Chapter 1

Elementary Materials

1.1 Random Variables

- If X is a random variable, the cumulative distribution function $F : \mathbb{R} \to [0,1]$ of X is defined by for all $t \in \mathbb{R}$

$$F(t) = P(X \leq t)$$

- If X is a random variable with cdf F, for any $a < b$

$$P[a < X \leq b] = F(b) - F(a)$$

- If $F : \mathbb{R} \to \mathbb{R}$ is a cumulative distribution function, then

  1. $F(t) \in [0,1]\ \forall t \in \mathbb{R}$
  2. $F(t_1) \leq F(t_2)\ \forall t_1 \leq t_2$
  3. $\lim_{t \to -\infty} F(t) = 0$ and $\lim_{t \to \infty} F(t) = 1$
  4. F is right continuous, i.e. $\lim_{\Delta \downarrow 0} F(t + \Delta) = F(t)$

- The right continuity is a consequence of the definition of $F(t) = P(X \leq t)$ involving the weak inequality "$\leq$". (If the definition were given in terms of strict inequality "$<$", it would have been left continuous.) In order to convince yourself that it is indeed right (not left) continuous, just consider a random variable X which is equal to 0 with probability 1, and note that $F(t) = 0$ for $t < 0$, and $F(t) = 1$ for $t \geq 0$.

- If X is a random variable with cdf F, for any x

$$P[X = x] = F(x) - \lim_{\Delta \downarrow 0} F(x - \Delta)$$

- A random variable is discrete if its distribution function is a step function.

  - If a random variable is discrete, its space, S_X, is either finite or countable.

- If X is discrete, the probability mass function (pmf) of X is

$$p_X(x) = \Pr[X = x] \quad \text{for} \quad x \in S_X$$

- p_X is a pmf on a finite or countable S_X iff
  1. $0 \leq p_X(x) \leq 1 \ \forall x \in S_X$
  2. $\sum_{s \in S_X} p_X(s) = 1$
- If X is a discrete random variable, F_X is a step function with steps at each point x in the support of X.

- A random variable X is continuous if its support is uncountable and there exists a function f_X such that for all A

$$P(X \in A) = \int_A f_X(t)\, dt$$

- If X is a continuous random variable, its distribution function is (absolutely) continuous, and its probability density function (pdf), f_X can be obtained from F_X by

$$f_X(x) = \begin{cases} \frac{dF_X(x)}{dx} & \text{if } F_X \text{ is differentiable at } x \\ \text{an arbitrary number} & \text{otherwise} \end{cases}$$

- Then, for any $a < b$, $P(a < X \leq b) = F_X(b) - F(a)$
- f_X is the pdf of a continuous X iff
  1. $0 \leq f_X(x)$, and
  2. $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- The support of X is the closure set of all values such that $f_X(x) > 0$.

1.2 Transformations

- Suppose $Y = g(X)$.
- In the more rigorous language introduced in the next chapter, we should consider a probability space (S_Y, Γ_Y, P_Y) induced by Y, where $S_Y = g(S_X)$, Γ_Y is a σ-field on S_Y such that for all $D \in \Gamma_Y$, $g^{-1}(D) \in \Gamma_X$, and P_Y satisfies

$$\forall D \in \Gamma_Y, \quad P_Y(D) = P\left(Y^{-1}(D)\right) = P_X\left(g^{-1}(D)\right)$$

  For now, you can ignore this comment.
- We want to determine the distribution of Y

- If X is discrete with pmf p_X, for all $D \in \Gamma_Y$

$$p_Y(D) = \sum_{x \in \{ x \in S_X | g(x) \in D \}} p_X(x)$$

- If X is continuous, pdf f_X, for all $D \in \Gamma_Y$

$$\Pr(Y \in D) = \int_{g^{-1}(D)} f_X(x)\, dx$$

Theorem 1 *Suppose X is a continuous random variable with pdf f_X and support S_X. Let $Y = g(X)$, where g is a 1-1 differentiable function. Then,*

$$f_Y(y) = f_X\left(g^{-1}(y)\right) \left| \frac{dg^{-1}(y)}{dy} \right|$$

Proof. (Sketch)

$$F_Y(y) = \Pr(Y \leq y) = \Pr(g(X) \leq y) = \Pr\left(X \leq g^{-1}(y)\right) = F_X\left(g^{-1}(y)\right)$$

Differentiate both sides w.r.t y to conclude the proof. ■

1.3 Expectations

Definition 1 *If X is a continuous random variable such that $\int_{-\infty}^{\infty} |x| f(x)\, dx < \infty$, then the expectation (mean, expected value) of X is*

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

If X is a discrete random variable such that $\sum_x |x| p_X(x) < \infty$, then the expectation (mean, expected value) of X is

$$E[X] = \sum_x x\ p_X(x)$$

Theorem 2 *If $Y = g(X)$, then*

$$E[Y] = \int_{-\infty}^{\infty} g(x)\ \ f_X(x)\ \ dx$$

if X is continuous, and

$$E[Y] = \sum_x g(x)\, p_X(x)$$

if X is discrete.

Theorem 3 *For* c_1, c_2, g_1, g_2

$$E\left[c_1 g_1(X) + c_2 g_2(X)\right] = c_1 E\left[g_1(X)\right] + c_2 E\left[g_2(X)\right]$$

Definition 2 $\operatorname{Var}(X) = \sigma^2 \equiv E\left[(X-\mu)^2\right]$

Theorem 4 $\operatorname{Var}(X) = E\left[X^2\right] - \left(E[X]\right)^2$

Theorem 5 *For any numbers* a *and* b*, we have that* $\operatorname{Var}(a+bX) = b^2 \operatorname{Var}(X)$.

