a_2, \ldots, a_n$ are real constants then

$$Var\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_ia_jCov\left(X_i,X_j\right).$$

(3) If $X_1, X_2, \ldots, X_n$ are independents random variables and $a_1, a_2, \ldots, a_n$ are real constants then

$$Var\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2\sigma_i^2.$$

3.6.8 Definition

The *correlation coefficient* of random variables X and Y is given by

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

3.6.9 Exercise

In Example 3.3.4 and Exercise 3.3.5 find $\rho(X,Y)$.

3.6.10 Theorem

If $\rho(X,Y)$ is the correlation coefficient of random variables X and Y then

$$-1 \leq \rho(X,Y) \leq 1.$$

$\rho(X,Y) = 1$ if and only if $Y = aX + b$ for some $a > 0$ and $\rho(X,Y) = -1$ if and only if $Y = aX + b$ for some $a < 0$.

3.7 Conditional Expectation

3.7.1 Definition

The *conditional expectation of* $g(Y)$ *given* $X = x$ is given by

$$E\left[g(Y)|x\right] = \sum_y g(y) f_2(y|x)$$

if Y is a discrete random variable and

$$E[g(Y)|x] = \int_{-\infty}^{\infty} g(y) f_2(y|x) dy$$

if Y is a continuous random variable provided the sum/integral converges absolutely. The conditional expectation of $h(X)$ given $Y = y$ is defined in a similar manner.

3.7.2 Special Cases

(1) The *conditional mean of* Y *given* $X = x$ is denoted by $E(Y|x)$.
(2) The *conditional variance of* Y *given* $X = x$ is denoted by $Var(Y|x)$ and is given by

$$Var(Y|x) = E\left\{[Y - E(Y|x)]^2|x\right\} = E(Y^2|x) - [E(Y|x)]^2.$$

3.7.3 Example

In Example 3.4.9 find $E(Y|x)$, $E(Y^2|x)$ and $Var(Y|x)$.

3.7.4 Exercise

In Example 3.3.4 and Exercise 3.3.5 find $E(Y|x)$, $Var(Y|x)$, $E(X|y)$ and $Var(X|y)$.

3.7.5 Theorem

If X and Y are independent random variables then $E[g(Y)|x] = E[g(Y)]$ and $E[h(X)|y] = E[h(X)]$.

3.7.6 Definition

$E[g(Y)|X]$ is the function of the random variable X whose value is $E[g(Y)|x]$ when $X = x$. This means of course that $E[g(Y)|X]$ is a random variable.

3.7.7 Theorem

Suppose X and Y are random variables then

$$E\{E[g(Y)|X]\} = E[g(Y)].$$

3.7.8 Corollary

Suppose X and Y are random variables then

$$E[E(Y|X)] = E(Y).$$

3.7.9 Theorem

Suppose X and Y are random variables then

$$Var(Y) = E[Var(Y|X)] + Var[E(Y|X)].$$

3.7.10 Exercise

Prove Theorem 3.7.9.

3.7.11 Example

Suppose $P \sim \text{UNIF}(0, 0.1)$ and $Y|P = p \sim \text{BIN}(10, p)$. Find $E(Y)$ and $Var(Y)$.

3.7.12 Exercise

In Examples 3.5.6 and 3.5.7 find $E(X)$ and $Var(X)$ using Corollary 3.7.8 and Theorem 3.7.9.

3.7.13 Exercise

Suppose $P \sim \text{BETA}(3,2)$ and $Y|P = p \sim \text{GEO}(p)$. Find $E(Y)$ and $Var(Y)$.

3.8 Joint Moment Generating Functions

3.8.1 Definition

If X and Y are random variables then

$$M(t_1, t_2) = E(e^{t_1 X + t_2 Y})$$

is called the *joint m.g.f. of X and Y* if this expectation exists (joint sum/integral converges absolutely) for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for some $h_1,\ h_2 > 0$.

More generally if $X_1, X_2, \ldots, X_n$ are random variables then

$$M(t_1, t_2, \ldots, t_n) = E\left[\exp\left(\sum_{i=1}^{n} t_i X_i\right)\right]$$

is called the *joint m.g.f. of $X_1, X_2, \ldots, X_n$* if this expectation exists for all $t_i \in (-h_i, h_i)$ for some $h_i > 0$, $i = 1, \ldots, n$.

3.8.2 Important Note

If $M(t_1, t_2)$ exists for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for some $h_1,\ h_2 > 0$ then the m.g.f. of X is given by

$$M_X(t) = E(e^{tX}) = M(t, 0), \quad t \in (-h_1, h_1)$$

and the m.g.f. of Y is given by

$$M_Y(t) = E(e^{tY}) = M(0, t), \quad t \in (-h_2, h_2).$$

3.8.3 Independence Theorem for m.g.f.'s

Suppose X and Y are random variables with joint m.g.f. $M(t_1, t_2)$ which exists for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ for some $h_1,\ h_2 > 0$. Then X and Y are independent random variables if and only if

$$M(t_1, t_2) = M_X(t_1) M_Y(t_2)$$

for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ where $M_X(t_1) = M(t_1, 0)$ and $M_Y(t_2) = M(0, t_2)$.

3.8.4 Example

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x, y) = e^{-y}, \qquad 0 < x < y < \infty.$$

Find the joint m.g.f. of X and Y. Are X and Y independent random variables? What is the marginal distribution of X? What is the marginal distribution of Y?

3.8.5 Exercise

Suppose $X_1, X_2, \ldots, X_n$ are idependent and identically distributed random varibles each with m.g.f. $M(t),\ t \in (-h, h)$ for some $h > 0$. Find $M(t_1, t_2, \ldots, t_n)$ the joint m.g.f. of $X_1, X_2, \ldots, X_n$. Find the m.g.f. of $T = \sum_{i=1}^{n} X_i$.

3.9 Multinomial Distribution

3.9.1 Definition

Suppose $(X_1, \ldots, X_k)$ are discrete random variables with joint p.f.

$$f(x_1, \ldots, x_k) = \frac{n!}{x_1! x_2! \cdots x_{k+1}!}\, p_1^{x_1} p_2^{x_2} \cdots p_{k+1}^{x_{k+1}}$$

$x_i = 0, \ldots, n,\ i = 1, \ldots, k+1,\ x_{k+1} = n - \sum_{i=1}^{k} x_i,\ 0 < p_i < 1,$

$i = 1, \ldots, k+1$, and $p_{k+1} = 1 - \sum_{i=1}^{k} p_i$. Then $(X_1, \ldots, X_k)$ is said to have a *multinomial distribution*. We write $(X_1, \ldots, X_k) \sim \text{MULT}(n, p_1, \ldots, p_k)$.

3.9.2 Theorem - Properties of the Multinomial Distribution

Suppose $(X_1, \ldots, X_k) \sim \text{MULT}(n, p_1, \ldots, p_k)$, then

(1) $(X_1, \ldots, X_k)$ has joint m.g.f.

$$\begin{aligned} M(t_1, \ldots, t_k) &= E\left(e^{t_1 X_1 + \cdots + t_k X_k}\right) \\ &= (p_1 e^{t_1} + \cdots + p_k e^{t_k} + p_{k+1})^n, \quad (t_1, \ldots, t_k) \in \Re^k. \end{aligned}$$

(2) Any subset of $X_1, \ldots, X_{k+1}$ also has a multinomial distribution. In particular

$$X_i \backsim \text{BIN}(n,\ p_i), \quad i = 1, \ldots, k+1.$$

(3) If $T = X_i + X_j$, $i \neq j$, then

$$T \backsim \text{BIN}(n,\ p_i + p_j).$$

(4)

$$Cov\,(X_i, X_j) = -np_ip_j, \quad i \neq j.$$

(5) The conditional distribution of any subset of $(X_1, \ldots, X_{k+1})$ given the rest of the coordinates is a multinomial distribution. In particular the conditional p.f. of X_i given $X_j = x_j$, $i \neq j$, is

$$X_i | X_j = x_j \sim \text{BIN}\left(n - x_j,\ \frac{p_i}{1 - p_j}\right).$$

(6) The conditional distribution of X_i given $T = X_i + X_j = t$, $i \neq j$, is

$$X_i | X_i + X_j = t \sim \text{BIN}\left(t,\ \frac{p_i}{p_i + p_j}\right).$$

3.9.3 Example

Prove property (2) in Theorem 3.9.2.

3.9.4 Exercise

Prove properties (1) and (3) in Theorem 3.9.2.

3.10 Bivariate Normal Distribution

3.10.1 Definition

Suppose X_1 and X_2 are random variables with joint p.d.f.

$$f(x_1, x_2) = \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\}, \qquad (x_1, x_2) \in \Re^2$$

where

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \qquad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

and Σ is an nonsingular matrix. Then $X = (X_1, X_2)^T$ is said to have a *bivariate normal distribution*. We write $X \backsim \text{BVN}(\mu, \Sigma)$.

3.10.2 Theorem - Properties of the BVN Distribution

Suppose $X \backsim \text{BVN}(\mu, \Sigma)$, then

(1) X has joint m.g.f.

$$M(t_1, t_2) = E\left[\exp\left(t^T X\right)\right] = E\left(e^{t_1X_1+t_2X_2}\right) = \exp\left(\mu^T t + \frac{1}{2}t^T\Sigma t\right)$$

for all $(t_1, t_2) \in \Re^2$.

(2) $X_1 \sim \text{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \text{N}(\mu_2, \sigma_2^2)$.

(3) $Cov\,(X_1, X_2) = \rho\sigma_1\sigma_2$ and $Corr\,(X_1, X_2) = \rho$ where $-1 \le \rho \le 1$.

(4) X_1 and X_2 are independent random variables if and only if $\rho = 0$.

(5) If $c = (c_1, c_2)^T$ is a nonzero vector of constants then

$$c^T X = \sum_{i=1}^{2} c_i X_i \backsim \text{N}\left(c^T\mu, c^T\Sigma c\right).$$

(6) If A is a 2×2 nonsingular matrix and b is a 2×1 vector then $Y = AX + b \backsim \text{BVN}\left(A\mu + b, A\Sigma A^T\right)$.

(7)

$$X_2|X_1 = x_1 \backsim \text{N}\left(\mu_2 + \rho\sigma_2(x_1 - \mu_1)/\sigma_1,\ \sigma_2^2(1-\rho^2)\right)$$

and

$$X_1|X_2 = x_2 \backsim \text{N}\left(\mu_1 + \rho\sigma_1(x_2 - \mu_2)/\sigma_2,\ \sigma_1^2(1-\rho^2)\right).$$

(8) $(X-\mu)^T\Sigma^{-1}(X-\mu) \backsim \chi^2(2)$.

3.10.3 Example

Prove property (5) in Theorem 3.10.2.

3.10.4 Exercise

Prove property (6) in Theorem 3.10.2.

In the Figures $3.2 - 3.4$ the BVN joint p.d.f. is graphed. The graphs all have the same mean vector $\mu = [0\ 0]^T$ but different variance/covariance matrices Σ. The axes all have the same scale.

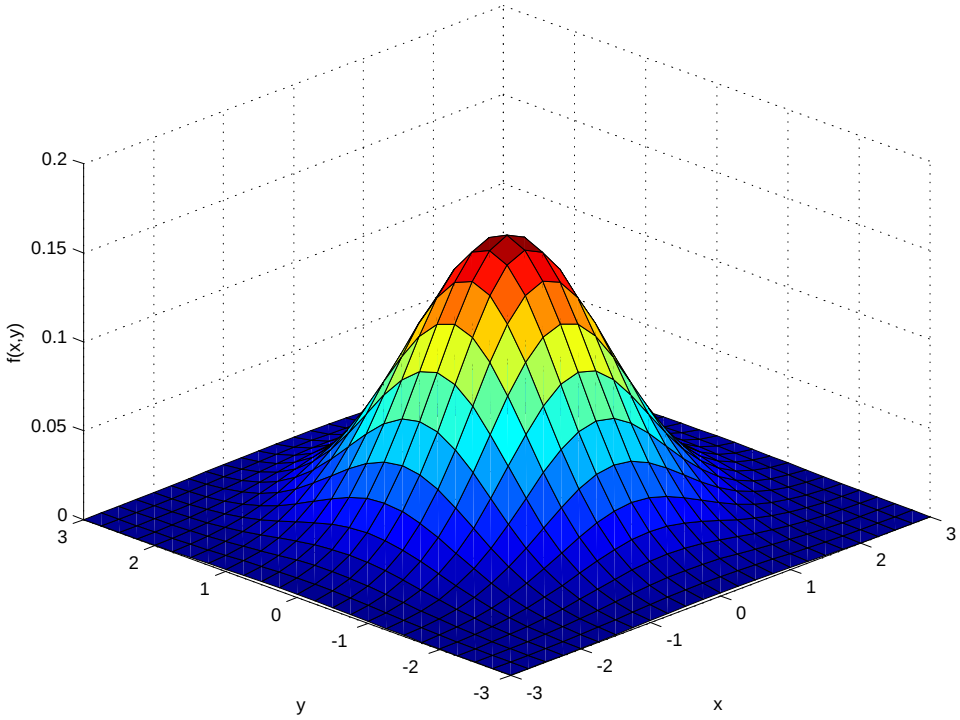
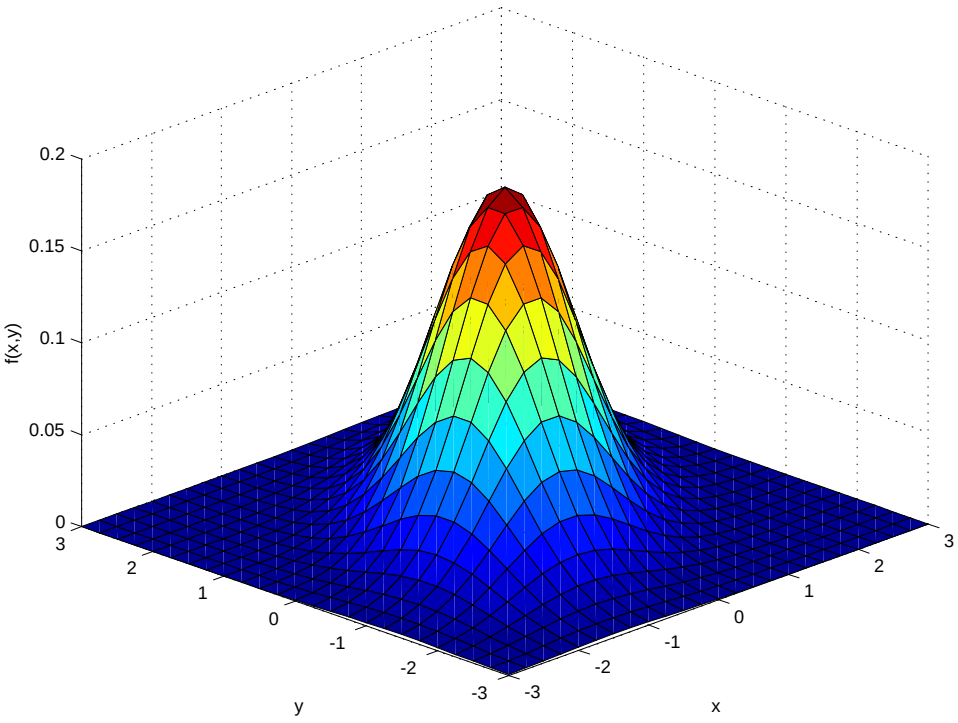


Figure 3.2:

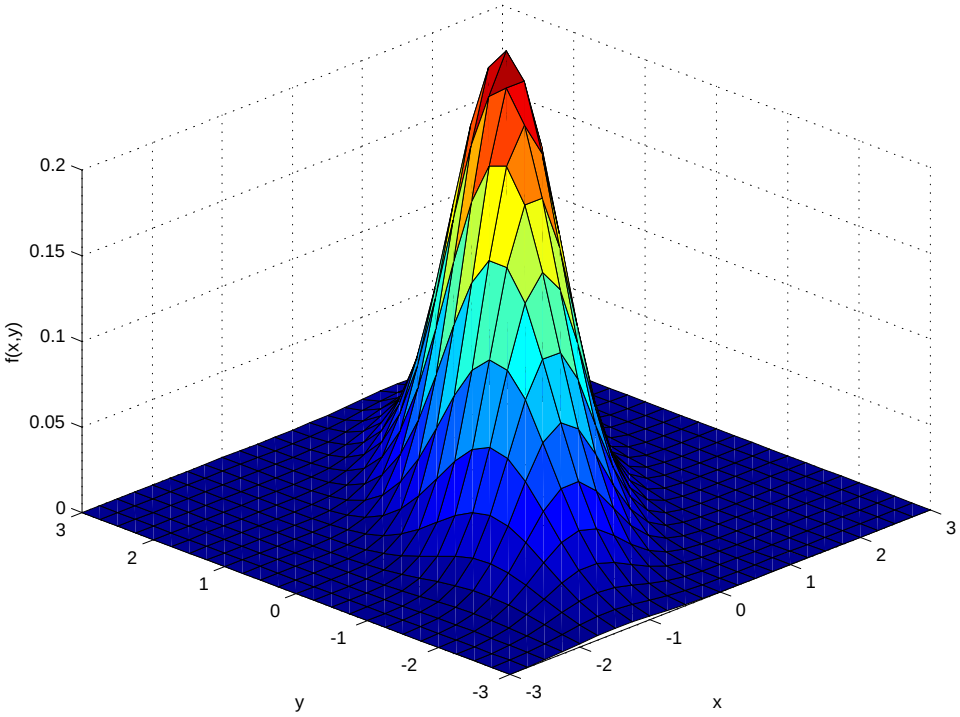
Graph of BVN p.d.f. with $\mu = [0\ 0]^T$ and $\Sigma = [1\ 0\ ;0\ 1]$.

