STAT 330 Notes

Thomas Liu

November 28, 2024

Contents

1 Univariate Random Variables

1.1 Probability

1.1.1 Sample Space

sample space (S) is a set of all distinct outcomes for random experiment, with property that in a single trial, one and only one of these outcome occurs

1.1.2 Sigma Algebra

a collection of subsets of set S is called σ algebra or σ field, denoted by $\mathcal{B}$, it satisfies following properties:

- $\emptyset \in \mathcal{B}$ and $S \in \mathcal{B}$
- $\mathcal{B}$ is closed under complementation
- $\mathcal{B}$ is closed under countable union

The pair $(S, \mathcal{B})$ is called measurable space. Define a probability measure on this space

1.1.3 Probability Set Function

A probability set function is a function P with domain $\mathcal{B}$ that satisfies following axioms:

- $P(A) \geq 0$ for all $A \in \mathcal{B}$
- $P(S) = 1$
- if $A_1, A_2, \cdots \in \mathcal{B}$ are mutually exclusive events, then

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

We call $(S, \mathcal{B}, P)$ a probability space, the three conditions are called Kolmogorove axioms of probability
When writing $P(x)$, we must define $P(x) = P(\{x\})$ to show it is a function on set

1.1.4 Properties

- $P(\emptyset) = 0$
- $P(A) \leq 1$
- $P(A') = 1 - P(A)$
- if $A, B \in \mathcal{B}$, then $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- if $A \subset B$, then $P(A) \leq P(B)$
- Boole's inequality: if $A_1, A_2, \cdots$ is a sequence of events

$$P(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$$

- Bonferroni's inequality: if $A_1, A_2, \cdots, A_k$ are events

$$P(\bigcap_{i=1}^{k} A_i) \geq 1 - \sum_{i=1}^{k} P(A_i^c)$$

- Continuity Property: if $A_1 \subseteq A_2 \subseteq \cdots$ is sequence of nested sets where $A = \bigcup_{i=1}^{\infty} A_i$

$$\lim_{n \to \infty} P(\bigcup_{i=1}^{n} A_i) = P(A)$$

1.1.5 Conditional Probability

conditional probability of event A given B is

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

1.1.6 Independent Events

two events A and B are independent if

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

1.2 Random Variables

random variable X is a function from sample space S to real numbers $\mathbb{R}$

$$X : S \to \mathbb{R}$$

such that $P(X \leq x)$ is defined for all $x \in \mathbb{R}$

1.2.1 Cumulative Distribution Function

cumulative distribution function (CDF) of random variable X is defined as

$$F_X(x) = P(X \leq x) \quad x \in \mathbb{R}$$

It is defined for all real numbers

1.2.2 Properties for CDF

- F is non-decreasing function
$$F(x_1) \leq F(x_2)$$
for all $x_1 < x_2$
- $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
- F is right-continuous function
$$\lim_{x \to a^+} F(x) = F(a)$$
- for all $a < b$
$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$
- for all b
$$P(X = b) = F(b) - \lim_{a \to b^-} F(a)$$

1.3 Discrete Random Variables

random variable X defined on sample space S is discrete random variable if there is a countable subset $A \subset \mathbb{R}$ such that $P(X \in A) = 1$

1.3.1 Probability Function

if X is discrete random variable, then probability function (PF) of X is

$$\begin{aligned} f(x) &= P(X = x) \\ &= F(x) - \lim_{\epsilon \to 0} F(x - \epsilon) \quad x \in \mathbb{R} \end{aligned}$$

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

1.3.2 Properties of PF

- $f(x) \geq 0$ for $x \in \mathbb{R}$
- $\sum_{x \in A} f(x) = 1$

1.4 Continuous Random Variables

X is random variable with CFD F. If F is continuous function for $x \in \mathbb{R}$ and F is differentiable except possibly coutably many points, then X is continuous random variable

1.4.1 PDF

PDF of X is $f(x) = F'(x)$ if F is differentiable at x

1.4.2 Properties

- $f(x) \geq 0$ for $x \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f(x)dx = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1$
- $f(x) = \lim_{h\to 0} \frac{F(x+h)-F(x)}{h} = \lim_{h\to 0} \frac{P(x\leq X\leq x+h)}{h}$ if limit exists
- $F(x) = \int_{-\infty}^{x} f(t)dt$, $x \in \mathbb{R}$
- $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a) = \int_a^b f(x)dx$
- $P(X = b) = F(b) - \lim_{a\to b^-} F(a) = F(b) - F(b) = 0$

1.4.3 Gamma Function

gamma function defined as

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

1.4.4 Properties of Gamma Function

- $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$, $\alpha > 1$
- $\Gamma(n) = (n-1)!$, $n = 1, 2, \cdots$
- $\Gamma(\frac{1}{2}) = \sqrt{\pi}$

1.5 Location and Scale Parameters

1.5.1 Location Parameter

Suppose X is continuous random variable with PDF $f(x;\theta)$ where θ is parameter of distribution. Let $F_0(x) = F(x;\theta = 0)$ and $f_0(x) = f(x;\theta = 0)$. Parameter θ is called location parameter of distribution if

$$F(x;\theta) = F_0(x-\theta) \quad \theta \in \mathbb{R}$$

or

$$f(x;\theta) = f_0(x-\theta) \quad \theta \in \mathbb{R}$$

1.5.2 Scale Parameter

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

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

or

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

1.6 Functions of Random Variables

Theorem

If $Z \sim N(0,1)$ then $Z^2 \sim \chi^2(1)$

1.6.1 Probability Integral Transformation

If X is continuous random variable with cumulative distribution function F then random variable

$$Y = F(X) = \int_{-\infty}^{X} f(t)dt$$

has Uniform(0,1) distribution

Theorem

Suppose F is cdf for continuous random variable. If $U \sim$ Uniform(0,1) then random variable $X = F^{-1}(U)$ also has cdf F

1.6.2 One-to-One Transformation of Random Variable

Suppose X is continuous random variable with pdf f and support set $A = \{x : f(x) > 0\}$. Let $Y = h(x)$ where h is real-valued function. Let $B = \{y : g(y) > 0\}$ be support set of random variable Y. If h is one-to-one function from A to B and $\frac{d}{dx}h(x)$ is continuous for $x \in A$, then pdf of Y is