Definition 3 *Suppose* X *is a random variable such that for some* $h > 0$*, the expectation of* e^{tX} *exists for* $-h < t < h$*. Then, the moment generating function of* X *is defined as*

$$M(t) \equiv E\left[e^{tX}\right]$$

Theorem 6 *For any positive integer* m*,*

$$E[X^m] = M^{(m)}(0)$$

Remark 1 *As for intuition, note that*

$$\begin{aligned} \frac{de^{tX}}{dt} &= Xe^{tX} \\ \frac{d^2e^{tX}}{dt^2} &= X^2e^{tX} \\ &\vdots \end{aligned}$$

from which we obtain

$$\begin{aligned} \left.\frac{de^{tX}}{dt}\right|_{t=0} &= X \\ \left.\frac{d^2e^{tX}}{dt^2}\right|_{t=0} &= X^2 \\ &\vdots \end{aligned}$$

Therefore, we "expect"

$$M^{(m)}(0) = \left.\frac{d^m M(t)}{dt^m}\right|_{t=0} = \left.\frac{d^m E\left[e^{tX}\right]}{dt^m}\right|_{t=0} = E\left[\left.\frac{d^m e^{tX}}{dt^m}\right|_{t=0}\right] = E\left[X^m\right],$$

the only dubious part of the argument being the third equality.

Theorem 7 *For any two random variables* X *and* Y*, with mgf* M_X *and* M_Y *existing in open intervals around* 0

$$F_X(t) = F_Y(t) \quad \text{for all } t \in \mathbb{R} \iff M_X(t) = M_Y(t) \text{ for all } t \in (-h, h)$$

1.4 Homework

1. Let X have the pdf $f(x) = 3x^2$ if $0 < x < 1$, $f(x) = 0$ otherwise.

   (a) Calculate $E[X^3]$

   (b) Derive the density of $Y = X^3$. Hint: Use Theorem 1.

   (c) Calculate $E[Y]$ using (b).

2. Let $f(x) = 2x$ if $0 < x < 1$, and $f(x) = 0$ otherwise be the pdf of X

   (a) Compute $E[1/X]$.

   (b) Find the pdf of $Y = 1/X$. Hint: Use Theorem 1.

   (c) Find the cdf of $Y = 1/X$. Differentiate it and verify that it is the same as derived in (b).

   (d) Compute $E[Y]$ using (b) or (c).

3. A random variable has the density

$$f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \qquad \text{for all } x \in \mathbb{R}$$

   (a) Derive the density of $Y = X^2$. Hint: Note that the cdf of Y is

$$F_Y(y) = P(Y \leq y) = P\left(X^2 \leq y\right) = P\left(-\sqrt{y} \leq X \leq \sqrt{y}\right)$$

   and the density of Y is the derivative of its cdf.

   (b) Derive the density of $Y = aX + b$. Hint: Use Theorem 1.

4. Suppose that X is a positive continuous random variable, with density f_x that is positive only on $(0, b)$. Show that

$$E[X] = \int_0^b (1 - F_X(x))\, dx$$

   Hint: Note that

$$\frac{d\{-(1 - F_X(x))\}}{dx} = f_X(x)$$

   and use integration-by-parts.

1.5 Bernoulli distribution

- A Bernoulli trial is an experiment with two, and only two possible outcome. A variable X has a Bernoulli(p) distribution if:

$$X = \begin{cases} 1 & \text{with probability } \ p \\ 0 & \text{with probability } 1-p \end{cases}, \quad 0 \leq p \leq 1$$

  where $X = 1$ is often termed "success" and p referred to as the probability of success.

- The pmf of X can be written as:

$$p(x) = p^x (1-p)^{1-x}, \quad x = 0, 1$$

-

$$M(t) = \sum_{x=0}^{1} e^{tx} p^x (1-p)^{1-x} = e^t p + 1 - p$$

$$\mu = E[X] = \left. \frac{\partial \left(e^t p + 1 - p\right)}{\partial t} \right|_{t=0} = p$$

$$E\left[X^2\right] = \left. \frac{\partial^2 \left(e^t p + 1 - p\right)}{\partial t^2} \right|_{t=0} = p$$

$$\sigma^2 = \text{Var}(X) = p - p^2$$

1.6 Binomial distribution

- Let X equal the number of observed successes in n (independent) Bernoulli trials, each with (constant) probability of success p. Then X has a *Binomial distribution*, denoted as $b(n, p)$, with pmf

$$p(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \ldots, n \\ 0 & \text{elsewhere.} \end{cases}$$

  where

$$\binom{n}{r} = \frac{n!}{(n-r)!\ r!}$$

- The mgf of a binomial distribution is given by:

$$\begin{aligned} M(t) &= \sum_x e^{tx} p(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} \left(pe^t\right)^x \ (1-p)^{n-x} \\ &= \left[(1-p) + pe^t\right]^n. \end{aligned}$$

  for all real values of t.