Figure 3.3:



Graph of BVN p.d.f. with $\mu = [0\ 0]^T$ and $\Sigma = [1\ 0.5;\ 0.5\ 1]$.

Figure 3.4:



Graph of BVN p.d.f. with $\mu = [0\ 0]^T$ and $\Sigma = [0.6\ 0.5;\ 0.5\ 1]$.

3.11 Calculus Review

Consider the region R in the xy-plane in Figure 3.5. Suppose $f(x, y) \geq 0$

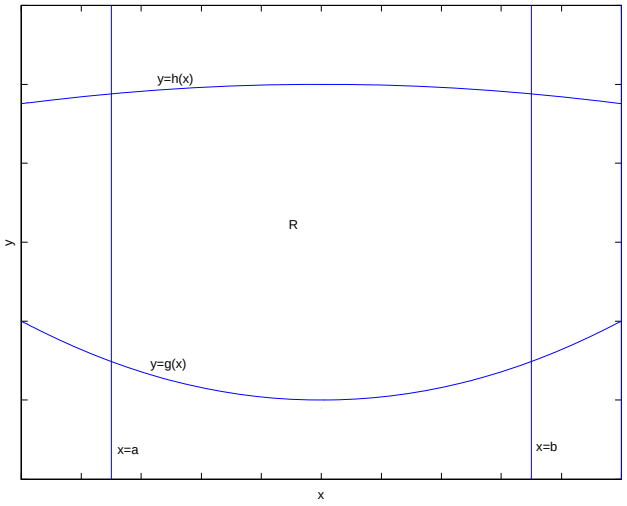


Figure 3.5:

for all $(x, y) \in \Re^2$. The graph of $z = f(x, y)$ is a surface in 3-space lying above or touching the xy-plane. The volume of the solid bounded by the surface $z = f(x, y)$ and the xy-plane above the region R is given by

$$\text{Volume} = \int_{x=a}^{b} \int_{y=g(x)}^{h(x)} f(x, y) dy dx.$$

If R is the region in Figure 3.6 then the volume is given by

$$\text{Volume} = \int_{y=c}^{d} \int_{x=g(y)}^{h(y)} f(x, y) dx dy.$$

Give an expression for the volume of the solid bounded by the surface $z = f(x, y)$ and the xy-plane above the region R in Figure 3.7.

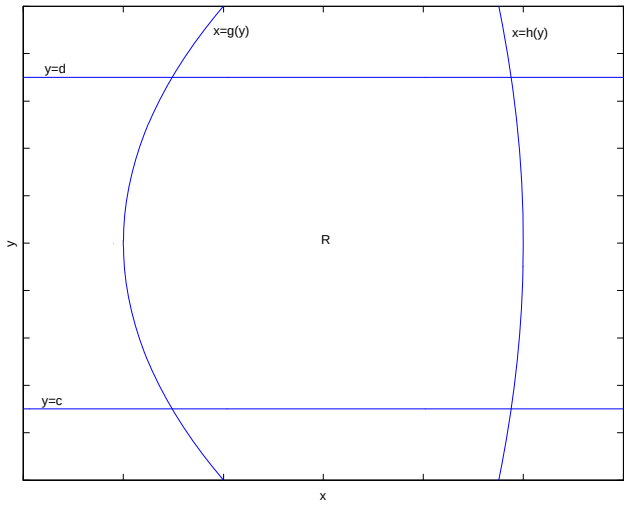


Figure 3.6:

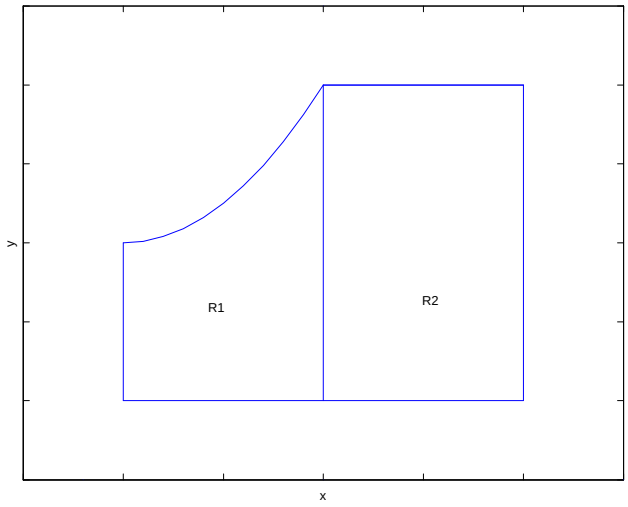


Figure 3.7:

Chapter 4

Functions of Random Variables

4.1 C.D.F. Technique

Suppose $(X_1, \ldots, X_n)$ are continuous random variables with joint p.d.f. $f(x_1, \ldots, x_n)$. We can find the p.d.f. of $Y = h(X_1, \ldots, X_n)$ using the c.d.f. technique that was used in Section 2.4 for the case $n = 1$.

4.1.1 Example

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x, y) = 3y, \quad 0 \leq x \leq y \leq 1$$

and 0 otherwise. Find the p.d.f. of $T = XY$.

4.1.2 Exercise

For the previous example show that p.d.f. of $S = Y/X$ is $g(s) = s^{-2}$, $s \geq 1$.

4.1.3 Example

Suppose $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) continuous random variables each with p.d.f. $f(x)$ and c.d.f. $F(x)$. Find the p.d.f. of $Y = \max(X_1, \ldots, X_n) = X_{(n)}$ and $T = \min(X_1, \ldots, X_n) = X_{(1)}$.

4.2 One-to-One Bivariate Transformations

Suppose the transformation S defined by

$$\begin{aligned} u &= h_1(x,y) \\ v &= h_2(x,y) \end{aligned}$$

is a one-to-one transformation for all $(x,y) \in R_{XY}$ and that S maps the region R_{XY} into the region R_{UV} in the uv-plane. Since $S:(x,y) \to (u,v)$ is a one-to-one transformation there exists a inverse transformation T defined by

$$\begin{aligned} x &= w_1(u,v) \\ y &= w_2(u,v) \end{aligned}$$

such that $T = S^{-1}:(u,v) \to (x,y)$ for all $(u,v) \in R_{UV}$. The Jacobian of the transformation T is

$$\frac{\partial(x,y)}{\partial(u,v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \left[\frac{\partial(u,v)}{\partial(x,y)}\right]^{-1}$$

where $\frac{\partial(u,v)}{\partial(x,y)}$ is the Jacobian of the transformation S.

4.2.1 Inverse Mapping Theorem

Consider the transformation S defined by

$$\begin{aligned} u &= h_1(x,y) \\ v &= h_2(x,y). \end{aligned}$$

If $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$ and $\frac{\partial v}{\partial y}$ are continuous functions and $\frac{\partial(u,v)}{\partial(x,y)} \neq 0$ for all $(x,y) \in R$ then S is one-to-one on R and S^{-1} exists.

Note: These are sufficient but not necessary conditions for the inverse to exist.

4.2.2 Theorem - One-to-One Bivariate Transformations

Let X and Y be continuous random variables with joint p.d.f. $f(x, y)$ and let $R_{XY} = \{(x, y) : f(x, y) > 0\}$ be the support set of (X, Y). Suppose the transformation S defined by

$$\begin{aligned} U &= h_1(X, Y) \\ V &= h_2(X, Y) \end{aligned}$$

is a one-to-one transformation with inverse transformation

$$\begin{aligned} X &= w_1(U, V) \\ Y &= w_2(U, V). \end{aligned}$$

Suppose also that S maps R_{XY} into R_{UV}. Then $g(u, v)$, the joint joint p.d.f. of U and V, is given by

$$g(u, v) = f(w_1(u, v), w_2(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|$$

for all $(u, v) \in R_{UV}$. (Compare Theorem 2.4.4 for univariate random variables.)

4.2.3 Proof

We want to find $g(u, v)$, the joint p.d.f. of the random variables U and V. Suppose S^{-1} maps the region $B \subset R_{UV}$ into the region $A \subset R_{XY}$ then

$$\begin{aligned} P[(U, V) &\in B] \\ &= \iint_B g(u, v) du dv \qquad (4.1) \\ &= P[(X, Y) \in A] \\ &= \iint_A f(x, y) dx dy \\ &= \iint_B f(w_1(u, v), w_2(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du dv \qquad (4.2) \end{aligned}$$

where the last line follows by the Change of Variable Theorem. Since this is true for all $B \subset R_{UV}$ we have, by comparing (4.1) and (4.2), that the joint p.d.f. of U and V is given by

$$g(u, v) = f(w_1(u, v), w_2(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right|$$

for all $(u, v) \in R_{UV}$.

4.2.4 Example

Suppose $X \sim \text{GAM}(a, 1)$ and $Y \sim \text{GAM}(b, 1)$ independently. Find the joint p.d.f. of $U = X + Y$ and $V = \frac{X}{X+Y}$. Show that $U \backsim \text{GAM}(a + b, 1)$ and $V \sim \text{BETA}(a, b)$ independently. Find $E(V)$. See Figure 4.1.

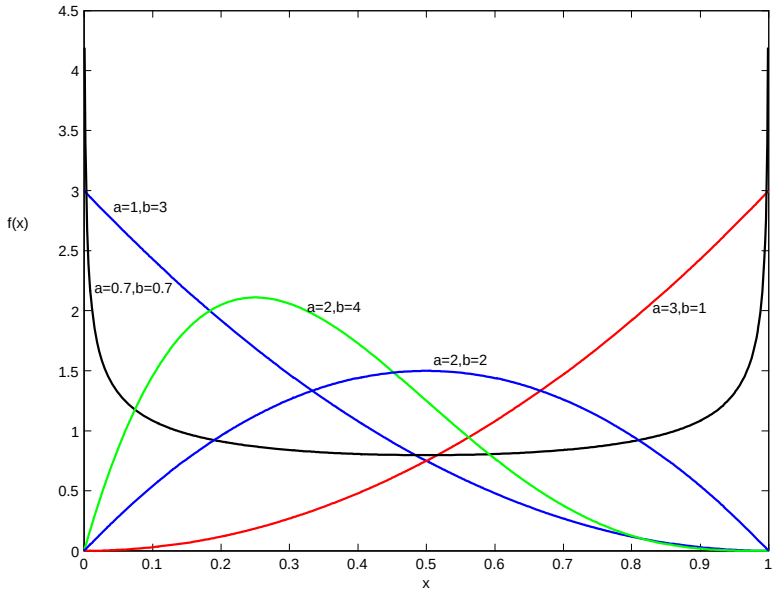


Figure 4.1: BETA(a, b) p.d.f.'s

4.2.5 Exercise

Suppose $X \sim \text{BETA}(a, b)$ and $Y \sim \text{BETA}(a+b, c)$ independently. Find the joint p.d.f. of $U = XY$ and $V = X$. Show that $U \sim \text{BETA}(a, b + c)$.

4.2.6 Example - Box-Mueller Transformation

Suppose $X \sim \text{UNIF}(0, 1)$ and $Y \sim \text{UNIF}(0, 1)$ independently. Find the joint p.d.f. of

$$\begin{aligned} U &= (-2 \log X)^{1/2} \cos(2\pi Y) \\ V &= (-2 \log X)^{1/2} \sin(2\pi Y). \end{aligned}$$

Explain how you could use this result to generate independent observations from a N(0, 1) distribution.

4.2.7 Exercise

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x,y) = e^{-x-y}, \qquad 0 < x < \infty,\ 0 < y < \infty.$$

Let $U = X + Y$ and $V = X$. Find the joint p.d.f. of U and V and the marginal p.d.f. of U. What is the distribution of U?

4.2.8 Exercise

Suppose X and Y are continuous random variables with joint p.d.f.

$$f(x,y) = e^{-x-y}, \qquad 0 < x < \infty,\ 0 < y < \infty.$$

Let $U = X + Y$ and $V = X - Y$. Find the joint p.d.f. of U and V. Be sure to specify the support of (U, V). Show that $U \sim \text{GAM}(2, 1)$ and $V \sim \text{DE}(1, 0)$.

4.2.9 Theorem

If $Z \sim \text{N}(0, 1)$ independently of $X \sim \chi^2(n)$ then

$$T = \frac{Z}{\sqrt{X/n}} \sim \text{t}(n).$$

4.2.10 Theorem

If $X \sim \chi^2(n)$ independently of $Y \sim \chi^2(n)$ then

$$U = \frac{X/n}{Y/m} \sim \text{F}(n, m).$$

4.2.11 Exercise

(a) Prove Theorem 4.2.10. (Hint: Complete the transformation with $V = Y$.)

(b) Find $E(U)$ and $Var(U)$ and note for what values of n and m that these exist. (Hint: Since X and Y are independent random variables $E(X/Y) = E(X) \cdot E(Y^{-1})$.)

4.3 Moment Generating Function Method

This method is particularly useful in finding distributions of sums of independent random variables.

4.3.1 Theorem

Suppose $X_1, \ldots, X_n$ are independent random variables and X_i has m.g.f. $M_i(t)$ which exists for $t \in (-h, h)$ for some $h > 0$. The m.g.f. of $Y = \sum_{i=1}^{n} X_i$ is given by

$$M_Y(t) = \prod_{i=1}^{n} M_i(t)$$

for $t \in (-h, h)$.

Notes:

(1) If the X_i's are i.i.d. random variables each with m.g.f. $M(t)$ then Y has m.g.f.

$$M_Y(t) = [M(t)]^n$$

for $t \in (-h, h)$.

(2) This theorem in conjunction with the Uniqueness Theorem for m.g.f.'s can be used to find the distribution of Y.

4.3.2 Special Results

(1) If $X \sim \text{GAM}(\alpha, \beta)$, where α is a positive integer, then $\frac{2X}{\beta} \sim \chi^2(2\alpha)$.

(2) If $X_i \sim \text{GAM}(\alpha_i, \beta)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} X_i \sim \text{GAM}\left(\sum_{i=1}^{n} \alpha_i,\ \beta\right).$$

(3) If $X_i \sim \text{GAM}(1, \beta) = \text{EXP}(\beta)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} X_i \sim \text{GAM}(n, \beta).$$

(4) If $X_i \sim \text{GAM}(\frac{k_i}{2}, 2) = \chi^2(k_i)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} X_i \sim \chi^2\left(\sum_{i=1}^{n} k_i\right).$$

(5) If $X_i \sim \mathrm{N}(\mu, \sigma^2)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n).$$

(6) If $X_i \sim \mathrm{POI}(\mu_i)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} X_i \sim \mathrm{POI}\left(\sum_{i=1}^{n} \mu_i \right).$$

(7) If $X_i \sim \mathrm{BIN}(n_i, p)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} X_i \sim \mathrm{BIN}\left(\sum_{i=1}^{n} n_i,\ p \right).$$

(8) If $X_i \sim \mathrm{NB}(k_i, p)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} X_i \sim \mathrm{NB}\left(\sum_{i=1}^{n} k_i,\ p \right).$$

4.3.3 Exercise

Prove (1) – (8) in 4.3.2.