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

1.7 Expectation

suppose $h(x)$ is real-valued function
If X is discrete random variable with probability function $f(x)$, then expectation of $h(X)$ is

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

provided the sum converges absolutely

$$E(|h(X)|) = \sum_{x \in A} |h(x)|f(x) \leq \infty$$

If X is continuous random variable with pdf $f(x)$, then expectation of $h(X)$ is

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

provided the integral converges absolutely

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

If $E(|h(X)|) = \infty$ then we say $E[h(X)]$ does not exist
$E[h(X)]$ is also called expected value of random variable $h(X)$

1.7.1 Expectation is Linear Operator

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

1.7.2 Special Expectations

- mean of random variable
$$E(X) = \mu$$
- kth moment (about the origin) of random variable
$$E(X^k)$$
- kth moment about the mean of random variable
$$E[(X - \mu)^k]$$
- kth factorial moment of random variable
$$E(X^{(k)}) = E[X(X - 1) \cdots (X - k + 1)]$$
- variance of random variable
$$\text{Var}(X) = E[(X - \mu)^2] = \sigma^2 \quad \mu = E(X)$$

1.7.3 Properties of Variance

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

and

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

1.8 Inequalities

1.8.1 Markov's Inequality

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

1.8.2 Chebyshev's Inequality

$$P(|X-\mu| \geq k\sigma) \leq \frac{1}{k^2} \quad \forall k > 0$$

1.9 Moment Generating Function

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

1.9.1 Moment Generating Function of Linear Function

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

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

1.9.2 Moments from Moment Generating Function

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

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

where

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

is kth derivative of $M(t)$

1.9.3 Uniqueness of Moment Generating Function

Suppose random variable X has mgf $M_X(t)$ and random variable Y has mgf $M_Y(t)$. Also $M_X(t) = M_Y(t)$ for all $t \in (-h, h)$ for some $h > 0$. Then X and Y have same distribution

$$P(X \leq s) = F_X(s) = F_Y(s) = P(Y \leq s) \quad \forall s \in \mathbb{R}$$

2 Joint Distributions

2.1 Definition of Joint Distribution (Bivariate)

Suppose X and Y are random variables defined on a sample space S. Then (X, Y) forms a random vector (bivariate) where the joint cdf of X and Y given by

$$F(x, y) = P(X \leq x, Y \leq y) = P([X \leq x] \cap [Y \leq y]), (x, y) \in \mathbb{R}^2$$

2.1.1 Properties of F

- F is non-decreasing in x for fixed y
- F is non-decreasing in y for fixed x
- $\lim_{x\to-\infty} F(x,y) = 0$ and $\lim_{y\to-\infty} F(x,y) = 0$
- $\lim_{(x,y)\to(-\infty,-\infty)} F(x,y) = 0$ and $\lim_{(x,y)\to(\infty,\infty)} F(x,y) = 1$
- $\lim_{x\to\infty} F(x,y) = F_Y(y)$ and $\lim_{y\to\infty} F(x,y) = F_X(x)$ where $F_X(x)$ and $F_Y(y)$ are cdf of random variables X and Y

2.2 Joint Random Variables

2.2.1 Joint Discrete Random Variables

Suppose X and Y are two discrete random variables. The joint pmf of X and Y is defined as

$$\begin{aligned} f(x,y) &= P(\{w \in S, X(w) = x, Y(w) = y\}) \\ &= P(X = x, Y = y) \qquad\qquad \forall (x,y) \in \mathbb{R}^2 \end{aligned}$$

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

- $f(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$
- $\sum\limits_{(x,y)\in A}\sum f(x,y) = 1$
- For all sets $B \subset \mathbb{R}^2$, $P[(X,Y) \in B] = \sum\limits_{(x,y)\in B}\sum f(x,y)$

2.2.2 Joint Continuous Random Variables

Two random variables X and Y are said to be jointly continuous if there exists a function $f(x,y)$ such that the joint cdf of X and Y is given by

$$F(X,Y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(t_1,t_2)dt_2dt_1 \quad \forall (x,y) \in \mathbb{R}^2$$

The function $f(x,y)$ is called joint density function of X and Y. It follows the definition above when second order partial derivative exists

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

$f(x,y) = 0$ when $\frac{\partial^2}{\partial x\partial y}F(x,y)$ doesn't exist

- $f(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$

- $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$
- For all sets $B \subset \mathbb{R}^2$

$$P[(X,Y) \in B] = \int\int_{(x,y)\in B} f(x,y)dxdy$$

2.3 Marginal Distributions

2.3.1 X and Y Discrete

Suppose X and Y are both discrete random variables with joint pmf $f(x,y)$. The marginal pmf of X and Y are

$$f_X(x) = P(X = x) = \sum_{y\in Supp(Y)} f(x,y) \quad x \in \mathbb{R}$$

$$f_Y(y) = P(Y = y) = \sum_{x\in Supp(X)} f(x,y) \quad y \in \mathbb{R}$$

2.3.2 X and Y Continuous

Suppose X and Y are both continuous random variables with joint pdf $f(x,y)$. The marginal pdf of X and Y are

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy \quad x \in \mathbb{R}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)dx \quad y \in \mathbb{R}$$

2.4 Independent Random Variables

2.4.1 Independence

Two r.v. X and Y with joint cdf $F(x,y)$ are independent iff

$$F(x,y) = F_X(x)F_Y(y) \quad \forall x,y \in \mathbb{R}$$

You can replace F with f, X and Y are independent iff

$$f(x,y) = f_X(x)f_Y(y) \quad \forall (x,y) \in Supp(X,Y)$$

2.4.2 Factorization Theorem for Independence

Suppose X and Y are random variables with joint pf/pdf $f(x,y)$, and marginal pf/pdf $f_X(x)$ and $f_Y(y)$. Suppose

- $A = \{(x,y) : f(x,y) > 0\}$ is support set of (X,Y)
- $A_X = \{x : f_X(x) > 0\}$ is support set of X

- $A_Y = \{y : f_Y(y) > 0\}$ is support set of Y

Then X and Y are independent r.v. iff $A = A_X \times A_Y$ and there exist non-negative functions $g(x)$ and $h(y)$ st

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

for all $(x, y) \in A_X \times A_Y$

2.5 Conditional Distributions

Let X and Y be both discrete with joint pmf $f(x, y) = P(X = x, Y = y)$, then conditional probability function of X given $Y = y$, for $P(Y = y) \neq 0$, is defined as