- The mean μ and the variance σ^2 of X may be computed from $M(t)$:

$$\begin{aligned}
\mu &= \left.\frac{dM(t)}{dt}\right|_{t=0} = \left.n\left[(1-p)+pe^t\right]^{n-1}pe^t\right|_{t=0} = np, \\
\sigma^2 &= \left.\frac{d^2M(t)}{dt^2}\right|_{t=0} - \mu^2 \\
&= \left.\left(n(n-1)\left[(1-p)+pe^t\right]^{n-2}\left(pe^t\right)^2 + n\left[(1-p)+pe^t\right]^{n-1}pe^t\right)\right|_{t=0} - (np)^2 \\
&= np(1-p).
\end{aligned}$$

- It will be shown later that "If $X_1, X_2, \ldots, X_n$ are independent random variables with MGF equal to $M_1(t), M_2(t), \ldots, M_n(t)$, the MGF of $\sum_{i=1}^n X_i$ is equal to $\prod_{i=1}^n M_i(t)$." This means that if X_i all have the identical MGF $(1-p)+pe^t$, then the MGF of $\sum_{i=1}^n X_i$ should be $[(1-p)+pe^t]^n$.

1.7 Poisson Distribution

- A random variable X, taking values in the non-negative integers, has a *Poisson distribution* with parameter m if it has a pmf of the form:

$$p(x) = \begin{cases} \dfrac{m^x\ e^{-m}}{x!} & x = 0, 1, 2, \ldots \\ 0 & \text{elsewhere} \end{cases}$$

  where $m > 0$.

  – Recall that

$$\sum_{x=0}^{\infty} \frac{m^x}{x!} = e^m$$

- The mgf of a Poisson distribution is given by:

$$\begin{aligned}
M(t) &= \sum_x e^{tx}p(x) = \sum_{x=0}^{\infty} e^{tx}\frac{m^x e^{-m}}{x!} = e^{-m}\sum_{x=0}^{\infty}\frac{(me^t)^x}{x!} = e^{-m}e^{me^t} \\
&= e^{m(e^t-1)}
\end{aligned}$$

  for all real values for t.

- We have that:

$$\begin{aligned}
\mu &= \left.\frac{dM(t)}{dt}\right|_{t=0} = \left.e^{m\left(e^t-1\right)}me^t\right|_{t=0} = m, \\
\sigma^2 &= \left.\frac{d^2M(t)}{dt^2}\right|_{t=0} - \mu^2 = \left.\left(e^{m\left(e^t-1\right)}\left(me^t\right)^2 + e^{m\left(e^t-1\right)}me^t\right)\right|_{t=0} - m^2 = m.
\end{aligned}$$

1.8 Uniform Distribution

- Suppose that $f_X(x) = 1$ for $x \in (0,1)$, and 0 otherwise. We then say that X has a uniform distribution.
- The definition extends in a natural way to the case where the support of X is some other interval of the form (a,b).

Theorem 8 *Suppose that Y is a continuous random variable with cdf equal to G. Then $U = G(Y)$ has a uniform$(0,1)$ distribution.*

Proof. It is because

$$\Pr[U \leq u] = \Pr[G(Y) \leq u] = \Pr\left[Y \leq G^{-1}(u)\right] = G\left(G^{-1}(u)\right) = u.$$

■

Theorem 9 *If U is uniformly distributed on $(0,1)$, then $Y = G^{-1}(U)$ is a random variable with cdf equal to G.*

Proof. It is because

$$\Pr[Y \leq y] = \Pr\left[G^{-1}(U) \leq y\right] = \Pr[U \leq G(y)] = G(y)$$

■

Remark 2 *These two results can be be useful if you want to generate a random variable with cdf equal to G from your computer; it is because many softwares can generate uniform$(0,1)$ random variables.*

1.9 Gamma Distribution

Definition 4 *If $\alpha > 0$, the value of*

$$\int_0^\infty y^{\alpha-1}e^{-y}dy$$

is a positive number. The integral is called the gamma function*, and we write*

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

Remark 3 *See Section 1.17 for properties of the gamma function.*

- In the integral that defines $\Gamma(\alpha)$, let us introduce a new variable by writing $y = \dfrac{x}{\beta}$, where $\beta > 0$. Then

$$\Gamma(\alpha) = \int_0^\infty \left(\frac{x}{\beta}\right)^{\alpha-1} e^{-x/\beta} \left(\frac{1}{\beta}\right) dx$$

or, equivalently,

$$1 = \int_0^\infty \frac{1}{\Gamma(\alpha)\ \beta^\alpha} x^{\alpha-1} e^{-x/\beta} dx.$$

- Since $\alpha > 0$, $\beta > 0$ and $\Gamma(\alpha) > 0$, we see that

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(\alpha)\ \beta^\alpha} x^{\alpha-1} e^{-x/\beta} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

is a pdf of a continuous random variable.

- A random variable X that has a pdf of this form is said to have a *gamma* distribution with parameters α and β. We write that X has a $\Gamma(\alpha, \beta)$ distribution.