4.3.4 Exercise

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with m.g.f. $M(t)$, $E(X_i) = \mu$, and $Var(X_i) = \sigma^2 < \infty$. Find the m.g.f. of $Z = \sqrt{n}\left(\bar{X} - \mu\right)/\sigma$ where $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$.

4.3.5 Theorem

If $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$ independently, then

$$\sum_{i=1}^{n} a_i X_i \sim \mathrm{N}\left(\sum_{i=1}^{n} a_i \mu_i,\ \sum_{i=1}^{n} a_i^2 \sigma_i^2 \right).$$

4.3.6 Corollary

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Then

$$\sum_{i=1}^{n} X_i \sim \mathrm{N}(n\mu, n\sigma^2) \quad \text{and} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \sim \mathrm{N}\left(\mu,\ \frac{\sigma^2}{n} \right).$$

4.3.7 Useful Identity

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 + n\left(\bar{X} - \mu\right)^2$$

4.3.8 Theorem

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Then

$$\bar{X} \sim \mathrm{N}\left(\mu,\ \frac{\sigma^2}{n}\right)$$

independently of

$$\frac{(n-1)\,S^2}{\sigma^2} = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{\sigma^2} \sim \chi^2(n-1)$$

where

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}.$$

4.3.9 Theorem

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \mathrm{t}(n-1).$$

4.3.10 Theorem

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu_1, \sigma_1^2)$ distribution and independently $Y_1, \ldots, Y_m$ is a random sample from the $\mathrm{N}(\mu_2, \sigma_2^2)$ distribution. Let

$$S_1^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1) \quad \text{and} \quad S_2^2 = \sum_{i=1}^{m} (Y_i - \bar{Y})^2/(m-1).$$

Then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \backsim \mathrm{F}(n-1, m-1).$$

4.3.11 Exerecise

Prove Theorem 4.3.10. Hint: Use Theorems 4.2.10 and 4.3.8.

Chapter 5

Limiting or Asymptotic Distributions

5.1 Convergence in Distribution

5.1.1 Definition - Convergence in Distribution

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables such that X_n has c.d.f. $F_n(x)$. Let X be a random variable with c.d.f. $F(x)$. We say X_n *converges in distribution* to X and we write

$$X_n \rightarrow_D X$$

if

$$\lim_{n\rightarrow\infty} F_n(x) = F(x)$$

at all points x at which $F(x)$ is continuous. We call F the *limiting or asymptotic distribution* of X_n.

Note:

(1) Although we talk of a sequence of random variables converging in distribution, it is really the c.d.f.'s that converge, not the random variables.

(2) This definition holds for both discrete and continuous random variables.

5.1.2 Theorem - e Limit

If b and c are real constants and $\lim_{n\rightarrow\infty} \psi(n) = 0$ then

$$\lim_{n\rightarrow\infty} \left[1 + \frac{b}{n} + \frac{\psi(n)}{n}\right]^{cn} = e^{bc}.$$

5.1.3 Corollary

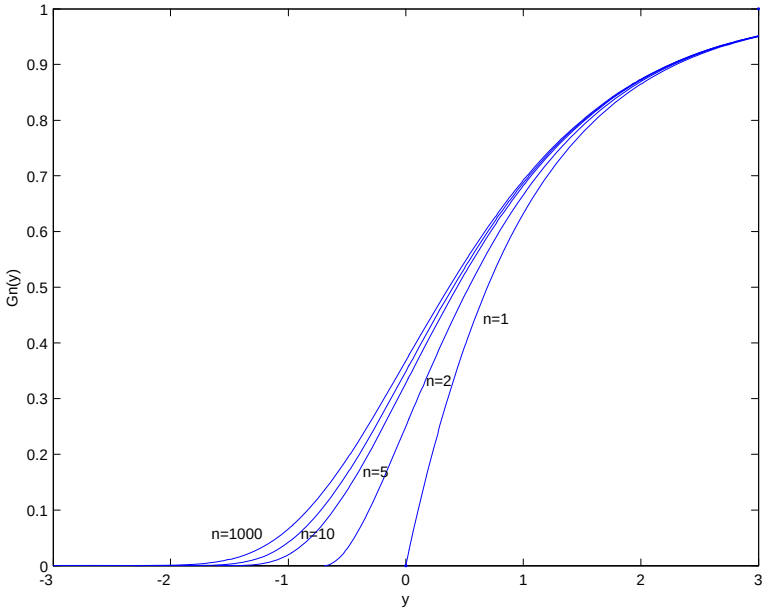
If b and c are real constants then

$$\lim_{n\to\infty}\left(1+\frac{b}{n}\right)^{cn}=e^{bc}.$$

5.1.4 Example

Suppose $X_i \backsim \text{EXP}(1)$, $i=1,2,\ldots$ independently. Consider the sequence of random variables $Y_1, Y_2, \ldots, Y_n, \ldots$ where $Y_n = \max(X_1,\ldots,X_n) - \log n$. Find the limiting distribution of Y_n. See Figure 5.1.

Figure 5.1:

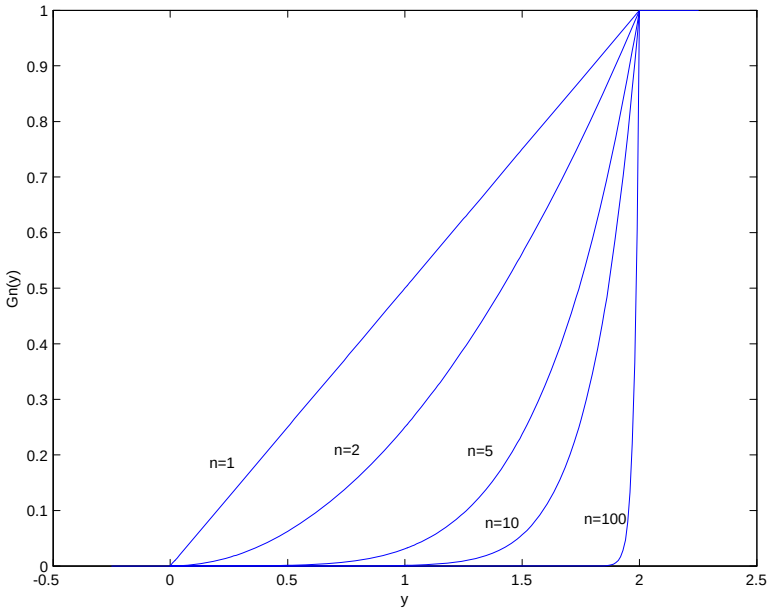


Graphs of $G_n(y) = \left(1-\frac{e^{-y}}{n}\right)^n$ for $n = 1, 2, 5, 10, 1000$.

5.1.5 Example

Suppose $X_i \backsim \text{UNIF}(0, \theta)$, $i = 1, 2, \ldots$ independently. Consider the sequence of random variables $Y_1, Y_2, \ldots, Y_n, \ldots$ where $Y_n = \max(X_1, \ldots, X_n)$. Find the limiting distribution of Y_n. See Figure 5.2.

Figure 5.2:



Graphs of $G_n(y) = \left(\frac{y}{\theta}\right)^n$ for $\theta = 2$ and $n = 1, 2, 5, 100$.

5.2 Convergence in Probability

5.2.1 Definition - Convergence in Probability

A sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ *converges in probability* to a random variable X if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \varepsilon\right) = 0$$

or equivalently

$$\lim_{n\to\infty} P\left(|X_n - X| < \varepsilon\right) = 1.$$

We write

$$X_n \to_p X.$$

5.2.2 Theorem - Convergence in Probability Implies Convergence in Distribution

If $X_n \to_p X$ then $X_n \to_D X$.

Many of the examples we consider involve convergence in probability to a constant.

5.2.3 Definition - Convergence in Probability to a Constant

A sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$ *converges in probability* to a constant b if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} P\left(|X_n - b| \geq \varepsilon\right) = 0$$

or equivalently

$$\lim_{n\to\infty} P\left(|X_n - b| < \varepsilon\right) = 1.$$

We write

$$X_n \to_p b.$$

The following theorem is one way to prove convergence in probability to a constant.

5.2.4 Theorem

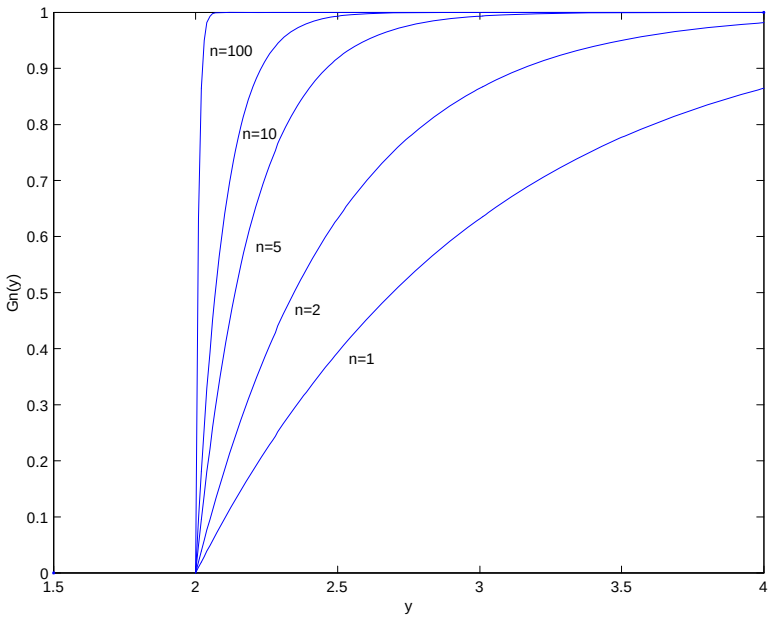
Suppose $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of random variables such that X_n has c.d.f. $F_n(x)$. If

$$\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} P(X_n \leq x) = \begin{cases} 0 & x < b \\ 1 & x > b \end{cases}$$

then $X_n \to_p b$.

Note:
We do not need to worry about whether $\lim_{n\to\infty} F_n(b)$ exists since $x = b$ is a point of discontinuity of the limiting distribution (see Definition 5.1.1).

Figure 5.3:



Graphs of $G_n(y) = 1 - e^{-n(y-\theta)}$ for $\theta = 2$ and $n = 1, 2, 5, 10, 100$.

5.2.5 Example

Suppose $X_i \backsim \text{EXP}(1, \theta)$, $i = 1, 2, \ldots$ independently. Consider the sequence of random variables $Y_1, Y_2, \ldots, Y_n, \ldots$ where $Y = \min(X_1, \ldots, X_n)$. Show that $Y_n = \min(X_1, \ldots, X_n) \rightarrow_p \theta$. See Figure 5.3.

5.2.6 Comment

Suppose $Y_n \rightarrow_D Y$. Then for large n we can use the approximation

$$P(Y_n \leq y) \approx P(Y \leq y).$$

If Y is degenerate at b then $P(Y = b) = 1$ and this approximation is not useful. However, if $Y_n \rightarrow_p b$ then we would use this result in another way. For example, if $X_1, \ldots, X_n$ is a random sample from the UNIF$(0, \theta)$ distribution then $Y_n = \max(X_1, \ldots, X_n) \rightarrow_p \theta$. If θ were unknown, this result suggests that we could use $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n) = \max(X_1, \ldots, X_n)$ as an estimator of θ. (**Note:** $\max(x_1, \ldots, x_n)$ is an estimate of θ.)

5.3 Limit Theorems

5.3.1 Limit Theorem for m.g.f.'s

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables such that X_n has m.g.f. $M_n(t)$ and let X be a random variable with m.g.f. $M(t)$. If there exists an $h > 0$ such that

$$\lim_{n \to \infty} M_n(t) = M(t)$$

for all $t \in (-h, h)$ then $X_n \rightarrow_D X$.

5.3.2 Example

Suppose $Y_k \sim \text{NB}(k, p)$. Find the limiting distribution of Y_k as $k \to \infty$, $p \to 1$ such that $kq/p = \mu$ remains constant where $q = 1 - p$.

5.3.3 Exercise - Poisson Approximation to the Binomial Distribution

Suppose $Y_n \sim \text{BIN}(n, p)$. Find the limiting distribution of Y_n as $n \to \infty$, $p \to 0$ such that $np = \mu$ remains constant.

5.3.4 Theorem

If $X_n \rightarrow_D X$ and X_n and X are non-negative integer-valued random variables then $\lim_{n\rightarrow\infty} P(X_n \leq x) = P(X \leq x)$ holds for all x.

Also $\lim_{n\rightarrow\infty} P(X_n = x) = P(X = x)$ holds for $x = 0, 1, \ldots$.

5.3.5 Central Limit Theorem

Suppose $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent random variables with $E(X_i) = \mu$ and
$Var(X_i) = \sigma^2 < \infty$. Then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow_D Z \backsim \mathrm{N}(0, 1)$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

5.3.6 Example - Normal Approximation to the χ^2 Distribution:

If $Y_n \sim \chi^2(n)$ then show

$$\frac{Y_n - n}{\sqrt{2n}} \rightarrow_D Z \sim \mathrm{N}(0, 1).$$

5.3.7 Exercise - Normal Approximation to the Binomial Distribution

If $Y_n \sim \mathrm{BIN}(n, p)$ then show

$$\frac{Y_n - np}{\sqrt{np(1-p)}} \rightarrow_D Z \sim \mathrm{N}(0, 1).$$

Hint: Let $X_i \sim \mathrm{BIN}(1, p)$, $i = 1, \ldots, n$ independently. Find the distribution of $\sum_{i=1}^{n} X_i$ and the limiting distribution of $\sum_{i=1}^{n} X_i$.

5.3.8 Weak Law of Large Numbers (WLLN)

Suppose $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent random variables with $E(X_i) = \mu$ and
$Var(X_i) = \sigma^2 < \infty$. Then

$$\bar{X}_n \rightarrow_p \mu.$$

5.3.9 Limit Theorems

(1) If $X_n \to_p a$ and g is continuous at $x = a$ then $g(X_n) \to_p g(a)$.

(2) If $X_n \to_p a$, $Y_n \to_p b$ and $g(x, y)$ is continuous at (a, b) then $g(X_n, Y_n) \to_p g(a, b)$.

(3) (Slutsky) If $X_n \to_D X$, $Y_n \to_p b$ and $g(x, b)$ is continuous for all $x \in$ support set of X then $g(X_n, Y_n) \to_D g(X, b)$.

5.3.10 Example

If $X_n \to_p a > 0$, $Y_n \to_p b \neq 0$ and $Z_n \to_D Z \sim \mathrm{N}(0, 1)$ then find the limiting distributions of each of the following:

$$(1)\ X_n^2 \quad (2)\ \sqrt{X_n} \quad (3)\ X_n Y_n \quad (4)\ X_n + Y_n \quad (5)\ X_n / Y_n$$
$$(6)\ 2Z_n \quad (7)\ Z_n + Y_n \quad (8)\ X_n Z_n \quad (9)\ Z_n^2 \quad (10)\ 1/Z_n$$

5.3.11 Example

Let $X_1, X_2, \ldots, X_n, \ldots$ is a sequence of independent POI(μ) random variables. Find the limiting distribution of $Z_n = \sqrt{n}\left(\bar{X}_n - \mu\right) / \sqrt{\bar{X}_n}$.