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

For continuous X and Y, notince $P(Y = y) = 0$, $\forall y \in \mathbb{R}$. So we define conditional pdf of X given $Y = y$ as

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

Conditional Distribution Properties

2.5.1 Discrete Case

- $\sum_x f(x|y) = 1$
- $F(x|y) = \sum_{w:w\leq x} f(w|y)$
- $f(x|y) = F(x|y) - F(x^-|y)$

2.5.2 Continuous Case

- $\int_x f(x|y)dx = 1$
- $F(x|y) = \int_{-\infty}^{x} f(t|y)dt$
- $f(x|y) = \frac{\partial}{\partial x}F(x|y)$

2.5.3 Theorem

Product Rule:
Suppose X and Y are random variables with joint pdf $f(x, y)$ and marginal pdf $f_X(x)$ and $f_Y(y)$, and conditional pdf $f(x|y)$ and $f(y|x)$. Then

$$f(x, y) = f_X(x|y)f_Y(y) = f_Y(y|x)f_X(x)$$

Independence:
X and Y are independent iff $f(x|y) = f_X(x)$ and $f(y|x) = f_Y(y)$

2.6 Expectation of Joint Random Variables

2.6.1 Joint Expectation

Suppose X and Y are random variables with joint pf $f(x, y)$ and support A. Also, suppose $h(x, y)$ is a real-valued function. Then

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

$$X\&Y \text{ continuous} : E[h(X, Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x, y)f(x, y)dxdy$$

provided the joint sum/integral converges absolutely

2.6.2 Linearity of Expectation in Bivariate Case

Suppose X and Y are random variables with joint pmf/pdf $f(x, y)$, a_i, b_i, $i = 1, \cdots, n$ are constants, and $g_i(x, y)$, $i = 1, \cdots, n$ are real-valued functions. Then

$$E[\sum_{i=1}^{n}(a_i g_i(X, Y) + b_i)] = \sum_{i=1}^{n}(a_i E[g_i(X, Y)]) + \sum_{i=1}^{n} b_i$$

provided $E[g_i(X, Y)]$ exist for all $i = 1, \cdots, n$

2.6.3 Covariance

No linear relationship between X and $Y \iff Cov(X, Y) = 0$
If two random variables are independent, then $Cov(X, Y) = 0$, but not the other way around
The covariance of random variables X and Y is defined as

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

where $\mu_X = E(X)$ and $\mu_Y = E(Y)$
By definition, $Cov(X, X) = Var(X)$

2.6.4 Variance of Linear Combination

Theorem 1

Suppose X and Y are random variables and a, b, c are real constants. Then

$$Var(aX + bY + c) = a^2 Var(X) + b^2 Var(Y) + 2ab\ Cov(X, Y)$$

Theorem 2

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

$$\begin{aligned} Var(\sum_{i=1}^{n} a_i X_i) &= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + 2 \sum_{i=1}^{n} \sum_{j=i+1}^{n} a_i a_j Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} a_i^2 \sigma_i^2 + \sum \sum_{i \neq j} a_i a_j Cov(X_i, X_j) \end{aligned}$$

If $X_1, X_2, \cdots, X_n$ are independent, then $Cov(X_i, X_j) = 0$ for all $i \neq j$, and

$$Var(\sum_{i=1}^{n} a_i X_i) = \sum_{i=1}^{n} a_i^2 \sigma_i^2$$

2.7 Correlation

Covariance is a real number depends on units of measurement of X and Y. THe informative part of covariance is its sign, unless it is puut into context.
TO put covariance into context, and to quantitatively measure the strength of a linear relationship, use correlation coefficient

2.7.1 Definition

The correlation coefficient of random variables X and Y is defined as

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

where $\sigma_X = \sqrt{Var(X)}$ and $\sigma_Y = \sqrt{Var(Y)}$

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

2.8 Conditional Expectation

Let g be a real-valued function. The conditional expectation of $g(Y)|X = x$ is given by

$$\begin{aligned} E[g(Y)|X = x] &= \sum_{y \in Supp(Y)} g(y) f(y|x) \quad \text{discrete} \\ &= \int_{-\infty}^{\infty} g(y) f(y|x) dy \quad \text{continuous} \end{aligned}$$

The conditional expecatation of $h(X)|Y = y$, for a real-valued function h, is defined in a similar manner
Based on definition above,

$$\begin{aligned} Var(Y|X = x) &= E[(Y - E(Y|X = x))^2|X = x] \\ &= E(Y^2|X = x) - [E(Y|X = x)]^2 \end{aligned}$$

is the conditional variance

2.8.1 Linearity

The linearity of the expected value applies to condtional expectation as well

- X and Y are two random variables with conditional distribution $f(y|x)$
- a_i and b_i are real constants for $i = 1, \cdots, n$
- $g_i(y)$, $i = 1, \cdots, n$ are real-valued functions

then

$$E[\sum_{i=1}^{n}(a_i g_i(Y) + b_i)|X = x] = \sum_{i=1}^{n}(a_i E[g_i(Y)|X = x]) + \sum_{i=1}^{n} b_i$$

provided $E[g_i(Y)|X = x]$ exists for all $i = 1, \cdots, n$
If X and Y are independent, then $E(g_i(Y)|X = x) = E(g_i(Y))$

Properties

Suppose X and Y are random variables then

$$E(E[g(Y)|x]) = E[g(Y)]$$

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

2.9 Joint Moment Generating Function

The joint moment generating function of random variables X and Y is defined as

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

if this expectation exists for all $(t_1, t_2) \in (-h_1, h_1) \times (-h_2, h_2)$ for some $h_1, h_2 > 0$
More generally, if $X_1, X_2, \cdots, X_n$ are random variables then

$$M(t_1, t_2, \cdots, t_n) = E(e^{t_1 X_1 + t_2 X_2 + \cdots + t_n X_n}) = E[exp \sum_{i=1}^{n} t_i X_i]$$

is called the joint mgf of $X_1, X_2, \cdots, X_n$ if this expectation exists for all $t_i \in \times(-h_i, h_i)$ for some $h_i > 0$, $i = 1, \cdots, n$