- The mgf of a gamma distribution is given by:

$$M(t) = \int_0^\infty e^{tx} \frac{1}{\Gamma(\alpha)\ \beta^\alpha} x^{\alpha-1} e^{-x/\beta}\ dx = \int_0^\infty \frac{1}{\Gamma(\alpha)\ \beta^\alpha} x^{\alpha-1}\ e^{-x(1-\beta t)/\beta}\ dx.$$

Setting $y = x(1-\beta t)/\beta$, $t < 1/\beta$, or $x = \beta y/(1-\beta t)$, we obtain

$$M(t) = \left(\frac{1}{1-\beta t}\right)^\alpha \int_0^\infty \frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y} dy = \left(\frac{1}{1-\beta t}\right)^\alpha, \text{ for } t < \frac{1}{\beta}.$$

- Hence, for a gamma distribution,

$$\mu = \alpha\beta,$$
$$\sigma^2 = \alpha\beta^2.$$

1.9.1 Exponential distribution

- A special case of the gamma distribution is the *exponential* distribution. Setting $\alpha = 1$, we have

$$f(x) = \begin{cases} \dfrac{1}{\beta} e^{-x/\beta} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

or, alternatively, with $\lambda = 1/\beta$,

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

- X is said to have an exponential distribution with parameter λ, with

$$\mu = E[X] = 1/\lambda,$$
$$\sigma^2 = \text{Var}(X) = 1/\lambda^2.$$

1.9.2 χ^2 distribution

- Another special case of the gamma distribution, in which $\alpha = r/2$, where r is a positive integer, and $\beta = 2$. A continuous random variable that has the pdf

$$f(x) = \begin{cases} \dfrac{1}{\Gamma(r/2)\ 2^{r/2}} x^{r/2-1}\, e^{-x/2} & 0 < x < \infty \\ 0 & \text{elsewhere} \end{cases}$$

  is said to have a χ^2 *distribution.* (Take $r = 1$, and compare with Question 3a.)

- The mgf is

$$M(t) = (1-2t)^{-r/2}, \quad t < 1/2$$

- The mean and the variance of a χ^2 distribution are:

$$\mu = (r/2)\ 2 = r,$$
$$\sigma^2 = (r/2)\, 2^2 = 2r.$$

- We call the parameter r the *number of degrees of freedom* of the chi-square distribution. We write that $X \sim \chi^2(r)$ to mean that X has a χ^2 distribution with r degrees of freedom.

1.10 Homework

1. Suppose that the density of X is

$$f_X(x) = \begin{cases} \frac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & \text{otherwise} \end{cases}$$

   for $\theta_1 < \theta_2$. Derive the cumulative distribution of X, the moment generating function of X, the expectation and the variance of X.

2. Suppose that the density of X is

$$f_X(x) = \begin{cases} \lambda\ \exp(-\lambda x) & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   for $\lambda > 0$.

   (a) Derive the cumulative distribution of X, the moment generating function, the expectation and the variance of X.

   (b) Show that X is "memoryless"; that is, for all $x, y > 0$, $\Pr(X > x + y \mid X > x) = \Pr(X > y)$. Recall that the conditional probability of A given B is defined to be

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

3. Suppose that the probability mass of Y is given by

$$\Pr(Y = y) = p\ (1-p)^y \qquad\qquad y = 0, 1, 2, 3, \ldots$$

for $0 < p < 1$.

(a) Derive the distribution function of Y, its moment generating function, expectation, and variance.

(b) Show that Y is memoryless; that is, for k, j nonnegative integers, $\Pr(Y \geq k + j | Y \geq k) = P(Y \geq j)$.

4. A random variable X has a Pareto distribution with parameters $\alpha > 0$ and $\beta > 0$ if its density is

$$f_X(x) = \begin{cases} \frac{\beta\ \alpha^\beta}{x^{\beta+1}} & \alpha < x < \infty \\ 0 & \text{otherwise} \end{cases}$$

(a) Verify that this is a pdf.

(b) Derive the mean and variance of X.

(c) Prove that the variance of X does not exist if $\beta \leq 2$.

1.11 Random vectors and joint distributions

Definition 5 *An n-dimensional random vector is a function from a sample space Ω into $\mathbb{R}^n$.*

Definition 6 *The joint cumulative distribution function (cdf) of (X_1, X_2) is defined as:*

$$F_{X_1,X_2}(x_1, x_2) = P(\{X_1 \leq x_1\} \cap \{X_2 \leq x_2\}) = P[X_1 \leq x_1, X_2 \leq x_2],$$

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

- Therefore,

$$\begin{aligned} P[a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2] = {} & F_{X_1,X_2}(b_1, b_2) - F_{X_1,X_2}(a_1, b_2) \\ & - F_{X_1,X_2}(b_1, a_2) + F_{X_1,X_2}(a_1, a_2) \end{aligned}$$

Definition 7 *A random vector (X_1, X_2) is a discrete random vector if its space, D, is either finite or countable. The joint probability mass function (pmf) of a discrete random vector (X_1, X_2) is defined by: $\forall\ (x_1, x_2) \in D$,*

$$p_{X_1,X_2}(x_1, x_2) = P[X_1 = x_1, X_2 = x_2]$$

Theorem 10 *The pmf uniquely defines the cdf, and has the following properties:*

1. $0 \leq p_{X_1,X_2}(x_1,x_2) \leq 1$

2. $\sum\limits_D \sum p_{X_1,X_2}(x_1,x_2) = \sum_{x_1}\sum_{x_2} p(x_1,x_2) = 1.$

3. *For an event* $A \subset D$, $P[(X_1,X_2) \in A] = \sum\limits_A \sum p_{X_1,X_2}(x_1,x_2)$.

Definition 8 *The support of a discrete random vector* (X_1,X_2), $S \subset D$, *are all the points* (x_1,x_2) *in the space of* (X_1,X_2), D, *such that* $p(x_1,x_2) > 0$.