5.3.12 Exercise

Suppose $X_i \backsim \mathrm{UNIF}(0, 1)$, $i = 1, 2, \ldots$ independently. Consider the sequence of random variables $U_1, U_2, \ldots, U_n, \ldots$ where $U_n = \max(X_1, \ldots, X_n)$. Show

$$U_n \to_p 1, \qquad n\,(1 - U_n) \to_D X \sim \mathrm{EXP}(1),$$
$$e^{U_n} \to_p e, \qquad \sin(1 - U_n) \to_p 0,$$
$$e^{-n(1-U_n)} \to_D e^{-X} \sim \mathrm{UNIF}(0, 1),$$
$$\text{and} \quad (U_n + 1)^2\,[n\,(1 - U_n)] \to_D 4X \sim \mathrm{EXP}(4).$$

5.3.13 Delta Method (∂-Method)

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables such that

$$n^b (X_n - a) \to_D X$$

for some $b > 0$. Suppose the function $g(x)$ is differentiable at a and $g'(a) \neq 0$. Then

$$n^b [g(X_n) - g(a)] \to_D g'(a) X.$$

5.3.14 Example

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent EXP(θ) random variables. Find the limiting distribution of $\bar{X}_n$,

$$\begin{aligned} Z_n &= \sqrt{n}\left(\bar{X}_n - \theta\right)/\bar{X}_n, \\ U_n &= \sqrt{n}\left(\bar{X}_n - \theta\right) \\ \text{and } V_n &= \sqrt{n}\left(\log(\bar{X}_n) - \log\theta\right). \end{aligned}$$

5.3.15 Exercise

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independent POI(θ) random variables. Show that

$$\begin{aligned} U_n &= \sqrt{n}\left(\bar{X}_n - \theta\right) \to_D U \sim \mathrm{N}(0, \theta) \\ \text{and } V_n &= \sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\theta}\right) \to_D V \sim \mathrm{N}\left(0, \frac{1}{4}\right). \end{aligned}$$

5.3.16 Theorem

Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of random variables such that

$$\sqrt{n}(X_n - a) \to_D X \sim \mathrm{N}\left(0, \sigma^2\right).$$

Suppose the function $g(x)$ is differentiable at a and $g'(a) \neq 0$. Then

$$\sqrt{n}[g(X_n) - g(a)] \to_D W \sim \mathrm{N}\left(0,\ \left[g'(a)\right]^2 \sigma^2\right).$$

5.3.17 Exercise

Prove Theorem 5.3.16. Hint: Use the ∂-method.

Chapter 6

Estimation

6.1 Introduction

Suppose $X_1, \ldots, X_n$ is a random sample, that is, $X_1, \ldots, X_n$ are i.i.d. random variables, from the distribution with p.f./p.d.f. $f(x;\theta)$. Suppose also that θ is unknown and $\theta \in \Omega$ where Ω is the parameter space or the set of possible values of θ. Note that θ could be a vector, $\theta = (\theta_1, \ldots, \theta_k)^T$. We are interested in making inferences about the unknown parameter θ, that is, we want to find estimators (point and interval) of θ and we want to test hypotheses about θ. The joint distribution of $X_1, \ldots, X_n$ is given by

$$\prod_{i=1}^{n} f(x_i;\theta).$$

We will sometimes denote the data more compactly by the random vector $X = (X_1, \ldots, X_n)$.

6.1.1 Definition

A *statistic*, $T = T(X) = T(X_1, \ldots, X_n)$, is a function of the data which does not depend on any unknown parameter(s).

Suppose $X_1, \ldots, X_n$ is a random sample from a distribution with $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ where μ and σ^2 are unknown. The sample mean $\bar{X}$ and the sample variance S^2 are statistics while $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is not a statistic.

6.1.2 Definition

A statistic $T = T(X) = T(X_1, \ldots, X_n)$ that is used to estimate $\tau(\theta)$, a function of θ, is called an *estimator* of $\tau(\theta)$ and an observed value of the statistic $t = t(x) = t(x_1, \ldots, x_n)$ is called an *estimate* of $\tau(\theta)$.

Suppose $X_1, \ldots, X_n$ are i.i.d. random variables with $E(X_i) = \mu$. The random variable $\bar{X}$ is an estimator of μ. For a given set of observations $x_1, \ldots, x_n$, the number $\bar{x}$ is an estimate of μ.

6.2 Maximum Likelihood Method - One Parameter

Suppose X is discrete random variable with probability function $P(X = x; \theta) = f(x; \theta)$, $\theta \in \Omega$ where the scalar parameter θ is unknown. Suppose x is an observed value of the random variable X. Then the probability of observing this value is, $P(X = x; \theta) = f(x; \theta)$. With the observed value of x substituted into $f(x; \theta)$ we have a function of the parameter θ only, referred to as the *likelihood function* and denoted $L(\theta)$. In the absence of any other information, it seems logical that we should estimate the parameter θ using a value most compatible with the data. For example we might choose the value of θ which maximizes the probability of the observed data or equivalently the value of θ which maximizes the likelihood function $L(\theta)$.

6.2.1 Definition

Suppose X is a random variable with p.f. $f(x; \theta)$, where θ is a scalar and $\theta \in \Omega$. If x is the observed data, then the *likelihood function* for θ based on x is

$$\begin{aligned} L(\theta) &= L(\theta; x) \\ &= P(\text{observing the data } x; \theta) \\ &= P(X = x; \theta) \\ &= f(x; \theta), \quad \theta \in \Omega. \end{aligned}$$

Suppose $X_1, \ldots, X_n$ is a random sample from a distribution with p.f. $f(x;\theta)$ and $x_1, \ldots, x_n$ are the observed data. The *likelihood function* for θ based on $x_1, \ldots, x_n$ is

$$\begin{aligned} L(\theta) &= L(\theta; x_1, \ldots, x_n) \\ &= P(\text{observing the data } x_1, \ldots, x_n; \theta) \\ &= P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n; \theta) \\ &= \prod_{i=1}^{n} f(x_i;\theta), \quad \theta \in \Omega. \end{aligned}$$

The value of θ which maximizes the likelihood $L(\theta)$ also maximizes the logarithm of the likelihood function. (Why?) Since it is easier to find the derivative of the sum of n terms rather than the product, it is usually easier to determine the maximum of the logarithm of the likelihood function.

6.2.2 Definition

The *log likelihood function* is defined as

$$l(\theta) = \log L(\theta), \quad \theta \in \Omega$$

where log is the natural logarithmic function.

6.2.3 Definition

The value of θ that maximizes the likelihood function $L(\theta)$ or equivalently the log likelihood function $l(\theta)$ is called the *maximum likelihood* (M.L.) *estimate*. The M.L. estimate is a function of the data x and we write $\hat{\theta} = \hat{\theta}(x)$. The corresponding M.L. estimator is denoted $\hat{\theta} = \hat{\theta}(X)$.

6.2.4 Example

Suppose in a sequence of n Bernoulli trials the probability of success is equal to θ and we have observed x successes. Find the likelihood function, the log likelihood function, the M.L. estimate of θ and the M.L. estimator of θ.

6.2.5 Example

Suppose we have collected data $x_1, \ldots, x_n$ and we believe these observations are independent observations from a POI(θ) distribution. Find the likelihood function, the log likelihood function, the M.L. estimate of θ and the M.L. estimator of θ.

6.2.6 Exercise

Suppose we have collected data $x_1, \ldots, x_n$ and we believe these observations are independent observations from the DU(θ) distribution. Find the likelihood function, the M.L. estimate of θ and the M.L. estimator of θ.

6.2.7 Definition

The *score function* is defined as

$$S(\theta) = S(\theta; x) = \frac{d}{d\theta} l(\theta) = \frac{d}{d\theta} \log L(\theta), \quad \theta \in \Omega.$$

6.2.8 Definition

The *information function* is defined as

$$I(\theta) = I(\theta; x) = -\frac{d^2}{d\theta^2} l(\theta) = -\frac{d^2}{d\theta^2} \log L(\theta), \quad \theta \in \Omega.$$

$I(\hat{\theta})$ is called the observed information.

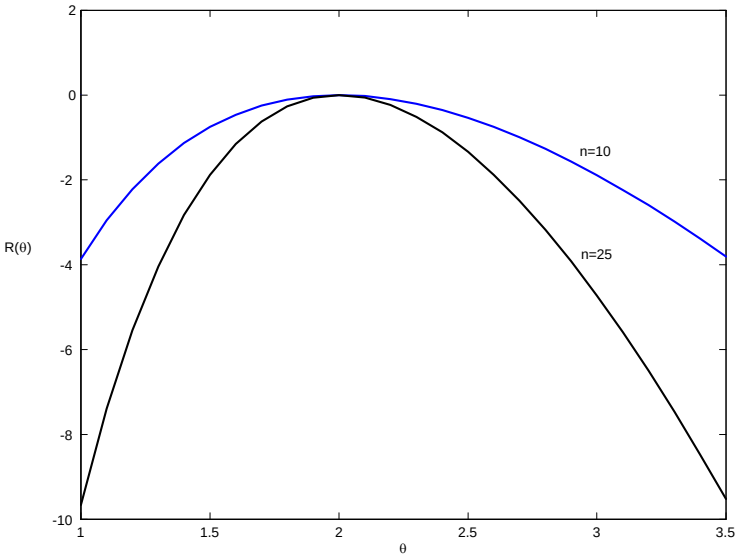


Figure 6.1: Poisson Log Likelihoods for $n = 10$ and $n = 25$

In Section 6.5 we will see how the observed information $I(\hat{\theta})$ can be used to construct approximate confidence intervals for the unknown parameter θ. $I(\hat{\theta})$ also tells us about the concavity of the log likelihood function.

Suppose in Example 6.2.5 the M.L. estimate of θ was $\hat{\theta} = 2$. If $n = 10$ then $I(\hat{\theta}) = I(2) = n/\hat{\theta} = 10/2 = 5$. If $n = 25$ then $I(\hat{\theta}) = I(2) = 25/\hat{\theta} = 25/2 = 12.5$. See Figure 6.1. The log likelihood function is more concave down for $n = 25$ than for $n = 10$ which reflects the fact that as the number of observations increases we have more "information" about the unknown parameter θ.

Although we view the likelihood, log likelihood, score and information functions as functions of θ they are, of course, also functions of the observed data x. When it is important to emphasize the dependence on the data x we will write $L(\theta; x)$, $S(\theta; x)$, and $I(\theta; x)$. Also when we wish to determine the sampling properties of these functions as functions of the random variable X we will write $L(\theta; X)$, $S(\theta; X)$, and $I(\theta; X)$.

6.2.9 Definition

If θ is a scalar then the *expected* or *Fisher information (function)* is given by

$$J(\theta) = E[I(\theta; X)] = E\left[-\frac{\partial^2}{\partial\theta^2} l(\theta; X)\right] \qquad \theta \in \Omega.$$

Note:
If $X_1, \ldots, X_n$ is a random sample from $f(x; \theta)$ then

$$J(\theta) = E\left[-\frac{\partial^2}{\partial\theta^2} l(\theta; X)\right] = nE\left[-\frac{\partial^2}{\partial\theta^2} \log f(X; \theta)\right]$$

where X has p.d.f. $f(x; \theta)$.

6.2.10 Example

In Examples 6.2.4 and 6.2.5 find the Fisher information and compare it with the variance of the M.L. estimator of θ.

6.2.11 Likelihood Functions for Continuous Models

Suppose X is a continuous random variable with probability density function $f(x;\theta)$. We will often observe only the value of X rounded to some degree of precision (say one decimal place) in which case the actual observation is a discrete random variable. For example, suppose we observe X correct to one decimal place. Then

$$P(we \ \ observe \ \ 1.1\ ;\theta) = \int_{1.05}^{1.15} f(x;\theta)dx \approx (0.1)f(1.1;\theta)$$

assuming the function $f(x;\theta)$ is quite smooth over the interval. More generally, if we observe X rounded to the nearest Δ (assumed small) then the likelihood of the observation is approximately $\Delta f(observation;\theta)$. Since the precision Δ of the observation does not depend on the parameter, then maximizing the discrete likelihood of the observation is essentially equivalent to maximizing the probability density function $f(observation;\theta)$ over the parameter. This partially justifies the use of the probability density function in the continuous case as the likelihood function.

6.2.12 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the distribution with p.d.f.

$$f(x;\theta) = \theta x^{\theta-1}, \quad 0 \leq x \leq 1, \quad \theta > 0.$$

Find the score function, the M.L. estimator of θ, the information function and the observed information.

6.2.13 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the UNIF$(0,\theta)$ distribution. Find the M.L. estimator of θ.

6.2.14 Finding M.L. Estimates

If $X_1, \ldots, X_n$ is a random sample from a distribution whose support set *does not* depend on θ then we usually find $\hat{\theta}$ by solving $S(\theta) = 0$. It is important to verify that $\hat{\theta}$ is the value of θ which maximizes $L(\theta)$ or equivalently $l(\theta)$. This can be done using the First Derivative Test. Note that the condition $I(\hat{\theta}) > 0$ only checks for a local maximum.

Often $S(\theta) = 0$ must be solved numerically using an iterative method such as *Newton's Method*.

6.2.15 Newton's Method

Let $\theta^{(0)}$ be an initial estimate of θ. The estimate $\theta^{(i)}$ can be updated using

$$\theta^{(i+1)} = \theta^{(i)} + \frac{S(\theta^{(i)})}{I(\theta^{(i)})}, \quad \text{for } i = 0, 1, \ldots$$

Notes:

(1) The initial estimate, $\theta^{(0)}$, may be determined by graphing $L(\theta)$ or $l(\theta)$.

(2) The algorithm is usually run until the value of $\theta^{(i)}$ no longer changes to a reasonable number of decimal places. When the algorithm is stopped it is always important to check that the value of θ obtained does indeed maximize $L(\theta)$.

(3) This algorithm is also called the Newton-Raphson Method.

(4) $I(\theta)$ can be replaced by $J(\theta)$ for a similar algorithm which is called the method of scoring or Fisher's method of scoring.

(5) The value of $\hat{\theta}$ may also be found by maximizing $L(\theta)$ or $l(\theta)$ using the maximization (minimization) routines available in various statistical software packages such as Maple, S-Plus, Matlab, R etc.

(6) If the support of X depends on θ (e.g. UNIF$(0, \theta)$) then $\hat{\theta}$ is not found by solving $S(\theta) = 0$.

6.2.16 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the WEI$(1, \theta)$ distribution. Explain how you would find the M.L. estimate of θ using Newton's Method.

6.2.17 Theorem - Invariance of the M.L. Estimator

Suppose $\tau = h(\theta)$ is a one-to-one function of θ. Suppose also that $\hat{\theta}$ is the M.L. estimator of θ. Then $\hat{\tau} = h(\hat{\theta})$ is the M.L. estimator of τ.

Note: The invariance property of the M.L. estimator means that if we know the M.L estimator of θ then we know the M.L. estimator of any one-to-one function of θ. This property is one reason why this estimator is so widely used.

6.2.18 Example

In Example 6.2.12 find the M.L. estimator of the median of the distribution.

6.2.19 Exercise

Suppose $X_1, \ldots, X_n$ is a random sample from the EXP(θ) distribution. Find the score function, the M.L. estimator of θ, the information function, the observed information and the Fisher information. Find the M.L. estimator of $Var\,(X_i)$.

6.2.20 Definition

The *relative likelihood function* $R(\theta)$ is defined by

$$R(\theta) = R\,(\theta; x) = \frac{L(\theta)}{L(\hat{\theta})}, \qquad \theta \in \Omega.$$

The relative likelihood function takes on values between 0 and 1 and can be used to rank parameter values according to their plausibilities in light of the data. If $R(\theta_1) = 0.1$, say, then θ_1 is rather an implausible parameter value because the data are ten times more probable when $\theta = \hat{\theta}$ than they are when $\theta = \theta_1$. However, if $R(\theta_1) = 0.5$, say, then θ_1 is a fairly plausible value because it gives the data 50% of the maximum possible probability under the model.

6.2.21 Definition

The set of θ values for which $R(\theta) \geq p$ is called a $100p\%$ *likelihood region* for θ. If the region is an interval of real values then it is called a $100p\%$ *likelihood interval (L.I.)* for θ.