2.9.1 Joint Moments and Marginal MGF

Given the joint moment generating function $M(t_1, t_2)$, the joint moments of X and Y are

$$E(X^j Y^k) = \frac{\partial^{j+k}}{\partial t_1^j \partial t_2^k} M(t_1, t_2)|_{\ (t_1,t_2)=(0,0)}$$

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 marginal mgf of X and Y are

$$M_X(t_1) = E(e^{t_1 X}) = M(t_1, 0)$$
$$M_Y(t_2) = E(e^{t_2 Y}) = M(0, t_2)$$

2.9.2 Independence and Joint MGF

Suppose X and Y are random variables with joint mgf $M(t_1, t_2)$ 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 iff

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

for all $t_1 \in (-h_1, h_1)$ and $t_2 \in (-h_2, h_2)$ where

- $M_X(t_1) = M(t_1, 0)$
- $M_Y(t_2) = M(0, t_2)$

2.10 Multinomial Distribution

Suppose $(X_1, X_2, \cdots, X_k)$ are discrete random variables with joint p.f.

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

$x_i = 0, \cdots, n$, $i = 1, \cdots, k+1$, and $x_{k+1} = n - \sum_{i=1}^k x_i$
$0 < p_i < 1$, $i = 1, \cdots, k+1$, and $p_{k+1} = 1 - \sum_{i=1}^k p_i$
Under these conditions, $(X_1, X_2, \cdots, X_k)$ have a multinomial distribution,
$(X_1, \cdots, X_k) \sim MULT(n, p_1, \cdots, p_k)$

2.10.1 Properties of Multinomial Distritbution

Suppose $(X_1, X_2, \cdots, X_k) \sim MULT(n, p_1, \cdots, p_k)$, then

- For all $(t_1, \cdots, t_k) \in \mathbb{R}^k$, the random variable $(X_1, \cdots, X_k)$ has joint mgf

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

- Any set of $X_1, \cdots, X_{k+1}$ also has a multinomial distribution. In particular, $X_i \sim BIN(n, p_i)$, $i = 1, \cdots, k+1$

- If $T = X_i + X_j$, $i \neq j$, then $T \sim BIN(n, p_i + p_j)$
- $Cov(X_i, X_j) = -np_ip_j$, $i \neq j$
- the conditional distirbution of any subset of $(X_1, \cdots, X_{k+1})$ given the rest of coordinates is a multinomial distribution. In particular, the conditional p.f. of X_i given $X_j = x_j$, $i \neq j$, is

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

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

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

2.11 Bivariate Normal Distribution

Let X_1 and X_2 be random variables with joint pdf

$$\begin{aligned} f(x_1, x_2) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\} \end{aligned}$$

where $(x_1, x_2) \in \mathbb{R}^2$ and

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

and Σ is an nonsignular matrix. Then $X = (X_1, X_2)^T$ is said to have a bivariate normal distribution. Write is as $X \sim BVN(\mu, \Sigma)$

- $\mu = \begin{bmatrix} \mu_1 & \mu_2 \end{bmatrix}^T$ is called the mean vector
- Σ is called the variance-covariance matrix, or simply the covariance matrix

2.11.1 Properties of Bivariate Normal Distribution

Suppose $X \sim BVN(\mu, \Sigma)$, where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

- X has joint mgf

$$\begin{aligned} M(t_1, t_2) &= E(e^{t_1X_1 + t_2X_2}) \\ &= \exp\left\{\mu^T \begin{bmatrix} t_1 \\ t_2 \end{bmatrix} + \frac{1}{2}\begin{bmatrix} t_1 & t_2 \end{bmatrix}\Sigma\begin{bmatrix} t_1 \\ t_2 \end{bmatrix}\right\} \qquad \forall (t_1, t_2) \in \mathbb{R}^2 \end{aligned}$$

- $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$
- $Cov(X_1, X_2) = \rho\sigma_1\sigma_2$ and $Corr(X_1, X_2) = \rho$ where $-1 \leq \rho \leq 1$
- X_1 and X_2 are independent iff $\rho = 0$
- if $c = \begin{bmatrix} c_1 & c_2 \end{bmatrix}^T$ is nonzero vector of constants, then

$$c^T X = \sum_{i=1}^{2} c_i X_i \sim N(c^T\mu, c^T\Sigma c)$$

- if A is a 2×2 nonsignular matrix and b is a 2×1 vector then $Y = AX + b \sim BVN(A\mu + b, A\Sigma A^T)$
- $X_2|X_1 = x_1 \sim N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x_1 - \mu_1), \sigma_2^2(1 - \rho^2))$ and $X_1|X_2 = x_2 \sim N(\mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2), \sigma_1^2(1 - \rho^2))$
- $(X - \mu)^T\Sigma^{-1}(X - \mu) \sim \chi^2(2)$

3 Functions of Two or More Random Variables

3.1 One-to-One Transformation

3.1.1 Jacobian of a Bivariate One-To-One Transformation

Consider the one-to-one transformation $S : (x, y) \to (u, v)$ mapping $(x, y) \in R_{XY} = supp((X, Y))$ onto $(u, v) \in R_{UV} = supp((U, V))$. We have

$$u = h_1(x, y) \quad v = h_2(x, y)$$

Since S is one-to-one transformation, there exists a inverse transformation T defined by

$$x = w_1(u, v) \quad y = w_2(u, v)$$

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

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

where $\left|\frac{\partial(x,y)}{\partial(u,v)}\right|$ is the Jacobian of transformation S. Assume all functions are continuously differentiable

3.1.2 Change of Variables

Consider continuous random variables X and Y with joint pdf $f(x, y)$. Define $U = h_1(X, Y)$ and $V = h_2(X, Y)$, where $S : (x, y) \to (u, v)$ is a one-to-one transformation

$$X = w_1(U, V) \quad Y = w_2(U, V)$$

The joint pdf of U and V is

$$g(u, v) = f(w_1(u, v), w_2(u, v)) \times |J| \quad \forall (u, v) \in supp[(U, V)]$$

where $|J| = \left|\frac{\partial(x,y)}{\partial(u,v)}\right|$ is the Jacobian of transformation S^{-1}