Definition 9 *A random vector* (X_1,X_2) *with space* S *is continuous if its cdf* $F_{X_1,X_2}(x_1,x_2)$ *can be expressed as:* $\forall\ (x_1,x_2) \in \mathbb{R}^2$,

$$F_{X_1,X_2}(x_1,x_2) = \int_{-\infty}^{x_2}\int_{-\infty}^{x_1} f_{X_1,X_2}(w_1,w_2)\,dw_1dw_2.$$

Theorem 11

$$\frac{\partial^2 F_{X_1,X_2}(x_1,x_2)}{\partial x_1 \partial x_2} = f_{X_1,X_2}(x_1,x_2)$$

Theorem 12 *We call the integrand* $f_{X_1,X_2}(w_1,w_2)$ *the joint probability density function (pdf) of* (X_1,X_2), *and it has the following properties:*

1. $f_{X_1,X_2}(x_1,x_2) \geq 0$

2. $\int\limits_D \int f_{X_1,X_2}(w_1,w_2)\,dw_1dw_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(w_1,w_2)\,dw_1dw_2 = 1.$

3. *For an event* $B \subset D$, $P[(X_1,X_2) \in B] = \int\limits_B \int f_{X_1,X_2}(w_1,w_2)\,dw_1dw_2$.

Definition 10 *The support* $S \subset D$ *of* (X_1,X_2) *contains all points* (x_1,x_2) *such that* $f(x_1,x_2) > 0$.

1.12 Marginal distributions

- Let (X_1,X_2) be a random vector. Then, both X_1 and X_2 are random variables and we can obtain their distributions in terms of the joint distribution of (X_1,X_2).
- $F_{X_1}(x_1) = \lim_{x_2\uparrow\infty} F(x_1,x_2)$
- In the discrete case, $\forall x_1 \in S_{X_1}$, where S_{X_1} is the support of X_1, we have

$$F_{X_1}(x_1) = \sum_{w_1\leq x_1}\sum_{-\infty<x_2<\infty} p_{X_1,X_2}(w_1,x_2) = \sum_{w_1\leq x_1}\left(\sum_{-\infty<x_2<\infty} p_{X_1,X_2}(w_1,x_2)\right),$$
$$p_{X_1}(x_1) = \sum_{-\infty<x_2<\infty} p_{X_1,X_2}(x_1,x_2).$$

- In the continuous case, $\forall x_1 \in S_{X_1}$, we have

$$F_{X_1}(x_1) = \int_{-\infty}^{x_1} \int_{-\infty}^{\infty} f_{X_1,X_2}(w_1, x_2)\ dx_2 dw_1 = \int_{-\infty}^{x_1} \left(\int_{-\infty}^{\infty} f_{X_1,X_2}(w_1, x_2)\, dx_2 \right) dw_1,$$
$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} f_{X_1,X_2}(x_1, x_2)\ dx_2.$$

1.13 Expectation

- Let (X_1, X_2) be a continuous random vector and $Y = g(X_1, X_2)$, where $g : \mathbb{R}^2 \to \mathbb{R}$.
- If

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |g(x_1, x_2)|\ f_{X_1,X_2}(x_1, x_2)\, dx_1 dx_2 < \infty$$

  then $E[Y]$ exists, and

$$E[Y] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_1, x_2)\ f_{X_1,X_2}(x_1, x_2)\ dx_1 dx_2$$

- Likewise, if (X_1, X_2) is discrete, then $E[Y]$ exists if

$$\sum_{x_1} \sum_{x_2} |g(x_1, x_2)|\ p_{X_1,X_2}(x_1, x_2) \ < \infty$$

  and

$$E[Y] = \sum_{x_1} \sum_{x_2} g(x_1, x_2)\, p_{X_1,X_2}(x_1, x_2)$$

- E is a linear operator: let $Y_1 = g_1(X_1, X_2)$ and $Y_2 = g_2(X_1, X_2)$ be random variables whose expectations exist. Then, for any real numbers k_1 and k_2,

$$E[k_1 Y_1 + k_2 Y_2] = k_1 E[Y_1] + k_2 E[Y_2]$$

- The expected value of any function $g(X_2)$ of X_2 can be found in two ways:

$$E[g(X_2)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x_2)\, f(x_1, x_2)\, dx_1 dx_2 = \int_{-\infty}^{\infty} g(x_2)\, f_{X_2}(x_2)\, dx_2$$

- The expected value of the random vector $X = (X_1, X_2)'$ exists if $E[X_1]$ and $E[X_2]$ exist, and is given by:

$$E[X] = \begin{bmatrix} E[X_1] \\ E[X_2] \end{bmatrix}$$

Definition 11 *Let $X = (X_1, X_2)'$ be a random vector. If $E\left[e^{t_1X_2+t_2X_2}\right]$ exists for $|t_1| < h_1$ and $|t_2| < h_2$, where $h_i > 0$, $i = 1, 2$, then*

$$M_{X_1,X_2}(t_1, t_2) \equiv E\left[e^{t_1X_1+t_2X_2}\right]$$

is the moment generating function (mgf) of X. The mgf of a random vector uniquely determines the distribution of the random vector.