Values inside the 10% L.I. are referred to as plausible and values outside this interval as implausible. Values inside a 50% L.I. are very plausible and outside a 1% L.I. are very implausible in light of the data.

6.2.22 Definition

The *log relative likelihood function* is the natural logarithm of the relative likelihood function:

$$r(\theta) = r\,(\theta; x) = \log[R(\theta)] = \log[L(\theta] - \log[L(\hat{\theta})] = l(\theta) - l(\hat{\theta}), \qquad \theta \in \Omega.$$

Likelihood regions or intervals may be determined from a graph of $R(\theta)$ or $r(\theta)$ and usually it is more convenient to work with $r(\theta)$. Alternatively, they can be found by solving $r(\theta) - \log p = 0$. Usually this must be done numerically.

6.2.23 Example

Plot the relative likelihood function for θ in Example 6.2.4 if $n = 30$ and $\hat{\theta} = 5$. Find 10% and 50% L.I.'s for θ. See Figure 6.2.

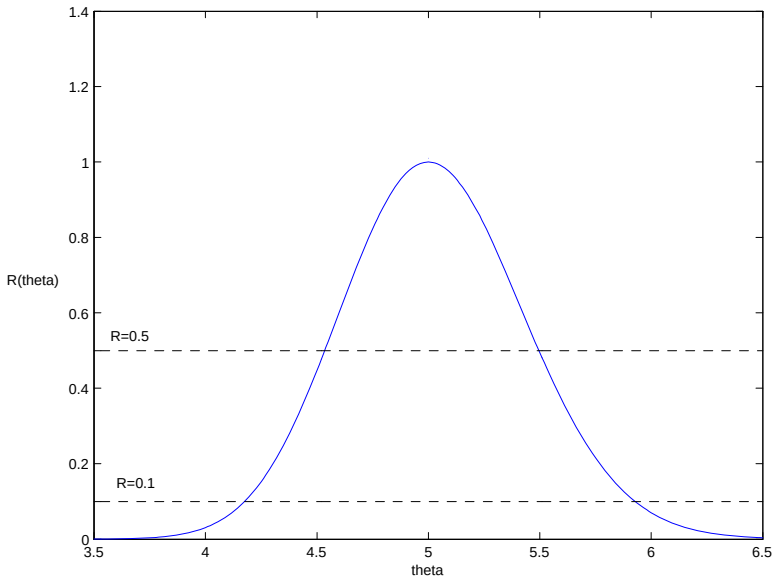


Figure 6.2: Poisson Relative Likelihood Function

6.2.24 Exercise

Suppose $(X_1, \ldots, X_n)$ is a random sample from the EXP$(1, \theta)$ distribution. Plot the relative likelihood function for θ if $n = 20$ and $x_{(1)} = 1$. Find 10% and 50% L.I.'s for θ.

6.3 Maximum Likelihood Method - Multiparameter

The case of several parameters is exactly analogous to the one parameter case. Suppose $\theta = (\theta_1, \ldots, \theta_k)^T$. In this case the "parameter" can be thought of as a column vector of k scalar parameters. The likelihood function $l(\theta_1, \ldots, \theta_k) = \log L(\theta_1, \ldots, \theta_k)$ is a function of k parameters. The M.L. estimate of θ, $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_k)^T$ is usually found by solving $\frac{\partial l}{\partial \theta_j} = 0$, $j = 1, \ldots, k$ simultaneously.

The invariance property of the M.L. estimator also holds in the multiparameter case.

6.3.1 Definition

If $\theta = (\theta_1, \ldots, \theta_k)^T$ then the *score vector* is defined as

$$S(\theta) = S(\theta; x) = \left[\frac{\partial l}{\partial \theta_1}, \ldots, \frac{\partial l}{\partial \theta_k}\right]^T \qquad \theta \in \Omega.$$

6.3.2 Definition

If $\theta = (\theta_1, \ldots, \theta_k)^T$ then the *information matrix* $I(\theta) = I(\theta; x)$ is a $k \times k$ symmetric matrix whose (i, j) entry is given by

$$-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta).$$

$I(\hat{\theta})$ is called the *observed information matrix*.

6.3.3 Definition

If $\theta = (\theta_1, \ldots, \theta_k)^T$ then the *expected* or *Fisher information matrix* $J(\theta)$ is a $k \times k$ symmetric matrix whose (i, j) entry is given by

$$E\left[-\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\theta; X)\right].$$

6.3.4 Likelihood Regions

The set of θ values for which $R(\theta) \geq p$ is called a $100p\%$ *likelihood region* for θ.

6.3.5 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Find the score vector, the information matrix, the Fisher information matrix and the M.L. estimator of $\theta = \left(\mu, \sigma^2\right)^T$. Find the observed information matrix $I\left(\hat{\mu}, \hat{\sigma}^2\right)$ and thus verify that $\left(\hat{\mu}, \hat{\sigma}^2\right)$ is the M.L. estimator of $\left(\mu, \sigma^2\right)$. What is the M.L. estimator of the parameter $\tau = \tau\left(\mu, \sigma^2\right) = \mu/\sigma$ which is called the coefficient of variation?

Since $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}\left(\mu, \sigma^2\right)$ distribution the likelihood function is

$$\begin{aligned} L\left(\mu, \sigma^2\right) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[\frac{-1}{2\sigma^2}\left(x_i - \mu\right)^2\right] \\ &= \left(2\pi\right)^{-n/2}\left(\sigma^2\right)^{-n/2} \exp\left[\frac{-1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2\right]. \end{aligned}$$

The log likelihood function is

$$\begin{aligned} l\left(\mu, \sigma^2\right) &= -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i - \mu\right)^2 \\ &= -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2}\left(\sigma^2\right)^{-1}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + n\left(\bar{x} - \mu\right)^2\right] \\ &= -\frac{n}{2}\log\left(2\pi\right) - \frac{n}{2}\log\left(\sigma^2\right) - \frac{1}{2}\left(\sigma^2\right)^{-1}\left[\left(n-1\right)s^2 + n\left(\bar{x} - \mu\right)^2\right] \end{aligned}$$

where

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2.$$

Now

$$\frac{\partial l}{\partial \mu} = \frac{n}{\sigma^2}\left(\bar{x} - \mu\right) = n\left(\sigma^2\right)^{-1}\left(\bar{x} - \mu\right)$$

and

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2}\left(\sigma^2\right)^{-1} + \frac{1}{2}\left(\sigma^2\right)^{-2}\left[\left(n-1\right)s^2 + n\left(\bar{x} - \mu\right)^2\right].$$

The equations

$$\frac{\partial l}{\partial \mu} = 0, \qquad \frac{\partial l}{\partial \sigma^2} = 0$$

are solved simultaneously for

$$\mu = \bar{x} \quad \text{and} \quad \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{(n-1)}{n}s^2.$$

Since

$$\begin{aligned} -\frac{\partial^2 l}{\partial \mu^2} &= \frac{n}{\sigma^2}, \quad -\frac{\partial^2 l}{\partial \sigma^2 \partial \mu} = \frac{n\left(\bar{x}-\mu\right)}{\sigma^4} \\ -\frac{\partial^2 l}{\partial \left(\sigma^2\right)^2} &= -\frac{n}{2}\frac{1}{\sigma^4} + \frac{1}{\sigma^6}\left[(n-1)s^2 + n\left(\bar{x}-\mu\right)^2\right] \end{aligned}$$

the information matrix is

$$I\left(\mu,\sigma^2\right) = \begin{bmatrix} n/\sigma^2 & n\left(\bar{x}-\mu\right)/\sigma^4 \\ n\left(\bar{x}-\mu\right)/\sigma^4 & -\frac{n}{2}\frac{1}{\sigma^4} + \frac{1}{\sigma^6}\left[(n-1)s^2 + n\left(\bar{x}-\mu\right)^2\right] \end{bmatrix}.$$

Since

$$I_{11}\left(\hat{\mu},\hat{\sigma}^2\right) = \frac{n}{\hat{\sigma}^2} > 0 \quad \text{and} \quad \det I\left(\hat{\mu},\hat{\sigma}^2\right) = \frac{n^2}{2\hat{\sigma}^6} > 0$$

then by the Second Derivative Test the M.L. estimates of of μ and σ^2 are

$$\hat{\mu} = \bar{x} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2 = \frac{(n-1)}{n}s^2$$

and the M.L. estimators are

$$\hat{\mu} = \bar{X} \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2 = \frac{(n-1)}{n}S^2.$$

The observed information is

$$I\left(\hat{\mu},\hat{\sigma}^2\right) = \begin{bmatrix} n/\hat{\sigma}^2 & 0 \\ 0 & \frac{1}{2}\left(n/\hat{\sigma}^4\right) \end{bmatrix}.$$

Now

$$E\left(\frac{n}{\sigma^2}\right) = \frac{n}{\sigma^2}, \qquad E\left[\frac{n\left(\bar{X}-\mu\right)}{\sigma^4}\right] = 0,$$

and

$$\begin{aligned} & E\left\{-\frac{n}{2}\frac{1}{\sigma^4} + \frac{1}{\sigma^6}\left[(n-1)S^2 + n\left(\bar{X}-\mu\right)^2\right]\right\} \\ = & -\frac{n}{2}\frac{1}{\sigma^4} + \frac{1}{\sigma^6}\left\{(n-1)E(S^2) + nE\left[\left(\bar{X}-\mu\right)^2\right]\right\} \\ = & -\frac{n}{2}\frac{1}{\sigma^4} + \frac{1}{\sigma^6}\left\{(n-1)\sigma^2 + \sigma^2\right\} \\ = & \frac{n}{2\sigma^4} \end{aligned}$$

since

$$E\left(\bar{X}-\mu\right)=0,\ E\left[\left(\bar{X}-\mu\right)^2\right]=Var\left(\bar{X}\right)=\frac{\sigma^2}{n}\ \text{ and }\ E\left(S^2\right)=\sigma^2.$$

Therefore the Fisher information matrix is

$$J\left(\mu,\sigma^2\right)=\begin{bmatrix} n/\sigma^2 & 0 \\ 0 & \frac{1}{2}\left(n/\sigma^4\right) \end{bmatrix}$$

and the inverse of the Fisher information matrix is

$$\left[J\left(\mu,\sigma^2\right)\right]^{-1}=\begin{bmatrix} \sigma^2/n & 0 \\ 0 & \left(2\sigma^4\right)/n \end{bmatrix}$$

Note that

$$\begin{aligned} Var\left(\bar{X}\right) &= \frac{\sigma^2}{n} \\ Var\left(\hat{\sigma}^2\right) &= Var\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2\right]=\frac{2(n-1)\sigma^4}{n^2}\approx\frac{2\sigma^4}{n} \end{aligned}$$

and

$$Cov(\bar{X},\hat{\sigma}^2)=\frac{1}{n}Cov(\bar{X},\ \sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2)=0$$

since $\bar{X}$ and $\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ are independent random variables.

By the invariance property of M.L. estimators the M.L. estimator of $\tau=\mu/\sigma$ is $\hat{\tau}=\hat{\mu}/\hat{\sigma}$.

Recall from STAT 231 that inferences for μ and σ^2 are made using the pivotal quantities

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}\backsim\ t\left(n-1\right)\quad\text{and}\quad\frac{(n-1)S^2}{\sigma^2}\backsim\chi^2\left(n-1\right).$$

See Figure 6.3 for a graph of $R\left(\mu,\sigma^2\right)$ for $n=350$, $\hat{\mu}=160$ and $\hat{\sigma}^2=36$.

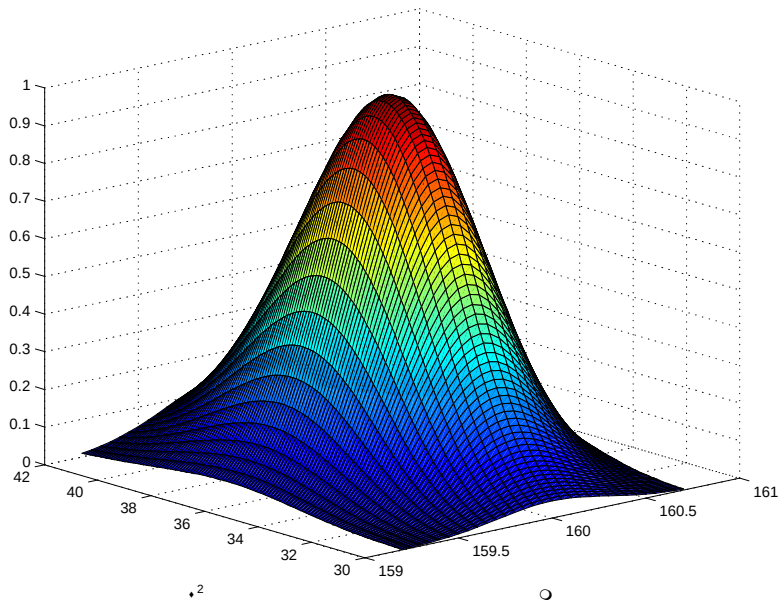


Figure 6.3: Normal Relative Likelihood Function for $n = 350$, $\hat{\mu} = 160$ and $\hat{\sigma}^2 = 36$

Often $S(\theta) = (0, \ldots, 0)^T$ must be solved numerically using a method such as Newton's Method.

6.3.6 Newton's Method

Let $\theta^{(0)}$ be an initial estimate of $\theta = (\theta_1, \ldots, \theta_k)^T$. The estimate $\theta^{(i)}$ can be updated using

$$\theta^{(i+1)} = \theta^{(i)} + \left[I(\theta^{(i)})\right]^{-1} S(\theta^{(i)}), \qquad i = 0, 1, \ldots$$

Note: The initial estimate, $\theta^{(0)}$, may be determined by calculating $L(\theta)$ for a grid of values to determine the region in which $L(\theta)$ obtains a maximum.

6.3.7 Example

The following data are 30 independent observations from a BETA(a, b) distribution:

0.2326, 0.0465, 0.2159, 0.2447, 0.0674, 0.3729, 0.3247, 0.3910, 0.3150, 0.3049, 0.4195, 0.3473, 0.2709, 0.4302, 0.3232, 0.2354, 0.4014, 0.3720, 0.5297, 0.1508, 0.4253, 0.0710, 0.3212, 0.3373, 0.1322, 0.4712, 0.4111, 0.1079, 0.0819, 0.3556

The likelihood function for observations $x_1, x_2, ..., x_n$ is

$$\begin{aligned} L(a,b) &= \prod_{i=1}^{n} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x_i^{a-1} (1-x_i)^{b-1}, \quad a>0,\ b>0 \\ &= \left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\right]^n \left[\prod_{i=1}^{n} x_i\right]^{a-1} \left[\prod_{i=1}^{n}(1-x_i)\right]^{b-1}. \end{aligned}$$

The log likelihood function is

$$l(a,b) = n\left[\log\Gamma(a+b) - \log\Gamma(a) - \log\Gamma(b) + (a-1)t_1 + (b-1)t_2\right]$$

where

$$t_1 = \frac{1}{n}\sum_{i=1}^{n} \log x_i \quad \text{and} \quad t_2 = \frac{1}{n}\sum_{i=1}^{n} \log(1-x_i).$$

Let

$$\Psi(z) = \frac{d\log\Gamma(z)}{dz} = \frac{\Gamma'(z)}{\Gamma(z)}$$

which is called the digamma function. The score vector is

$$S(a,b) = \begin{bmatrix} \partial l/\partial a \\ \partial l/\partial b \end{bmatrix} = n \begin{bmatrix} \Psi(a+b) - \Psi(a) + t_1 \\ \Psi(a+b) - \Psi(b) + t_2 \end{bmatrix}.$$

$S(a,b) = [0\ 0]^T$ must be solved numerically to find the M.L. estimates of a and b.

Let

$$\Psi'(z) = \frac{d}{dz}\Psi(z)$$

which is called the trigamma function. The information matrix is

$$I(a,b) = n \begin{bmatrix} \Psi'(a) - \Psi'(a+b) & -\Psi'(a+b) \\ -\Psi'(a+b) & \Psi'(b) - \Psi'(a+b) \end{bmatrix}$$

which is also the Fisher or expected information matrix.