3.2 Moment Generating Function Method

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

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

If X_i are independent random variables each with mgf $M(t)$, then Y has mgf

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

3.2.1 Properties

- If $X \sim GAM(\alpha, \beta)$, where α is positive integer, then

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

- If $X_i \sim GAM(\alpha_i, \beta)$, $i = 1, \cdots, n$ independently, then

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

- If $X_i \sim GAM(1, \beta) = EXP(\beta)$, $i = 1, \cdots, n$ independently, then

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

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

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

- If $X_i \sim N(\mu, \sigma^2)$, $i = 1, \cdots, n$ independently, then

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

- If $X_i \sim POI(\mu_i)$, $i = 1, \cdots, n$ independently, then

$$\sum_{i=1}^{n} X_i \sim POII(\sum_{i=1}^{n} \mu_i)$$

- If $X_i \sim BIN(n_i, p)$, $i = 1, \cdots, n$ independently, then

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

- If $X_i \sim NB(k_i, p)$, $i = 1, \cdots, n$ independently, then

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

3.2.2 Gaussian Distribution

If $X_i \sim N(\mu_i, \sigma_i^2)$, $i = 1, \cdots, n$ independently, then

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

Assume $X_i \sim N(\mu, \sigma^2)$, $i = 1, \cdots, n$ independently, and $\overline{X} = \dfrac{1}{n} \sum_{i=1}^{n} X_i$, and $S^2 = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$, then

- $\overline{X} \sim N(\mu, \dfrac{\sigma^2}{n})$
- $\overline{X}$ and S^2 are independent
- $\dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$
- $\dfrac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$

3.2.3 Snedecor's F Distribution

If $X \sim \chi^2(n)$ and $Y \sim \chi^2(m)$ independently, then

$$U = \frac{X/n}{Y/m} \sim F_{n,m}$$

Suppose $X_1, \cdots, X_n$ is random sample from $N(\mu_1, \sigma_1^2)$ and $Y_1, \cdots, Y_m$ is random sample from $N(\mu_2, \sigma_2^2)$. Let $S_1^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$ and $S_2^2 = \sum_{i=1}^m (Y_i - \overline{Y})^2/(m-1)$, then

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{n-1,m-1}$$

4 Limiting or Asymptotic Distributions

4.1 Convergence in Distribution

4.1.1 Definition

Let $X_1, X_2, \cdots$ be a sequence of random variables such that X_n has the CDF $F_n(x)$, $n = 1, 2, \cdots$. Let X be a random variable with CDF $F(x)$. We say X_n converges in distribution to a random variable X and write

$$X_n \xrightarrow{D} X$$

if

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

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

Notes

$X_n \xrightarrow{D} X$ means

$$P(X_n \le a) \approx P(X \le a) \text{ for large n}$$

but does not mean $X_n \approx X$

4.1.2 Taylor Series with Remainder

Suppose $f : [a, b] \to \mathbb{R}$ is infinitely differentiable, and $c \in [a, b]$. Then $\forall x \in [a, b]$ and positive integer k,

$$f(x) = \sum_{i=1}^{k} \frac{f^{(i)}(c)}{i!}(x-c)^i + \frac{f^{(k+1)}(\zeta_x)(x-c)^{k+1}}{(k+1)!}$$

where ζ_x is in the interval $[c, x]$

Two Equations

- If $b, c \in \mathbb{R}$ are constants and $\lim_{n\to\infty} \psi(n) = 0$, then

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

- If $b, c \in \mathbb{R}$ are constants, then

$$\lim_{n\to\infty} \left[1 + \frac{b}{n}\right]^{cn} = e^{bc}$$

4.2 Convergence in Probability

4.2.1 Degenrate Distribution

The function $F(y)$ is the CDF of a degenerate distribution at $y = c$ if

$$F(y) = \begin{cases} 0 & y < c \\ 1 & y \geq c \end{cases}$$

In other words, $F(y)$ is the CDF of a discrete distribution where

$$P(Y = y) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases}$$

If y is degenerate at c, then $E(Y) = c$, $Var(Y) = 0$

4.2.2 Convergence in Probability

A sequence of random variables $X_1, X_2, \cdots$ converges in probability to a random variable X if, for every $\epsilon > 0$,

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

or equivalently

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1$$

We show convergence in probability by

$$X_n \xrightarrow{P} X$$

4.2.3 Relationship Between the Two Convergences

Convergence in probability implies convergence in distribution,

$$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{D} X$$

This shows that convergence in probability is a stronger form of convergence than convergence in distribution

4.3 Weak Law of Large Numbers (WLLN)

4.3.1 Theorem

If $X_1, \cdots, X_n$ is a random sample from a distribution with finite mean μ and variance σ^2, then the sequence of sample means converges in probability to μ, $\overline{X} \xrightarrow{P} \mu$

4.4 MGF Technique For Limiting Distributions

4.4.1 Theorem

Let $Y_1, Y_2, \cdots$ be a sequence of random variables with respective MGFs $M_1(t), M_2(t), \cdots$ defined on a common neighborhood about 0. Then

$$Y_n \xrightarrow{D} Y$$

if and only if

$$\lim_{n\to\infty} M_n(t) = M(t) \quad t \in (-h, h)$$

where $M(t)$ is the MGF of the limiting random variable Y

4.4.2 Central Limit Theorem (CLT)

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

$$\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} Z$$

where $\overline{X}_n = \sum_{i=1}^n X_i/n$ and $Z \sim N(0,1)$

4.5 Limit Theorems

- If $X_n \xrightarrow{D} a$ and $g(x)$ is continuous at $x = a$, then

$$g(X_n) \xrightarrow{D} g(a)$$

- If $X_n \xrightarrow{D} a$ and $Y_n \xrightarrow{D} b$, and $g(x, y)$ are continuous at (a, b) then

$$g(X_n, Y_n) \xrightarrow{D} g(a, b)$$

- (Slutsky's Theorem) If $X_n \xrightarrow{D} X$, $Y_n \xrightarrow{P} b$, and $g(x, b)$ is continuous for all $x \in$ support set of X, then

$$g(X_n, Y_n) \xrightarrow{D} g(X, b)$$

- If $X_n \xrightarrow{D} X$ and $g(x)$ is continuous for all $x \in$ support set of X, then

$$g(X_n) \xrightarrow{D} g(X)$$

4.5.1 Delta Method

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

$$n^b(X_n - a) \xrightarrow{D} X$$

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

$$n^b \left[g(X_n) - g(a)\right] \xrightarrow{D} g'(a)X$$

Corollary

Let $X_1, X_2, \cdots$ be a sequence of IID random variables with mean μ and variance σ^2. Suppose the function $g(x)$ is differentiable at μ and $g'(\mu) \neq 0$, then

$$\sqrt{n}\left[g(\overline{X}_n) - g(\mu)\right] \xrightarrow{D} Z \sim N(0, [g'(\mu)]^2\sigma^2)$$