$$M_{X_1}(t_1) = M_{X_1,X_2}(t_1, 0)$$

1.14 Multinomial Distribution

- I will not discuss this in class. You are required to read this note and the related discussion on pp.138-139 of the textbook.
- The binomial distribution is generalized to the multinomial distribution as follows.
- Let a random experiment be repeated n independent times. On each repetition, the experiment results in one of k mutually exclusive outcomes, $C_1, C_2, \ldots, C_k$. Let p_i, $i = 1, \ldots k$ be the probability that the outcome is an element of C_i, p_i constant through the n repetitions. Define the random variable X_i to be equal to the number of outcomes that are elements of C_i, $i = 1, 2, \ldots k-1$. Let $x_1, x_2, \ldots, x_{k-1}$ be non-negative integers such that $x_1 + x_2 + \cdots + x_{k-1} \leq n$. Then the probability that exactly x_1 terminations of the experiments are in C_1, exactly x_{k-1} are in $C_{k-1}$and $n - (x_1 + x_2 + \cdots + x_{k-1}) = x_k$ are in C_k is:
$$\frac{n!}{x_1!\, x_2! \ldots x_{k-1}!\, x_k!} p_1^{x_1}\, p_2^{x_2} \cdots p_{k-1}^{x_{k-1}} p_k^{x_k}.$$
This is the *multinomial pmf* of $k-1$ discrete random variables $X_1, X_2, \ldots, X_{k-1}$.
- The mgf of a multinomial distribution is given by
$$M(t_1, \ldots, t_{k-1}) = (p_1e^{t_1} + p_2e^{t_2} + \ldots + p_{k-1}e^{t_{k-1}} + p_k)^n$$
for all real values of $t_1, t_2, \ldots, t_{k-1}$.
- There is a bit of a confusion in the definition. Sometimes it is easier to talk about the distribution of the k-dimensional random vector $X = (X_1, X_2, \ldots, X_k)$, with the understanding that $X_1 + X_2 + \cdots + X_k = n$. The MGF of such X is
$$M(t_1, \ldots, t_k) = \left(p_1e^{t_1} + p_2e^{t_2} + \cdots + p_ke^{t_k}\right)^n$$
which is symmetric and more convenient to work with. Letting $t_2 = \cdots = t_k = 0$, we can see that the MGF of X_1 is equal to
$$\left(p_1e^{t_1} + p_2 + \cdots + p_k\right)^n = \left(p_1e^{t_1} + 1 - p_1\right)^n$$

i.e., the MGF of $b(n, p_1)$. Therefore, X_1 is binomial. By the same reasoning, we can conclude that each $X_m \sim b\left(n, p_m\right)$

- Note that

$$\frac{\partial^2 \left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k}\right)^n}{\partial t_1 \partial t_2} = n\left(n-1\right)\left(p_1 e^{t_1} + p_2 e^{t_2} + \cdots + p_k e^{t_k}\right)^{n-2} p_1 e^{t_1} p_2 e^{t_2}$$

so

$$E\left[X_1 X_2\right] = n\left(n-1\right) p_1 p_2$$

from which we obtain

$$\mathrm{Cov}\left(X_1, X_2\right) = n\left(n-1\right) p_1 p_2 - E\left[X_1\right] E\left[X_2\right] = n\left(n-1\right) p_1 p_2 - n p_1 \cdot n p_2 = -n p_1 p_2$$

1.15 Transformations of Bivariate Random Variables

Theorem 13 *Let $p_{X_1,X_2}\left(x_1, x_2\right)$ be the joint pmf of the discrete random variables X_1 and X_2, with S the support of (X_1, X_2). Let $y_1 = u_1\left(x_1, x_2\right)$ and $y_2 = u_2\left(x_1, x_2\right)$ define a $1-1$ transformation mapping S onto T. Then the joint pmf of Y_1 and Y_2 is given by:*

$$p_{Y_1,Y_2}(y_1, y_2) = \begin{cases} p_{X_1,X_2}\left[w_1(y_1, y_2), w_2(y_1, y_2)\right] & (y_1, y_2) \in T \\ 0 & \textit{elsewhere} \end{cases}$$

where $x_1 = w_1(y_1, y_2)$, $x_2 = w_2(y_1, y_2)$ is the single-valued inverse of $y_1 = u\left(x_1, x_2\right)$, $y_2 = u_2\left(x_1, x_2\right)$. From $p_{Y_1,Y_2}(y_1, y_2)$, we can obtain the marginal pmf of Y_1 and Y_2.

Theorem 14 *In the continuous case, suppose $y_1 = u_1\left(x_1, x_2\right)$ and $y_2 = u_2\left(x_1, x_2\right)$ define a $1-1$ transformation mapping S onto T, the support of (Y_1, Y_2). Then, the joint pdf $f_{Y_1,Y_2}\left(y_1, y_2\right)$ of Y_1 and Y_2 is:*

$$f_{Y_1,Y_2}(y_1, y_2) = \begin{cases} f_{X_1,X_2}\left[w_1(y_1, y_2), w_2(y_1, y_2)\right] \ |J| & (y_1, y_2) \in T \\ 0 & \textit{elsewhere} \end{cases}$$

where J, the Jacobian of the transformation, is given by

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix}$$

Remark 4 *The theorem above is a bivariate counterpart of Theorem 1.*

Example 1 *Suppose that X_1 and X_2 have the joint pdf*

$$f_{X_1,X_2}(x_1, x_2) = \begin{cases} \frac{1}{4}\exp\left(-\frac{x_1+x_2}{2}\right) & x_1 > 0, \quad x_2 > 0 \\ 0 & \textit{otherwise} \end{cases}$$