For the data above

$$t_1 = \frac{1}{30}\sum_{i=1}^{n} \log x_i = -1.3929 \quad \text{and} \quad t_2 = \frac{1}{30}\log \sum_{i=1}^{n} \log(1 - x_i) = -0.3594.$$

The M.L. estimates of a and b can be found using Newton's Method given by

$$\begin{bmatrix} a^{(i+1)} \\ b^{(i+1)} \end{bmatrix} = \begin{bmatrix} a^{(i)} \\ b^{(i)} \end{bmatrix} + \left[I(a^{(i)}, b^{(i)})\right]^{-1} S(a^{(i)}, b^{(i)})$$

for $i = 0, 1, ...$ until convergence. Newton's Method converges after 8 interations beginning with the initial estimates $a^{(0)} = 2$, $b^{(0)} = 2$. The iterations are given below:

$$\begin{bmatrix} 0.6449 \\ 2.2475 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix} + \begin{bmatrix} 10.8333 & -8.5147 \\ -8.5147 & 10.8333 \end{bmatrix}^{-1} \begin{bmatrix} -16.7871 \\ 14.2190 \end{bmatrix}$$

$$\begin{bmatrix} 1.0852 \\ 3.1413 \end{bmatrix} = \begin{bmatrix} 0.6449 \\ 2.2475 \end{bmatrix} + \begin{bmatrix} 84.5929 & -12.3668 \\ -12.3668 & 4.3759 \end{bmatrix}^{-1} \begin{bmatrix} 26.1919 \\ -1.5338 \end{bmatrix}$$

$$\begin{bmatrix} 1.6973 \\ 4.4923 \end{bmatrix} = \begin{bmatrix} 1.0852 \\ 3.1413 \end{bmatrix} + \begin{bmatrix} 35.8351 & -8.0032 \\ -8.0032 & 3.2253 \end{bmatrix}^{-1} \begin{bmatrix} 11.1198 \\ -0.5408 \end{bmatrix}$$

$$\begin{bmatrix} 2.3133 \\ 5.8674 \end{bmatrix} = \begin{bmatrix} 1.6973 \\ 4.4923 \end{bmatrix} + \begin{bmatrix} 18.5872 & -5.2594 \\ -5.2594 & 2.2166 \end{bmatrix}^{-1} \begin{bmatrix} 4.2191 \\ -0.1922 \end{bmatrix}$$

$$\begin{bmatrix} 2.6471 \\ 6.6146 \end{bmatrix} = \begin{bmatrix} 2.3133 \\ 5.8674 \end{bmatrix} + \begin{bmatrix} 12.2612 & -3.9004 \\ -3.9004 & 1.6730 \end{bmatrix}^{-1} \begin{bmatrix} 1.1779 \\ -0.0518 \end{bmatrix}$$

$$\begin{bmatrix} 2.7058 \\ 6.7461 \end{bmatrix} = \begin{bmatrix} 2.6471 \\ 6.6146 \end{bmatrix} + \begin{bmatrix} 10.3161 & -3.4203 \\ -3.4203 & 1.4752 \end{bmatrix}^{-1} \begin{bmatrix} 0.1555 \\ -0.0067 \end{bmatrix}$$

$$\begin{bmatrix} 2.7072 \\ 6.7493 \end{bmatrix} = \begin{bmatrix} 2.7058 \\ 6.7461 \end{bmatrix} + \begin{bmatrix} 10.0345 & -3.3478 \\ -3.3478 & 1.4450 \end{bmatrix}^{-1} \begin{bmatrix} 0.0035 \\ -0.0001 \end{bmatrix}$$

$$\begin{bmatrix} 2.7072 \\ 6.7493 \end{bmatrix} = \begin{bmatrix} 2.7072 \\ 6.7493 \end{bmatrix} + \begin{bmatrix} 10.0280 & -3.3461 \\ -3.3461 & 1.4443 \end{bmatrix}^{-1} \begin{bmatrix} 0.0000 \\ 0.0000 \end{bmatrix}$$

The M.L. estimates are $\hat{a} = 2.7072$ and $\hat{b} = 6.7493$.

The observed information matrix is

$$I(\hat{a},\hat{b}) = \begin{bmatrix} 10.0280 & -3.3461 \\ -3.3461 & 1.4443 \end{bmatrix}$$

Note that since $\det[I(\hat{a},\hat{b})] = (10.0280)(1.4443) - (3.3461)^2 > 0$ and $[I(\hat{a},\hat{b})]_{11} = 10.0280 > 0$ and then by the Second Derivative Test we have found the M.L. estimates.

A graph of the relative likelihood function is given in Figure 6.4.

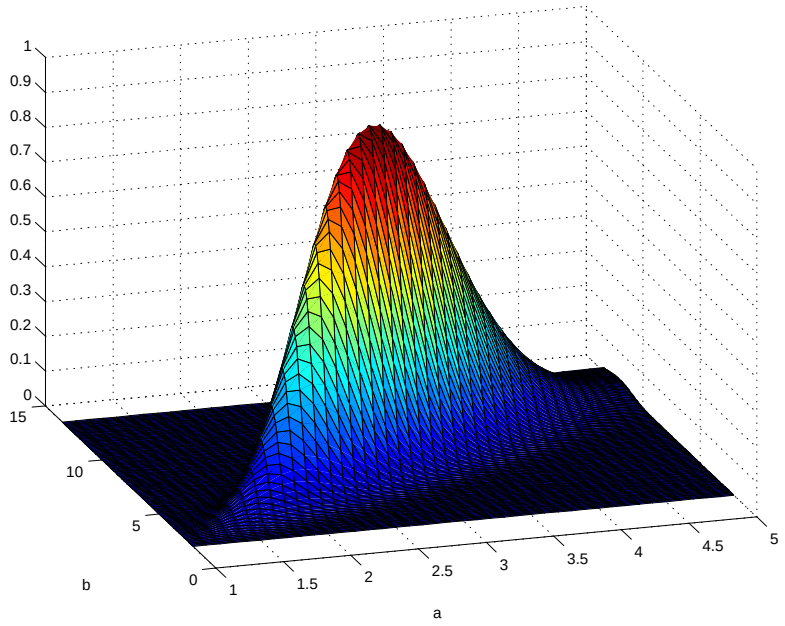


Figure 6.4: Beta Relative Likelihood Function

A $100p\%$ likelihood region for (a, b) is given by $\{(a, b)\,;\ R\,(a, b) \geq p\}$. The 1%, 5% and 10% likelihood regions for (a, b) are shown in Figure 6.5.

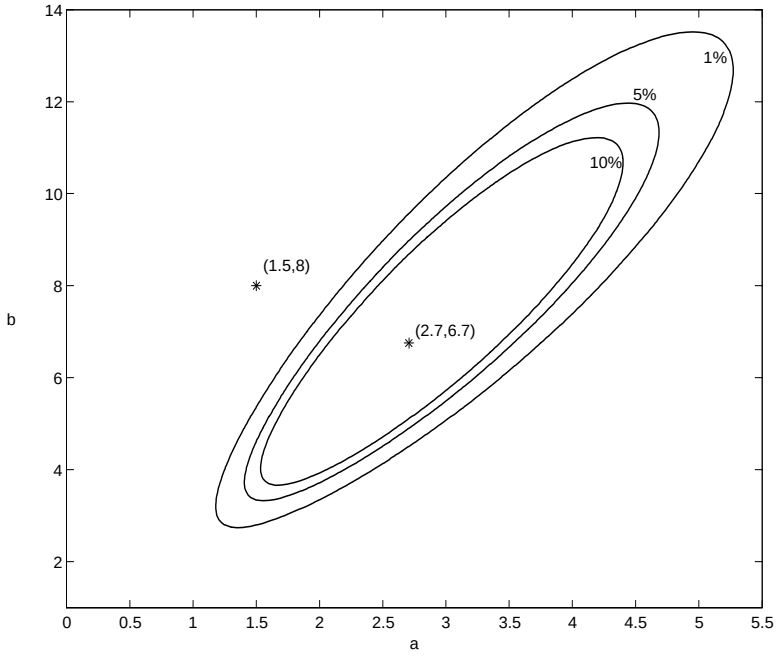


Figure 6.5: Likelihood Regions for BETA(a,b) Example

Note that the likelihood contours are elliptical in shape and are skewed relative to the ab coordinate axes. This follows since, for (a, b) sufficiently close to $(\hat{a}, \hat{b})$,

$$\begin{aligned} L\,(a, b) &\approx L(\hat{a}, \hat{b}) + S(\hat{a}, \hat{b}) \begin{bmatrix} \hat{a} - a \\ \hat{b} - b \end{bmatrix} + \frac{1}{2} \begin{bmatrix} \hat{a} - a & \hat{b} - b \end{bmatrix} I(\hat{a}, \hat{b}) \begin{bmatrix} \hat{a} - a \\ \hat{b} - b \end{bmatrix} \\ &= L(\hat{a}, \hat{b}) + \frac{1}{2} \begin{bmatrix} \hat{a} - a & \hat{b} - b \end{bmatrix} I(\hat{a}, \hat{b}) \begin{bmatrix} \hat{a} - a \\ \hat{b} - b \end{bmatrix} \quad \text{since } S(\hat{a}, \hat{b}) = 0. \end{aligned}$$

Therefore

$$\begin{aligned} R(a,b) &= \frac{L(a,b)}{L(\hat{a},\hat{b})} \\ &\approx 1 - \left[2L(\hat{a},\hat{b})\right]^{-1} \begin{bmatrix} \hat{a}-a & \hat{b}-b \end{bmatrix} I(\hat{a},\hat{b}) \begin{bmatrix} \hat{a}-a \\ \hat{b}-b \end{bmatrix} \\ &= 1 - \left[2L(\hat{a},\hat{b})\right]^{-1} \begin{bmatrix} \hat{a}-a & \hat{b}-b \end{bmatrix} \begin{bmatrix} \hat{I}_{11} & \hat{I}_{12} \\ \hat{I}_{12} & \hat{I}_{22} \end{bmatrix} \begin{bmatrix} \hat{a}-a \\ \hat{b}-b \end{bmatrix} \\ &= 1 - \left[2L(\hat{a},\hat{b})\right]^{-1} \left[(a-\hat{a})^2 \hat{I}_{11} + 2(a-\hat{a})(b-\hat{b})\hat{I}_{12} + (b-\hat{b})^2 \hat{I}_{22}\right]. \end{aligned}$$

The set of points (a,b) which satisfy $R(a,b) = p$ is approximately the set of points (a,b) which satisfy

$$(a-\hat{a})^2 \hat{I}_{11} + 2(a-\hat{a})(b-\hat{b})\hat{I}_{12} + (b-\hat{b})^2 \hat{I}_{22} = 2(1-p)L(\hat{a},\hat{b})$$

which we recognize as the points on an ellipse centred at $(\hat{a},\hat{b})$. The skewness of the likelihood contours relative to the ab coordinate axes is determined by the value of $\hat{I}_{12}$. If this value is close to zero the skewness will be small.

6.3.8 Exercise

Suppose $Y_i \sim \mathrm{N}(\alpha + \beta x_i, \sigma^2)$, $i = 1, \ldots, n$ independently where the x_i are known constants. Show that the M.L. estimators of α, β and σ^2 are given by

$$\begin{aligned} \hat{\alpha} &= \bar{Y} - \hat{\beta}\bar{x}, \\ \hat{\beta} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\alpha} - \hat{\beta}x_i)^2 \end{aligned}$$

Note: $\hat{\alpha}$ and $\hat{\beta}$ are also the least squares estimators of α and β.

6.4 Asymptotic Properties of M.L. Estimators - One Parameter

6.4.1 Theorem - Asymptotic Distribution of the M.L. Estimator

Suppose $X = (X_1, \ldots, X_n)$ be a random sample from $f(x;\theta)$. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be the M.L. estimator of θ based on X. Then under certain (regularity) conditions

$$\hat{\theta}_n \to_p \theta_0 \tag{6.1}$$

$$[J(\theta_0)]^{1/2}(\hat{\theta}_n - \theta_0) \to_D Z \backsim \mathrm{N}(0,1) \tag{6.2}$$

$$-2\log R(\theta_0; X) = 2[l(\hat{\theta}_n; X) - l(\theta_0; X)] \to_D W \backsim \chi^2(1) \tag{6.3}$$

where θ_0 is the true but unknown value of θ.

Since (6.1) holds $\hat{\theta}_n$ is called a *consistent estimator* of θ.

This theorem implies that for sufficiently large n, $\hat{\theta}_n$ has an approximately $\mathrm{N}(\theta_0, [J(\theta_0)]^{-1})$ distribution. $[J(\theta_0)]^{-1}$ is called the *asymptotic variance* of $\hat{\theta}_n$ and for sufficiently large n

$$Var(\hat{\theta}_n) \approx [J(\theta_0)]^{-1}.$$

Of course $J(\theta_0)$ is unknown because θ_0 is unknown. But (6.1), (6.2) and the Limit Theorems imply that

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \to_D Z \backsim \mathrm{N}(0,1) \tag{6.4}$$

and therefore for sufficiently large n

$$Var(\hat{\theta}_n) \approx \left[J(\hat{\theta}_n)\right]^{-1}.$$

We will see in the next section how these results can be used to construct approximate confidence intervals for θ.

By the WLLN

$$\frac{1}{n}I(\theta;X) = -\frac{1}{n}\sum_{i=1}^{n}\frac{d^2}{d\theta^2}l(\theta;X_i) \to_p E\left[-\frac{d^2}{d\theta^2}l(\theta;X_i)\right]. \tag{6.5}$$

Therefore by (6.1), (6.2), (6.5) and the Limit Theorems it follows that

$$[I(\hat{\theta}_n;X)]^{1/2}(\hat{\theta}_n - \theta_0) \to_D Z \backsim \text{N}(0,1) \tag{6.6}$$

so that for sufficiently large n

$$Var(\hat{\theta}_n) \approx \left[I(\hat{\theta}_n)\right]^{-1}$$

where $I(\hat{\theta}_n)$ is the observed information.

In Chapter 7 we will see how result (6.3) can be used in a test of hypothesis.

Note: These results do not hold if the support set of X depends on θ.

6.4.2 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the WEI$(\theta, 2)$ distribution. Verify that (6.1), (6.2), (6.4) and (6.6) hold for this distribution. Note: if $X \backsim$ WEI$(\theta, 2)$ then $E\left(X^k\right) = \theta^k\Gamma\left(\frac{k}{2}+1\right)$.

6.4.3 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the UNIF$(0, \theta)$ distribution. Since the support of X_i depends on θ Theorem 6.4.1 does not hold. Show however that $\hat{\theta}_n = X_{(n)}$ is still a consistent estimator of θ. See Example 5.1.5.

6.4.4 Exercise

In Example 6.4.3 show that $n(1 - \hat{\theta}_n/\theta_0) \to_D W \backsim$ EXP(1).

6.5 Interval Estimators

6.5.1 Definition

Suppose X is a random variable whose distribution depends on θ. Suppose that $A(x)$ and $B(x)$ are functions such that $A(x) \leq B(x)$ for all $x \in$ support of X and $\theta \in \Omega$. Let x be the observed data. Then $(A(x),\ B(x))$ is an *interval estimate* for θ. The interval $(A(X),\ B(X))$ is an *interval estimator* for θ.

Likelihood intervals are one type of interval estimator. Confidence intervals are another type of interval estimator.

We now consider a general approach for constructing confidence intervals based on pivotal quantities.

6.5.2 Definition

Suppose X is a random variable whose distribution depends on θ. The random variable $Q(X;\theta)$ is called a *pivotal quantity* if the distribution of Q does not depend on θ. $Q(X;\theta)$ is called an *asymptotic pivotal quantity* if the limiting distribution of Q as $n \to \infty$ does not depend on θ.

6.5.3 Exercise - Pivotal Quantities for the Normal Distribution (STAT 231)

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Show that

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}},\ \frac{\bar{X}-\mu}{S/\sqrt{n}},\ \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\sigma^2} \text{ and } \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{\sigma^2}$$

are all pivotal quantities.