5 Maximum Likelihood Estimation (One Parameter)

5.1 Terminology

5.1.1 Statistic

A statistic, $T = T(X) = T(X_1, \cdots, X_n)$, is a function of data which does not depend on any unknown parameters. In other words, $T(X)$ is calculated explicitly from realization of $X_1, \cdots, X_n$

5.1.2 Estimator/Estimate

A statistic that is used to estimate an unknown parameter θ like $\tau(\theta)$ is called an estimator. If $T(X)$ estimates $\tau(\theta)$, an observed value of the statistic $t = t(x) = t(x_1, \cdots, x_n)$ is called an estimate of $\tau(\theta)$

Estimator Properties

- Unbiasedness: $E(\tilde{\theta}) = \theta$
- Small variability: $Var(\tilde{\theta})$ small
- Consistency: $\tilde{\theta} \xrightarrow{P} \theta$

5.2 Maximum Likelihood Estimation

Suppose $X_1, \cdots, X_n$ form a simple random sample (IID) from a discrete distribution with pf $f(x;\theta)$. The joint distribution of $X_1, \cdots, X_n$ is

$$P(X_1 = x_1, \cdots, X_n = x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

If $x_1, \cdots, x_n$ is observed sample, then in discrete case:

$$L(\theta) = P(X_1 = x_1, \cdots, X_n = x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta)$$

A natural estimate of θ then is the value which maximizeds the probability of the observed sample

$$\hat{\theta} = arg \max_{\theta} L(\theta)$$

5.2.1 Likelihood and Log-Likelihood Function

Suppose $X_1, \cdots, X_n \overset{iid}{\sim} f(x; \theta)$, then the function:

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

is called the likelihood function of the parameter θ.
The function

$$l(\theta) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

is called the loglikelihood function of the parameter θ

5.2.2 Maximum Likelihood Estimate

The value of θ that maximizes the likelihood function $L(\theta)$, or the log-likelihood function $l(\theta)$, is called the maximum likelihood (M.L.) estimate of θ and is denoted by $\hat{\theta}_{ML}$

$$\hat{\theta}_{ML} = arg \max_{\theta} L(\theta) \ = arg \max_{\theta} l(\theta)$$

The corresponding estimator to $\hat{\theta}_{ML}$ is the maximum likelihood estimator shown by $\tilde{\theta}_{ML}$

5.3 Score and Information Function

5.3.1 Score Function

The score function is defined as

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

Notice that to find the MLE, usually set score function equal to 0 and solve for θ

5.3.2 Information Function

Suppose that $X_1, \cdots, X_n$ are IID random variables with pf/pdf $f(x;\theta)$, and log-likelihood $l(\theta)$. The information function of θ, denoted $I(\theta)$, is a (random) function defined by

$$I(\theta) = I(\theta; X_1, \cdots, X_n) = -\frac{d^2}{d\theta^2} l(\theta)$$

If $X_1, \cdots, X_n$ are replaced by realized values $x_1, \cdots, x_n$, then the information function is a real valued function of θ. If $\hat{\theta}$ is MLE of θ, then $I(\hat{\theta})$ is called the observed information

5.3.3 Expected Information

The information function $I(\theta)$ measures the curvature of the log-likelihood function. However, it is a function of both θ and data $X = (X_1, \cdots, X_n)$
If θ is scalar, then the expected information function is given by

$$J(\theta) = E[I(\theta; X)] = E[-\frac{d^2}{d\theta^2} l(\theta; X)]$$

5.3.4 Invariance properties of MLE

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

$$MLE(h(\theta)) = h(MLE(\theta))$$

5.4 Relative Likelihood and Likelihood Region/Interval

5.4.1 Relative Likelihood

Suppose $X_1, \cdots, X_n \overset{iid}{\sim} f(x;\theta)$ where the likelihood function is $L(\theta)$ and the MLE of θ is $\hat{\theta}$. The relative likelihood function $R(\theta)$ is defined by

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}$$

$0 \leq R(\theta) \leq 1$

5.4.2 Likelihood Region/Interval

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 (LI) for θ

- Commone value for p are $50\%, 10\%, 1\%$
- Given data, values inside the 10% LI are referred to as plausible and values outside this interval as implausible
- Values inside a 50% LI are very implausible
- Values outside a 1% LI are very implausible

5.5 Asymptotic Properties and Limiting Distribution of MLE

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

- Consistency:
$$\tilde{\theta}_n \xrightarrow{P} \theta$$
- Asymptotic Normality:
$$\sqrt{J(\theta)}(\tilde{\theta}_n - \theta) \xrightarrow{D} Z \sim N(0,1)$$
- Asymptotic Distribution of Relative Likelihood:
$$-2\log R(\theta) = 2[l(\tilde{\theta}_n; X) - l(\theta; X)] \xrightarrow{D} W \sim \chi^2(1)$$

where θ is the true but unknown value of the parameter

5.5.1 Asymptotic Variance of MLE

For large values of n, we have
$$Var(\tilde{\theta}_n) \approx [J(\theta)]^{-1}$$
$$[J(\tilde{\theta}_n)]^{1/2}(\tilde{\theta}_n - \theta) \xrightarrow{D} Z \sim N(0,1)$$

Therefore, for sufficiently large n,
$$Var(\tilde{\theta}_n) \approx \frac{1}{J(\hat{\theta}_n)}$$

Since
$$\frac{I(\tilde{\theta}_n)}{J(\theta)} \xrightarrow{P} 1$$

then
$$[I(\tilde{\theta}_n)]^{1/2}(\tilde{\theta}_n - \theta) \xrightarrow{D} Z \sim N(0,1)$$

Also,
$$Var(\tilde{\theta}_n) \approx \frac{1}{I(\tilde{\theta}_n)}$$

5.6 Confidence Interval

5.6.1 Interval Estimators

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

5.6.2 Pivotal Quantity

The random variable $Q(X;\theta)$, which is a function of the data X and unknown parameter θ, is called a pivotal quantity if the distribution of Q does not depend on θ