Let

$$Y_1 = \frac{1}{2}(X_1 - X_2)$$
$$Y_2 = X_2$$

Then

$$X_1 = 2Y_1 + Y_2$$
$$X_2 = Y_2$$

so

$$J = \det \begin{bmatrix} \frac{\partial(2y_1+y_2)}{\partial y_1} & \frac{\partial(2y_1+y_2)}{\partial y_2} \\ \frac{\partial y_2}{\partial y_1} & \frac{\partial y_2}{\partial y_2} \end{bmatrix} = \det \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} = 2$$

so

$$f_{Y_1,Y_2}(y_1, y_2) = \frac{1}{4} \exp\left(-\frac{(2y_1 + y_2) + y_2}{2}\right) |2| = \frac{1}{2} \exp(-y_1 - y_2)$$

for $T = \{(y_1, y_2) : 2y_1 + y_2 > 0, y_2 > 0\}$*, and 0 elsewhere.*

1.16 Homework

Questions #2, #3, and #4 may look lengthy, but you will see that they are solved for you for all practical purpose. Questions #2 and #3 are applications of Theorem 14.

1. The pmf of (X, Y) is

$$p_{X,Y}(x, y) = p^{x+y}(1-p)^{2-x-y} \quad x = 0, 1;\ y = 0, 1$$

for $p \in (0, 1)$

   (a) What is the moment generating function of (X, Y)?

   (b) What is $E[X^2Y]$?

   (c) Let $Z = X + Y$. What is the pmf of Z? What is the moment generating function of Z? What is the expectation of Z?

2. Let X_1 and X_2 be continuous random variables with the joint PDF f_{X_1,X_2}.

   (a) Let

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} X_1 + X_2 \\ X_2 \end{bmatrix}$$

   so

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} = \begin{bmatrix} Y_1 - Y_2 \\ Y_2 \end{bmatrix}$$

Prove that $J = 1$. Hint:

$$J = \det \begin{bmatrix} \frac{\partial(y_1 - y_2)}{\partial y_1} & \frac{\partial(y_1 - y_2)}{\partial y_2} \\ \frac{\partial y_2}{\partial y_1} & \frac{\partial y_2}{\partial y_2} \end{bmatrix} = \det \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix}.$$

(b) Prove that

$$f_{Y_1}(y_1) = \int_{-\infty}^{+\infty} f_{X_1,X_2}(y_1 - y_2, y_2)\, dy_2.$$

Hint: Theorem Theorem 14 and conclude that

$$f_{Y_1,Y_2}(y_1, y_2) = f_{X_1,X_2}(y_1 - y_2, y_2)$$

(c) Conclude that the PDF of $Z = X_1 + X_2$

$$f_Z(z) = \int_{-\infty}^{+\infty} f_{X_1,X_2}(z - y, y)\, dy$$

which is often called the *convolution formula*.

3. The *beta* family of distributions is a continuous family on $(0, 1)$ indexed by two parameters, α and β. The *beta* (α, β) pdf[1] is:

$$f(x) = \begin{cases} \dfrac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

In this question, you are asked to show that if X_1 and X_2 are independent and have $\Gamma(\alpha, 1)$ and $\Gamma(\beta, 1)$ distributions, then $X_1/(X_1 + X_2)$ has the beta(α, β) distribution.

(a) Prove that the joint PDF of X_1 and X_2 is given by

$$\frac{1}{\Gamma(\alpha)} x_1^{\alpha-1} e^{-x_1} \frac{1}{\Gamma(\beta)} x_2^{\beta-1} e^{-x_2} = \frac{1}{\Gamma(\alpha)\Gamma(\beta)} x_1^{\alpha-1} x_2^{\beta-1} e^{-x_1-x_2}$$

Hint: Use a fact that you learned in college, i.e., the joint PDF of two independent random variables is the product of their marginal PDFs.[2]

[1] The PDF is also be written as

$$f(x) = \begin{cases} \dfrac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} & 0 < x < 1 \\ 0 & \text{elsewhere} \end{cases}$$

where

$$B(\alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\,\Gamma(\beta)} = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1} dx.$$

denotes the beta function

[2] This is discussed in Section 3.6.

(b) Let

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} u_1(x_1, x_2) \\ u_2(x_1, x_2) \end{bmatrix} = \begin{bmatrix} x_1 + x_2 \\ \frac{x_1}{x_1 + x_2} \end{bmatrix}$$

Prove that $J = -y_1$. Hint: Write

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} w_1(y_1, y_2) \\ w_2(y_1, y_2) \end{bmatrix} = \begin{bmatrix} y_1 y_2 \\ y_1(1 - y_2) \end{bmatrix}$$

and note that

$$J = \det \begin{bmatrix} y_2 & y_1 \\ 1 - y_2 & -y_1 \end{bmatrix}.$$

(c) Conclude that the joint PDF of Y_1 and Y_2 is

$$\frac{1}{\Gamma(\alpha)\Gamma(\beta)} (y_1 y_2)^{\alpha - 1} [y_1(1 - y_2)]^{\beta - 1} e^{-y_1 y_2 - y_1(1 - y_2)} |-y_1| = \frac{y_2^{\alpha - 1}(1 - y_2)^{\beta - 1}}{\Gamma(\alpha)\Gamma(\beta)} y_1^{\alpha + \beta - 1} e^{-y_1}$$

(d) Prove that the marginal PDF of Y_2 is

$$\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} y_2^{\alpha - 1}(1 - y_2)^{\beta - 1}.$$

Hint: Use the definition of the gamma function, and conclude that

$$\int_0^\infty y_1^{\alpha + \beta - 1} e^{-y_1} dy_1 = \Gamma(\alpha + \beta)$$