6.5.4 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{POI}(\theta)$ distribution. Show that

$$\sqrt{n}\left(\bar{X}_n-\theta\right)/\bar{X}_n$$

is an asymptotic pivotal. See Example 5.3.11.

6.5.5 Definition

Suppose $A(X)$ and $B(X)$ are statistics. If $P[A(X) < \theta < B(X)] = p$, $0 < p < 1$ then $(a(x), b(x))$ is called a $100p\%$ *confidence interval* (C.I.) for θ.

Pivotal quantities can be used for constructing C.I.'s in the following way. Since the distribution of $Q(X;\theta)$ is known we can write down a probability statement of the form

$$P(q_1 \leq Q(X;\theta) \leq q_2) = p.$$

If Q is a monotone function of θ then this statement can be rewritten as

$$P\left[A(X) \leq \theta \leq B(X)\right] = p$$

and the interval $[a(x), b(x)]$ is a $100p\%$ C.I..

6.5.6 Exercise

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. Use the pivotal quantities in Example 6.5.5 to find:

(1) a $100p\%$ C.I. for μ if σ^2 is known

(2) a $100p\%$ C.I. for μ if σ^2 is unknown

(3) a $100p\%$ C.I. for σ^2 if μ is known

(4) a $100p\%$ C.I. for σ^2 if μ is unknown.

The following theorem gives the pivotal quantity in the case in which θ is either a location or scale parameter.

6.5.7 Theorem

Let $X = (X_1, \ldots, X_n)$ be a random sample from $f(x;\theta)$ and let $\hat{\theta} = \hat{\theta}(X)$ be the M.L. estimator of the scalar parameter θ based on X.

(1) If θ is a location parameter then $Q = \hat{\theta} - \theta$ is a pivotal quantity.

(2) If θ is a scale parameter then $Q = \hat{\theta}/\theta$ is a pivotal quantity.

6.5.8 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the EXP(θ) distribution. Show that θ is a scale parameter and $\hat{\theta} = \bar{X}$. Find the distribution of the pivotal quantity $Q = \hat{\theta}/\theta$ and show how it can be used to construct an exact equal tail $100p\%$ C.I. for θ. For the data $n = 15$ and $\sum_{i=1}^{15} x_i = 36$ find a 95% equal tail C.I. for θ.

6.5.9 Exercise

Suppose $X_1, \ldots, X_n$ is a random sample from the WEI($\theta, 2$) distribution. Show that θ is a scale parameter. Find the distribution of the pivotal quantity $Q = 2n(\hat{\theta}/\theta)^2$. (**Hint:** Show that $X_i^2 \backsim \text{EXP}\left(\theta^2\right)$ and then use the m.g.f. technique.) Use this pivotal quantity to construct an equal tail 95% C.I. for θ for the data $n = 12$ and $\sum_{i=1}^{12} x_i^2 = 24$. (Answer: $[1.1043, 1.9675]$)

6.5.10 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the EXP($1, \theta$) distribution. Show that θ is a location parameter and $\hat{\theta} = X_{(1)}$ is the M.L. estimator of θ. Show that

$$P(\hat{\theta} - \theta \leq q) = 1 - e^{-nq}, \quad q \geq 0$$

and thus show that

$$[\hat{\theta} + n^{-1} \log(1-p), \quad \hat{\theta}]$$

and

$$[\hat{\theta} + n^{-1} \log((1-p)/2), \quad \hat{\theta} + n^{-1} \log((1+p)/2)]$$

are both $100p\%$ C.I.'s for θ. Which C.I. seems more reasonable?

6.5.11 Exercise

Suppose $(X_1, \ldots, X_n)$ is a random sample from the UNIF($0, \theta$) distribution. Find the c.d.f. of $Q = \hat{\theta}/\theta$ and thus show that Q is a pivotal quantity. Determine a such that

$$[\hat{\theta}, \ a\hat{\theta}]$$

is a $100p\%$ C.I. for θ.

6.5.12 Asymptotic Pivotal Quantities and Approximate C.I.'s

In cases in which an exact pivotal quantity cannot be constructed we can use the limiting distribution of the M.L. estimator $\hat{\theta}_n = \hat{\theta}_n(x_1, \ldots, x_n)$ (see Section 6.4) to construct approximate C.I.'s. Since

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \mathrm{N}(0,1)$$

then $[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0)$ is an asymptotic pivotal quantity. An approximate $100p\%$ C.I. based on this asymptotic pivotal quantity is given by

$$[\hat{\theta}_n - a[J(\hat{\theta}_n)]^{-1/2},\ \hat{\theta}_n + a[J(\hat{\theta}_n)]^{-1/2}] \tag{6.7}$$

where $P(-a < Z < a) = p$ and $Z \backsim \mathrm{N}(0,1)$.

Simlarly since

$$[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \mathrm{N}(0,1)$$

then $[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0)$ is an asymptotic pivotal quantity. An approximate $100p\%$ C.I. based on this asymptotic pivotal quantity is given by

$$[\hat{\theta}_n - a[I(\hat{\theta}_n)]^{-1/2},\ \hat{\theta}_n + a[I(\hat{\theta}_n)]^{-1/2}] \tag{6.8}$$

where $I(\hat{\theta}_n)$ is the observed information.

6.5.13 Example

Suppose $X \sim \mathrm{BIN}(n, \theta)$. Show how you would construct an approximate $100p\%$ C.I. for θ.

6.5.14 Example

For the data in Example 6.5.8 find approximate 95% C.I.'s based on (6.7) and (6.8). Compare these intervals with the equal tail 95% C.I. in Example 6.5.8.

6.5.15 Exercise

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{POI}(\theta)$ distribution. Show how you would construct an approximate $100p\%$ C.I. for θ.

6.5.16 Exercise

For the data in Exercise 6.5.9 find approximate 95% C.I.'s based on (6.7) and (6.8). Compare these intervals with the equal tail 95% C.I. in Exercise 6.5.9.

6.5.17 Likelihood Intervals and Approximate C.I.'s

A 15% L.I. for θ is given by

$$\{\theta : R(\theta; x) \geq 0.15\}.$$

Since

$$-2\log R(\theta_0; X) \rightarrow_D W \backsim \chi^2(1),$$

$$\begin{aligned} P\left[R(\theta; X) \geq 0.15\right] &= P\left[-2\log R(\theta; X) \leq -2\log(0.15)\right] \\ &= P\left[-2\log R(\theta; X) \leq 3.79\right] \\ &\approx P\left(W \leq 3.79\right) = P\left(Z^2 \leq 3.79\right) \quad \text{where } Z \backsim \mathrm{N}(0,1) \\ &\approx P\left(-1.95 \leq Z \leq 1.95\right) \\ &\approx 0.95 \end{aligned}$$

and therefore a 15% L.I. is an approximate 95% C.I. for θ. Note that while the confidence intervals given by (6.7) or (6.8) are symmetric about the point estimate $\hat{\theta}_n$, this is not true in general for likelihood intervals.

6.5.18 Example

For the data in Example 6.5.8 find a 15% L.I. Compare this interval with the intervals in Example 6.5.13.

6.5.19 Exercise

For the data in Example 6.5.9 find a 15% L.I. Compare this interval with the intervals in Exercise 6.5.16.

6.6 Asymptotic Properties of M.L. Estimators - Multiparameter

To discuss the asymptotic properties of the M.L. estimator in the multiparameter case we first review the definition and properties of the multivariate normal distribution.

6.6.1 Definition - Multivariate Normal Distribution

Let $X = (X_1, \ldots, X_k)^T$ be a $k \times 1$ random vector with $E(X_i) = \mu_i$ and $Cov(X_i, X_j) = \sigma_{ij}$, $i, j = 1, \ldots, k$. (Note: $Cov(X_i, X_i) = \sigma_{ii} = Var(X_i) = \sigma_i^2$.) Let $\mu = (\mu_1, \ldots, \mu_k)^T$ be the mean vector and Σ be the $k \times k$ symmetric covariance matrix whose (i, j) entry is σ_{ij}. Suppose also that Σ^{-1} exists. If the joint p.d.f. of $(X_1, \ldots, X_k)$ is given by

$$f(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right], \quad x \in \Re^k$$

where $x = (x_1, \ldots, x_k)^T$ then X is said to have a *multivariate normal distribution*. We write $Y \backsim \text{MVN}(\mu, \Sigma)$.

6.6.2 Theorem

Suppose $X = (X_1, \ldots, X_k)^T \backsim \text{MVN}(\mu, \Sigma)$. Then
(1) X has joint m.g.f.

$$M(t_1, \ldots, t_k) = \exp(\mu^T t + \frac{1}{2}t^T\Sigma t), \quad t = (t_1, \ldots, t_k)^T \in \Re^k.$$

(2) Any subset of $X_1, \ldots, X_k$ also has a MVN distribution and in particular $X_i \backsim \text{N}\left(\mu_i, \sigma_i^2\right)$, $i = 1, \ldots, k$.
(3) $(X-\mu)^T\Sigma^{-1}(X-\mu) \backsim \chi^2(k)$.
(4) Let $c = (c_1, \ldots, c_k)^T$ be a nonzero vector of constants then

$$c^T X = \sum_{i=1}^{k} c_i X_i \backsim \text{N}\left(c^T\mu,\ c^T\Sigma c\right).$$

(5) Let A be a $k \times p$ vector of constants of rank p then

$$A^T X \backsim \text{N}(A^T\mu,\ A^T\Sigma A).$$

(6) The conditional distribution of any subset of $(X_1, \ldots, X_k)$ given the rest of the coordinates is a multivariate normal distribution. In particular the conditional p.d.f. of X_i given $X_j = x_j$, $i \neq j$, is

$$X_i | X_j = x_j \sim \text{N}(\mu_i + \rho_{ij}\sigma_i(x_j - \mu_j)/\sigma_j,\ (1-\rho_{ij}^2)\sigma_i^2).$$

6.6.3 Theorem - Asymptotic Distribution of the M.L. Estimator in the Multiparmeter Case

Suppose $X = (X_1, \ldots, X_n)$ is a random sample from $f(x;\theta)$ where $\theta = (\theta_1, \ldots, \theta_k)^T$ is a vector. Let $\hat{\theta}_n = \hat{\theta}_n(X_1, \ldots, X_n)$ be the M.L. estimator of θ based on X. Let 0_k be a $k \times 1$ vector of zeros and let I_k be the $k \times k$ identity matrix. Then under certain (regularity) conditions

$$\hat{\theta}_n \rightarrow_p \theta_0 \tag{6.9}$$

$$[J(\theta_0)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0_k, I_k) \tag{6.10}$$

$$-2\log R(\theta_0; X) = 2[l(\hat{\theta}_n; X) - l(\theta_0; X)] \rightarrow_D W \backsim \chi^2(k) \tag{6.11}$$

where θ_0 is the true but unknown value of θ.

This theorem implies that for sufficiently large n, $\hat{\theta}_n$ has an approximately $\text{MVN}(\theta_0, [J(\theta_0)]^{-1})$ distribution. $[J(\theta_0)]^{-1}$ is called the *asymptotic variance/covariance* matrix of $\hat{\theta}_n$ and for sufficiently large n

$$Var(\hat{\theta}_n) \approx [J(\theta_0)]^{-1}.$$

Of course $J(\theta_0)$ is unknown because θ_0 is unknown. But (6.9), (6.10) and the Limit Theorems imply that

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0_k, I_k) \tag{6.12}$$

and therefore for sufficiently large n

$$Var(\hat{\theta}_n) \approx \left[J(\hat{\theta}_n)\right]^{-1}.$$

These results can be used to construct approximate confidence regions for θ.

It is also possible to show that

$$[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D \text{MVN}(0_k, I_k) \tag{6.13}$$

so that for sufficiently large n we also have

$$Var(\hat{\theta}_n) \approx \left[I(\hat{\theta}_n)\right]^{-1}$$

where $I(\hat{\theta}_n)$ is the observed information matrix.

Note: These results do not hold if the support set of X depends on θ.

6.7 Confidence Regions

6.7.1 Definition

A $100p\%$ *confidence region* for the vector $\theta = (\theta_1, \ldots, \theta_k)^T$ based on $X = (X_1, \ldots, X_n)$ is a region $R(X) \subset R^k$ which satisfies

$$P[\theta \in R(X)] = p.$$

6.7.2 Aymptotic Pivotal Quantities and Approximate Confidence Regions

The limiting distribution of $\hat{\theta}_n$ can be used to obtain approximate confidence regions for θ. Since

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \to_D Z \backsim \text{MVN}(0_k, I_k)$$

it follows that

$$(\hat{\theta}_n - \theta_0)^T J(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) \to_D W \backsim \chi^2(k)$$

and an approximate $100p\%$ confidence region for θ based on this asymptotic pivotal quantity is the set of all θ vectors in the set

$$\{\theta : (\hat{\theta}_n - \theta)^T J(\hat{\theta}_n)(\hat{\theta}_n - \theta) \leq b\}$$

where c is the value such that $P(W < c) = p$ and $W \backsim \chi^2(k)$.

Similarly since

$$[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \to_D Z \backsim \text{MVN}(0_k, I_k)$$

it follow that

$$(\hat{\theta}_n - \theta_0)^T I(\hat{\theta}_n; X)(\hat{\theta}_n - \theta_0) \to_D W \backsim \chi^2(k)$$

and an approximate $100p\%$ confidence region for θ based on this asymptotic pivotal quantity is the set of all θ vectors in the set

$$\{\theta : (\hat{\theta}_n - \theta)^T I(\hat{\theta}_n)(\hat{\theta}_n - \theta) \leq c\}$$

where $I(\hat{\theta}_n)$ is the observed information.

Finally since

$$-2 \log R(\theta_0; X) \to_D W \backsim \chi^2(k)$$

an approximate $100p\%$ confidence region for θ based on this asymptotic pivotal quantity is the set of all θ vectors satisfying

$$\{\theta : -2 \log R(\theta; x) \leq c\}$$

where $(x_1, \ldots, x_n)$ are the observed data.

6.7.3 Approximate C.I.'s for a Single Parameter

Let θ_i be the ith entry in the vector $\theta = (\theta_1, \ldots, \theta_k)^T$. Approximate confidence intervals for θ_i can also be obtained. Since

$$[J(\hat{\theta}_n)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0_k, I_k)$$

it follows that an approximate $100p\%$ C.I. for θ_i is given by

$$[\hat{\theta}_i - a\sqrt{\hat{v}_{ii}},\ \hat{\theta}_i + a\sqrt{\hat{v}_{ii}}]$$

where $\hat{\theta}_i$ is the ith entry in the vector $\hat{\theta}_n$, $\hat{v}_{ii}$ is the (i,i) entry of $[J(\hat{\theta}_n)]^{-1}$ and a is the value such that $P(-a < Z < a) = p$ where $Z \backsim \text{N}(0,1)$.

Simlarly since

$$[I(\hat{\theta}_n; X)]^{1/2}(\hat{\theta}_n - \theta_0) \rightarrow_D Z \backsim \text{MVN}(0, I_k)$$

it follows that an approximate $100p\%$ C.I. for θ_i is given by

$$[\hat{\theta}_i - a\sqrt{\hat{v}_{ii}},\ \hat{\theta}_i + a\sqrt{\hat{v}_{ii}}]$$

where $\hat{v}_{ii}$ is now the (i,i) entry of $[I(\hat{\theta}_n)]^{-1}$.