5.6.3 Asymptotic Confidence Interval

The random variable $Q(X;\theta)$ is called an asymptotic pivotal quantity if the limiting distribution of Q as $n \to\to$ does not depend on θ
Based on asymptotic properties of MLE, both

$$Q_1 = \left[J(\tilde{\theta}_n)\right]^{1/2} (\tilde{\theta}_n - \theta) \overset{n\to\infty}{\sim} N(0,1)$$

and

$$Q_1 = \left[I(\tilde{\theta}_n)\right]^{1/2} (\tilde{\theta}_n - \theta) \overset{n\to\infty}{\sim} N(0,1)$$

are examples of asymptotic pivotal quantities, as they both have limiting $N(0,1)$ distribution

5.6.4 Building Confidence Interval

Basic Idea: Since distribution of Q is known, we find q_1 and q_2 such that $P(q_1 \le Q(X;\theta) \le q_2) = 1-\alpha$, and then solve the inequality for θ such that $P(A(X) \le \theta \le B(X)) = 1-\alpha$. Sometimes, $1-\alpha$ is shown by p. The random interval $(A(X), B(X))$ is called a $100(1-\alpha)\%$ confidence interval for θ

5.6.5 Pivotal Quantity in Location and Scale Families

Let $X = (X_1, \cdots, X_n)$ be a random sample from $f(x;\theta)$ and let $\tilde{\theta} = \tilde{\theta}(X)$ be the MLE of parameter θ based on X

- if θ is location parameter then $Q = \tilde{\theta} - \theta$ is a pivotal quantity
- if θ is scale parameter then $Q = \tilde{\theta}/\theta$ is a pivotal quantity

5.6.6 Asymptotic Pivotal Quantities and Confidence Intervals

When an exact pivotal quantity cannot be constructed, we can use the limiting distribution of the MLE to construct confidence intervalls

$$\sqrt{J(\tilde{\theta}_n)}(\tilde{\theta}_n - \theta) \overset{D}{\to} Z \sim N(0,1)$$

Therefore,

$$P(-z^* \le \sqrt{J(\tilde{\theta}_n)}(\tilde{\theta}_n - \theta) \le z^*) = 1-\alpha$$
$$P(\tilde{\theta}_n - z^*/\sqrt{J(\tilde{\theta}_n)} \le \theta \le \tilde{\theta}_n + z^*/\sqrt{J(\tilde{\theta}_n)}) = 1-\alpha$$

Therefore, one asymptotic $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta}_n \pm z^* \frac{1}{\sqrt{J(\hat{\theta}_n)}}$$

where $P(-z^* \leq Z \leq z^*) = 1-\alpha$ and $Z \sim N(0,1)$
Similarly, since

$$\sqrt{I(\tilde{\theta}_n)}(\tilde{\theta}_n - \theta) \xrightarrow{D} Z \sim N(0,1)$$

an asymptotic $100(1-\alpha)\%$ confidence interval for θ is

$$\hat{\theta}_n \pm z^* \frac{1}{\sqrt{I(\hat{\theta}_n)}}$$

where $P(-z^* \leq Z \leq z^*) = 1-\alpha$ and $Z \sim N(0,1)$

5.7 Confidence Interval VS Likelihood Interval

If a is a value such that $p = 2P(Z \leq a) - 1$ where $Z \sim N(0,1)$, then the likelihood interval $\{\theta : R(\theta) \geq e^{-a^2/2}\}$ is an approximate $100p\%$ confidence interval for θ
Recall that

$$-2\log R(\theta; X) \xrightarrow{D} \chi^2(1)$$

$$\begin{aligned}
P[R(\theta;X) \geq p] &= P[-2\log R(\theta;X) \leq -2\log p] \\
&\approx P(W \leq -2\log p) \\
&= P(Z^2 \leq -2\log p) && Z \sim N(0,1) \\
&= P(-\sqrt{-2\log p} \leq Z \leq \sqrt{-2\log p}) \\
&= 2P(Z \leq \sqrt{-2\log p}) - 1
\end{aligned}$$

6 Maximum Likelihood Estimation (Multi-Parameters)

6.1 Definition

The likelihood function is a k-variate function, where $l(\theta_1, \cdots, \theta_k) = \log L(\theta_1, \cdots, \theta_k)$. We have

$$L(\theta_1, \cdots, \theta_k) = \prod_{i=1}^{n} f(x_i; \theta_1, \cdots, \theta_k)$$

Note that X may or may not be multivariate. Here, the focus is on number of parameters $\theta_1, \cdots, \theta_k$

6.1.1 Multiparamter Score Function

Suppose $X_1, \cdots, X_n \overset{iid}{\sim} f(x;\theta)$ where $\theta = (\theta, \cdots, \theta_k)^T$. The score vector is defined as

$$S(\theta) = S(\theta; x) = \begin{bmatrix} \frac{\partial l}{\partial \theta_1} & \cdots & \frac{\partial l}{\partial \theta_k} \end{bmatrix}^T \quad \theta \in \Omega$$

The maximum likelihood estimator of θ is

$$\hat{\theta} = (\hat{\theta}_1, \cdots, \hat{\theta}_k)^T = arg\max_{\theta} l(\theta_1, \cdots, \theta_k)$$

which is usually calculated by solving simultaneously k estimating equation $S(\theta) = 0$

6.1.2 Observed and Expected/Fisher Information

Suppose $X_1, \cdots, X_n \overset{iid}{\sim} f(x;\theta)$ where $\theta = (\theta_1, \cdots, \theta_k)^T$

- 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, where $\hat{\theta} = MLE(\theta)$
- The expected or Fisher information matrix $J(\theta)$ is a $k \times k$ symmetric matrix whose (i, j) entry is givne by
$$E\left[-\frac{\partial^2}{\partial\theta_i \partial\theta_j} l(\theta)\right]$$
- The set of $\theta = (\theta_1, \cdots, \theta_k)$ values for which $R(\theta) \geq p$ is called a $100p\%$ likelihood region for θ, which is a region in the k-dimensional space $\mathbb{R}^k$