6.7.4 Example

Recall from Example 6.3.6 that for a random sample from the BETA(a,b) distribution the information matix and the Fisher information matrix are given by

$$I(a,b) = n \begin{bmatrix} \Psi'(a) - \Psi'(a+b) & -\Psi'(a+b) \\ -\Psi'(a+b) & \Psi'(b) - \Psi'(a+b) \end{bmatrix} = J(a,b).$$

Since

$$\begin{bmatrix} \hat{a} - a_0 & \hat{b} - b_0 \end{bmatrix} J(\hat{a},\hat{b}) \begin{bmatrix} \hat{a} - a_0 \\ \hat{b} - b_0 \end{bmatrix} \rightarrow_D W \backsim \chi^2(2),$$

an approximate $100p\%$ confidence region for (a,b) is given by

$$\{(a,b) : \begin{bmatrix} \hat{a} - a & \hat{b} - b \end{bmatrix} J(\hat{a},\hat{b}) \begin{bmatrix} \hat{a} - a \\ \hat{b} - b \end{bmatrix} \leq c\}$$

where $P(W \leq c) = p$. Since $\chi^2(2) = \text{GAM}(1,2) = \text{EXP}(2)$, c can be determined using

$$p = P(W \leq c) = \int_0^c \frac{1}{2} e^{-x/2} dx = 1 - e^{-c/2}$$

which gives

$$c = -2\log(1-p).$$

For $p = 0.95$, $c = -2\log(0.05) = 5.99$, an approximate 95% confidence region is given by

$$\{(a,b) : \left[\begin{array}{cc} \hat{a}-a & \hat{b}-b \end{array}\right] J(\hat{a},\hat{b}) \left[\begin{array}{c} \hat{a}-a \\ \hat{b}-b \end{array}\right] \leq 5.99\}.$$

If we let

$$J(\hat{a},\hat{b}) = \left[\begin{array}{cc} \hat{J}_{11} & \hat{J}_{12} \\ \hat{J}_{12} & \hat{J}_{22} \end{array}\right]$$

then the approximate confidence region can be written as

$$\{(a,b) : (\hat{a}-a)^2 \hat{J}_{11} + 2(\hat{a}-a)(\hat{b}-b)\hat{J}_{12} + (\hat{b}-b)^2\hat{J}_{22} \leq 5.99\}.$$

We note that the approximate confidence region is the set of points inside the ellipse

$$(\hat{a}-a)^2 \hat{J}_{11} + 2(\hat{a}-a)(\hat{b}-b)\hat{J}_{12} + (\hat{b}-b)^2\hat{J}_{22} = 5.99$$

which is centred at $(\hat{a},\hat{b})$.

For the data in Example 6.3.6, $\hat{a} = 2.7072$, $\hat{b} = 6.7493$ and

$$J(\hat{a},\hat{b}) = \left[\begin{array}{cc} 10.0280 & -3.3461 \\ -3.3461 & 1.4443 \end{array}\right].$$

Approximate 90%, 95% and 99% confidence regions are shown in Figure 6.3.

A 10% likelihood region for (a,b) is given by $\{(a,b) : R(a,b;x) \geq 0.1\}$. Since

$$-2\log R(a_0,b_0;X) \rightarrow_D W \backsim \chi^2(2) = \text{EXP}(2)$$

we have

$$\begin{array}{rcl} P[R(a,b;X) \geq 0.1] & = & P[-2\log R(a,b;X) \leq -2\log(0.1)] \\ & \approx & P(W \leq -2\log(0.1)) \\ & = & 1 - e^{-[-2\log(0.1)]/2} \\ & = & 1 - 0.1 = 0.9 \end{array}$$

and therefore a 10% likelihood region corresponds to an approximate 90% confidence region. Similarly 1% and 5% likelihood regions correspond to approximate 99% and 95% confidence regions respectively. Compare the

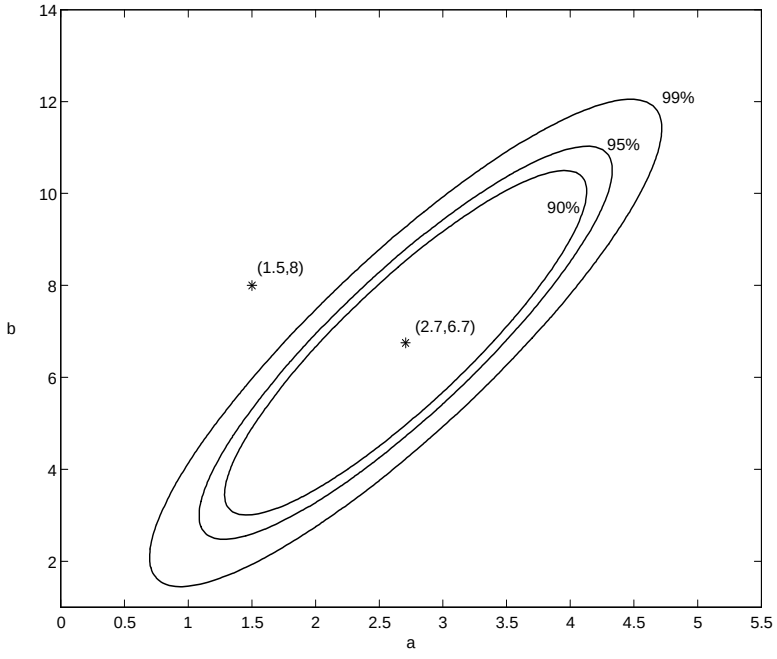


Figure 6.6: Approximate Confidence Regions for BETA(a,b) Example

likelihood regions in Figure 6.2 with the approximate confidence regions shown in Figure 6.3. What do you notice?

Let

$$\left[J(\hat{a},\hat{b})\right]^{-1} = \begin{bmatrix} \hat{v}_{11} & \hat{v}_{12} \\ \hat{v}_{12} & \hat{v}_{22} \end{bmatrix}.$$

Since

$$[J(\hat{a},\hat{b})]^{1/2} \begin{bmatrix} \hat{a}-a_0 \\ \hat{b}-b_0 \end{bmatrix} \rightarrow_D Z \backsim \text{BVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

then for large n, $Var(\hat{a}) \approx \hat{v}_{11}$, $Var(\hat{b}) \approx \hat{v}_{22}$ and $Cov(\hat{a},\hat{b}) \approx \hat{v}_{12}$. Therefore an approximate 95% C.I. for a is given by

$$[\hat{a} - 1.96\sqrt{\hat{v}_{11}},\ \hat{a} + 1.96\sqrt{\hat{v}_{11}}]$$

and an approximate 95% C.I. for b is given by

$$[\hat{b} - 1.96\sqrt{\hat{v}_{22}},\ \hat{b} + 1.96\sqrt{\hat{v}_{22}}].$$

For the data in Example 6.2.23, $\hat{a} = 2.7072$, $\hat{b} = 6.7493$ and

$$\left[J(\hat{a},\hat{b})\right]^{-1} = \begin{bmatrix} 0.4393 & 1.0178 \\ 1.0178 & 3.0503 \end{bmatrix}.$$

An approximate 95% C.I. for a is

$$[2.7072 + 1.96\sqrt{0.44393},\ 2.7072 - 1.96\sqrt{0.44393}] = [1.4080, 4.0063]$$

and an approximate 95% C.I. for b is

$$[6.7493 - 1.96\sqrt{3.0503},\ 6.7493 + 1.96\sqrt{3.0503}] = [3.3261, 10.1725].$$

Note that $a = 1.5$ is in the approximate 95% C.I. for a and $b = 8$ is in the approximate 95% C.I. for b and yet the point $(1.5, 8)$ is not in the approximate 95% joint confidence region for (a, b). Clearly these marginal C.I.'s for a and b must be used with care.

To obtain an approximate 95% C.I. for $a + b$ we note that

$$\begin{aligned} Var(\hat{a}+\hat{b}) &= Var(\hat{a}) + Var(\hat{b}) + 2Cov(\hat{a},\hat{b}) \\ &\approx \hat{v}_{11} + \hat{v}_{22} + 2\hat{v}_{12} = \hat{v} \end{aligned}$$

so that an approximate 95% C.I. for $a + b$ is given by

$$[\hat{a} + \hat{b} - 1.96\sqrt{\hat{v}},\ \hat{a} + \hat{b} + 1.96\sqrt{\hat{v}}].$$

For the data in Example 6.3.6

$$\begin{aligned} \hat{a} + \hat{b} &= 2.7072 + 6.7493 = 9.4565, \\ \hat{v} &= \hat{v}_{11} + \hat{v}_{22} + 2\hat{v}_{12} = 0.4393 + 3.0503 + 2(1.0178) = 5.5293 \end{aligned}$$

and an approximate 95% C.I. for $a + b$ is

$$[9.4565 + 1.96\sqrt{5.5293},\ 9.4565 - 1.96\sqrt{5.5293}] = [4.8493, 14.0636].$$

Chapter 7

Hypothesis Tests

7.1 Introduction

A *test of hypothesis* is a procedure for evaluating the strength of the evidence provided by the data against an hypothesis. In many cases the hypothesis can be formulated in terms of the parameters in the model $f(x;\theta)$ where $\theta = (\theta_1, \ldots, \theta_k)^T \in \Omega$ and Ω is the parameter space or set of possible values of θ. Usually we write

$$H : \theta \in \Omega_0$$

where Ω_0 is some subset of Ω.

To measure the evidence against H based on the observed data we use a test statistic or discrepancy measure. A small observed value of the test statistic shows close agreement between the observed data and the hypothesis and a large observed value of the test statistic indicates poor agreement. The test statistic is chosen before the data are examined and the choice reflects the type of departure from the hypothesis that we wish to detect. A general method for constructing test statistics can be based on the likelihood function as we will see in the next section.

After the data have been collected the observed value of the test statistic is calculated. Assuming the hypothesis H is true we compute the probability of observing a value of the test statistic at least as great as that observed. This probability is called the *significance level* (S.L.) or *p-value* of the data in relation to the hypothesis. The S.L. is the probability of observing such poor agreement between the hypothesis and the data if the hypothesis is true. If the S.L. is very small, then such poor agreement would occur very rarely if the hypothesis is true, and we have evidence against the hypothesis. The smaller the S.L. the stronger the evidence against the

hypothesis. A large S.L. does not mean that the hypothesis is true but only indicates a lack of evidence against the hypothesis based on the observed data.

7.1.1 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution where σ^2 is known and we wish to test $H : \mu = \mu_0$. The test statistic usually used for this case is

$$T = \frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}}.$$

What is the distribution of

$$\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

if H is true? Suppose $\sigma^2 = 1$. Find the S.L. if the observed data are $n = 25$, $\bar{x} = 0.5$ and $\mu_0 = 0$. What would you conclude?

7.2 Likelihood Ratio Tests for Simple Hypotheses

Suppose $X_1, \ldots, X_n$ is a random sample from $f(x; \theta)$ where $\theta = (\theta_1, \ldots, \theta_k)^T$. Suppose we wish to test $H : \theta = \theta_0$ where θ_0 is a completely known $k \times 1$ vector. This hypothesis is called a simple hypothesis since it specifies the values of all unknown parameters in the model.

The *likelihood ratio* (L.R.) statistic is defined as

$$\begin{aligned} -2 \log R(\theta_0; X) &= -2 \log \left[\frac{L(\theta_0; X)}{L(\hat{\theta}; X)} \right] \\ &= 2 \left[l(\hat{\theta}; X) - l\left(\theta_0; X\right) \right] \end{aligned}$$

where $X = (X_1, \ldots, X_n)$ are the data and $\hat{\theta} = \hat{\theta}\left(X_1, \ldots, X_n\right)$ is the M.L. estimator of θ. Under certain regularity conditions and assuming H is true,

$$-2 \log R(\theta_0; X) \rightarrow_D W \backsim \chi^2(k).$$

Therefore an approximate S.L. is given by

$$S.L. \approx P(W \geq -2 \log R(\theta_0; x))$$

where $x = (x_1, \ldots, x_n)$ are the observed data.

7.2.1 Example

In Example 7.1.1 find the L.R. statistic and compare it to the test statistic used there.

7.2.2 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution where μ is known. Find the L.R. statistic for testing $H : \sigma^2 = \sigma_0^2$.

7.2.3 Exercise

In the previous example show that the L.R. statistic takes on large values if $\hat{\sigma}^2 > \sigma_0^2$ or $\hat{\sigma}^2 < \sigma_0^2$. Hint: Compare the graphs of the functions $f(t) = t - 1$ and $g(t) = \log t$, $t > 0$ and let $t = (\hat{\sigma}/\sigma_0)^2$.

7.2.4 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{POI}(\theta)$ distribution. Find the L.R. statistic for testing $H : \theta = \theta_0$. Another test statistic which could be used is

$$T = \frac{|\bar{X} - \theta_0|}{\sqrt{\theta_0/n}}.$$

What is the approximate distribution of

$$\frac{\bar{X} - \theta_0}{\sqrt{\theta_0/n}}$$

for large n if H is true and how could you use this to find an approximate S.L.?

7.2.5 Exercise

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{EXP}(\theta)$ distribution. Show that the L.R. statistic for testing $H : \theta = \theta_0$ is given by

$$-2 \log R(\theta_0; X) = 2n \left[\frac{\bar{X}}{\theta_0} - \log \left(\frac{\bar{X}}{\theta_0} \right) - 1 \right]$$

Another test statistic which could be used is

$$T = \frac{|\bar{X} - \theta_0|}{\theta_0/\sqrt{n}}.$$

What is the approximate distribution of

$$\frac{\bar{X} - \theta_0}{\theta_0/\sqrt{n}}$$

for large n if H is true and how could you use this to find an approximate S.L.?

7.2.6 Example

The following table gives the observed frequencies of the six faces in 100 rolls of a die:

Face: i	1	2	3	4	5	6	Total
Obs. Freq.: f_i	16	15	14	20	22	13	100

Are these observations consistent with the hypothesis that the die is fair?

7.2.7 Exercise

In a long-term study of heart disease in a large group of men, it was noted that 63 men who had no previous record of heart problems died suddenly of heart attacks. The following table gives the number of such deaths recorded on each day of the week:

Day of Week	Mon.	Tues.	Wed.	Thurs.	Fri.	Sat.	Sun.
No. of Deaths	22	7	6	13	5	4	6

The hypothesis of interest for these data is that the deaths are equally likely to occur on any day of the week. Show that the observed value of the likelihood ratio statistic for testing this hypothesis is 23.3. What would you conclude?

7.3 Likelihood Ratio Tests for Composite Hypotheses

Suppose $X_1, \ldots, X_n$ is a random sample from $f(x;\theta)$ where $\theta \in \Omega$ and Ω is an open set in $\mathcal{R}^k$. Suppose we wish to test $H : \theta \in \Omega_0$ where Ω_0 is an open set in $\mathcal{R}^q$ where $0 < q < k$. The hypothesis H is a composite hypothesis since all the values of the unknown parameters are not specified. For testing composite hypotheses we use the *likelihood ratio statistic*

$$\begin{aligned}\Lambda(X) &= -2\log\left[\frac{\max\limits_{\theta\in\Omega_0} L(\theta;X)}{\max\limits_{\theta\in\Omega} L(\theta;X)}\right] \\ &= 2\left[l(\hat{\theta};X) - \max_{\theta\in\Omega_0} l(\theta;X)\right]\end{aligned}$$

where $X = (X_1, \ldots, X_n)$ are the data and $\hat{\theta} = \hat{\theta}\,(X_1, \ldots, X_n)$ is the M.L. estimator of θ. The asypmptotic distribution of the L.R. statistic if H is true is given by

$$\Lambda(X) \rightarrow_D W \backsim \chi^2(k-q)$$

and an approximate S.L. is given by

$$S.L. \approx P(W \geq -2\log R(\theta_0; x))$$

where $x = (x_1, \ldots, x_n)$ are the observed data.

Note:

The number of degrees of freedom is the difference between the number of paramters that need to be estimated in the model and the number of parameters left to be estimated under the restrictions imposed by H.

7.3.1 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the $\mathrm{N}(\mu, \sigma^2)$ distribution. We wish to test $H : \mu = \mu_0$ where σ^2 is unknown. Find the L.R. statistic for testing H. Consider the test statistic

$$T = \frac{|\bar{X} - \mu_0|}{S/\sqrt{n}}.$$

What is the distribution of

$$\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

if H is true? Compare the two test statistics.

7.3.2 Example

Suppose $X_1, \ldots, X_n$ is a random sample from the EXP(θ_1) distribution and independently $Y_1, \ldots, Y_n$ is a random sample from the EXP(θ_2) distribution. Find the L.R. statistic for testing $H : \theta_1 = \theta_2$. Find the approximate S.L. if the observed data are $n = 10$, $\sum_{i=1}^{10} x_i = 15$ and $\sum_{i=1}^{10} y_i = 20$. What would you conclude?