6.2 MLE and Positive Definite Observed Information

To make sure that the solution to the k simultaneous estimating equations $S(\theta) = 0$ is the MLE, the matrix of the second partial derivatives (aka the Hessian matrix H) must be negative definite when calculated at the MLE $\hat{\theta}$ for all non-zero vectors $a = (a_1, \cdots, a_k)^T$

$$a^T H a|_{\theta=\hat{\theta}} < 0$$

This means that the observed information matrix must be positive definite

$$a^T I(\hat{\theta}) a > 0$$

A sufficient condition for $I(\hat{\theta})$ to be positive definite is that $\det(I(\hat{\theta})) > 0$. Another sufficient conditions is that the all eigenvalues of $I(\hat{\theta})$ are positive

6.3 Asymptotic Properties of MLE

Suppose $X = (X_1, \cdots, X_n)$ be a random sample from $f(x;\theta)$, where $\theta = (\theta_1, \cdots, \theta_k)^T \in \Omega$ and the dimension of Ω is k. Let $\tilde{\theta}_n = \tilde{\theta}_n(X_1, \cdots, X_n)$ be the MLE of θ based on X. Also 0_k be a $1 \times k$ vecotr of zeros and let I_k be the $k \times k$ identity matrix. Then under certain (regularity) conditions:

- Consistency:
$$\tilde{\theta}_n \xrightarrow{P} \theta$$
- Asymptotic Normality:
$$(\tilde{\theta}_n - \theta)\,[J(\theta)]^{1/2} \xrightarrow{D} Z \sim MVN(0_k, I_k)$$
- Asymptotic Distribution of Relative Likelihood:
$$-2\log R(\theta; X) = 2[l(\tilde{\theta}_n; X) - l(\theta; X)] \xrightarrow{D} W \sim \chi^2(k)$$

where θ is true but unknown value of the parameter vector

6.4 Asymptotic Variance of MLE

For large values of n, we have

$$Var(\tilde{\theta}_n) \approx [J(\theta)]^{-1} \rightarrow \text{inverse of the Fisher information matrix}$$

For sufficiently large n,

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

6.5 Asymptotic Confidence Region for $\theta = (\theta_1, \cdots, \theta_k)$

Recall that if $Y = (Y_1, \cdots, Y_k)^T \sim MVN(\mu_k, \Sigma_{k\times k})$, then

$$(Y - \mu)^T \Sigma^{-1} (Y - \mu) \sim \chi^2(k)$$

Since

$$(\tilde{\theta} - \theta)\,[J(\theta)]^{1/2} \xrightarrow{D} Z \sim MVN(0_k, I_k)$$

then

$$(\tilde{\theta} - \theta)[J(\theta)](\tilde{\theta} - \theta)^T \xrightarrow{D} W \sim \chi^2(k)$$

Therefore, an approximate $100p\%$ confidence region for θ is

- $\{\theta : (\hat{\theta}_n - \theta)[J(\hat{\theta})](\hat{\theta}_n - \theta)^T \leq c\}$
- $\{\theta : (\hat{\theta}_n - \theta)[I(\hat{\theta})](\hat{\theta}_n - \theta)^T \leq c\}$

in which c is such that $P(W \leq c) = p$ where $W \sim \chi^2(k)$

7 Hypothesis Testing

7.1 Ingredients of Test of Hypothesis

Suppose based on a random sample $X_1, \cdots, X_n \sim f(x;\theta)$, we are interested in testing

$$H_0 : \theta \in \Omega_0 \quad vs \quad H_a : \theta \notin \Omega_0$$

- Discrepancy measure or test statistic: The idea is that the discrepency measure evaluates the consistency of the data with H_0 in some way. To interpret the output of the discrepancy measure, we must know its distribution under H_0. The test statistic is a pivotal quantity, hence its distribution is known
- p-value: If under H_0 the observed test statistic is "extreme", we conclude that the data does not support H_0. We measure this by $p-value$, which is the probability of observing the test statistic value we have observed or something more extreme under H_0
- Making a decision abuot H_0: deciding whether or not to reject H_0

7.2 Likelihood Ratio Tests for Simple Hypotheses

As a starting point, we first focus on testing simple hypotheses ($H_0 : \theta = \theta_0$) which means that under H_0 the value of the scaler/vector parameter θ, hence the distribution, is fully specified (as opposed to cases like $H_0 : \theta \geq 2$)
One popular method to propose a discrepancy measure and carry out the test is using the asymptotic property of the ML estimators. In particular, recall the asymptotic distribution of the likelihood ratio. Under $H_0 : \theta = \theta_0$ we have

$$-2\log(R(\theta_0;X)) = -2\log\left[\frac{L(\theta_0;X)}{L(\tilde{\theta};X)}\right] = 2\left[l(\tilde{\theta};X) - l(\theta_0;X)\right]$$

in which $X = (X_1, \cdots, X_n)$ is a random sample and $\tilde{\theta}$ is the MLE of θ
$\frac{L(\theta_0)}{L(\hat{\theta}_{ML})}$ close to 1 means no evidence against H_0, close to 0 means there is evidence against H_0
Under regularity conditions (one of which being that the support of X does not depend on θ), we have

$$\Lambda(\theta_0) = -2\log[R(\theta_0;X)] \xrightarrow{D} W_k \sim \chi^2(k)$$

where k is the number of parameters in the model (under the general hypothesis) minus the number of parameters under H_0

$$k = dim(\Omega) - dim(\Omega_0)$$

Notice that large values of $\Lambda(\theta_0)$ imply lack of support for H_0 in the data
The asymptotic p-value or the asymptotic significance level of the test is

$$p - value = SL \approx P(W_k \geq \lambda(\theta_0))$$

where $\lambda(\theta_0)$ is observed value $\Lambda(\theta_0)$ given the data

7.2.1 Steps of Likelihood Ration Test

1. Hypotheses: set up H_0 and H_a
2. MLE: find $\tilde{\theta}_{ML}$
3. Test statistic: calculate the discrepency measure

$$\lambda(\theta_0) = -2\log\left[\frac{L(\theta_0; X)}{L(\tilde{\theta}_{ML}; X)}\right]$$

4. Degrees of freedom: calculate the df of the χ^2 distribution

$$dim(\Omega) - dim(\Omega_0)$$

5. P-value: calculate

$$p - value = P(W_k > \lambda(\theta_0))$$

   where $W_k \sim \chi^2(k)$

6. Interpretation: Interpret the p-value and draw conclusion