4. Prove that the mean and variance of the $beta(\alpha, \beta)$ distribution are

$$\mu = \frac{\alpha}{\alpha + \beta}$$
$$\sigma^2 = \frac{\alpha\beta}{(\alpha + \beta + 1)(\alpha + \beta)^2}$$

Hint: Recognizing that the formula

$$\int_0^1 y^{\alpha - 1}(1 - y)^{\beta - 1} dy = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

for any α and β, we have

$$\int_0^1 y^{(\alpha + 1) - 1}(1 - y)^{\beta - 1} dy = \frac{\Gamma(\alpha + 1)\Gamma(\beta)}{\Gamma(\alpha + 1 + \beta)}$$
$$\int_0^1 y^{(\alpha + 2) - 1}(1 - y)^{\beta - 1} dy = \frac{\Gamma(\alpha + 2)\Gamma(\beta)}{\Gamma(\alpha + 2 + \beta)}$$

so

$$\int_0^1 y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+1)\Gamma(\beta)}{\Gamma(\alpha+1+\beta)}$$
$$\int_0^1 y^2\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\Gamma(\alpha+2)\Gamma(\beta)}{\Gamma(\alpha+2+\beta)}$$

Use (1.1), obtain that

$$\begin{aligned}
\Gamma(\alpha+1) &= \alpha\Gamma(\alpha),\\
\Gamma(\alpha+2) &= (\alpha+1)\Gamma(\alpha+1) = (\alpha+1)\alpha\Gamma(\alpha),\\
\Gamma(\alpha+1+\beta) &= (\alpha+\beta)\Gamma(\alpha+\beta),\\
\Gamma(\alpha+2+\beta) &= (\alpha+1+\beta)\Gamma(\alpha+1+\beta) = (\alpha+1+\beta)(\alpha+\beta)\Gamma(\alpha+\beta).
\end{aligned}$$

and conclude that

$$\int_0^1 y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+\beta)\Gamma(\alpha+\beta)},$$
$$\int_0^1 y\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}dy = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\frac{(\alpha+1)\alpha\Gamma(\alpha)\Gamma(\beta)}{(\alpha+1+\beta)(\alpha+\beta)\Gamma(\alpha+\beta)}.$$

1.17 Technical Details - Properties of Gamma Function

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

- If $\alpha > 1$, an integration by parts shows that:
$$\begin{aligned}
\Gamma(\alpha) &= -y^{\alpha-1}e^{-y}\big|_0^\infty + (\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}dy\\
&= (\alpha-1)\int_0^\infty y^{\alpha-2}e^{-y}dy\\
&= (\alpha-1)\Gamma(\alpha-1). \qquad (1.1)
\end{aligned}$$

  Accordingly, if α is a positive integer greater than 1,
$$\Gamma(\alpha) = (\alpha-1)!$$

- $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$
  - This proof would make sense after we discuss normal distribution later. In any case, we start with
$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty y^{\frac{1}{2}-1}e^{-y}dy = \int_0^\infty \frac{e^{-y}}{y^{\frac{1}{2}}}dy$$

With change of variable $x = y^{\frac{1}{2}}$ and $dx = \frac{1}{2}\frac{dy}{y^{\frac{1}{2}}}$, we have

$$\Gamma\left(\frac{1}{2}\right) = 2\int_0^\infty e^{-x^2}dx = \int_{-\infty}^\infty e^{-x^2}dx$$

where the second equality is by symmetry. Note that

$$\int_{-\infty}^\infty e^{-\frac{x^2}{2\sigma^2}}dx = \sqrt{2\pi}\sigma$$

so taking $\sigma^2 = \frac{1}{2}$, we obtain

$$\Gamma\left(\frac{1}{2}\right) = \int_{-\infty}^\infty e^{-x^2}dx = \sqrt{\pi}.$$

Chapter 2

Probability Spaces, Random Variables, Expectations

2.1 Elementary Set Theory - Mostly for Notations

Definition 12 (Sample Space) *Collection of all possible outcomes. We will often denote it by S.*

Definition 13 (Union) $A \cup B = \{x \in S \mid x \in A \text{ or } x \in B\}$

Definition 14 (Intersection) $A \cap B = \{x \in S \mid x \in A \text{ and } x \in B\}$

Definition 15 (Complement) $A^c = \{x \in S \mid x \notin A\}$

Remark 5 *We will not use the notation $\overline{A}$ to denote the complement of A. It looks confusing relative to the closure of A.*

Definition 16 (Null set) *The empty set, $\varnothing$, is the set with no elements.*

Definition 17 (Subset) $A \subset B$ *if* $[x \in A \Rightarrow x \in B]$. $A = B$ *if* $[A \subset B \text{ and } B \subset A]$.

Definition 18 (Partition) $C_1, \ldots, C_n$ *is a partition of* S *if* $\bigcup_{i=1}^n C_i = S$ *and for all* $i \neq j$, $C_i \cap C_j = \varnothing$

Theorem 15 (DeMorgan's Laws) $(C_1 \cup C_2)^c = C_1^c \cap C_2^c$. $(C_1 \cap C_2)^c = C_1^c \cup C_2^c$.

2.2 Probability Set Function

Definition 19 (σ-field) *A collection of subsets, Γ, in S is a σ-field (or σ-algebra) if*

1. $S \in \Gamma$

2. $A \in \Gamma \quad \Rightarrow \quad A^c \in \Gamma$

3. $A_1, A_2, \ldots \in \Gamma \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \Gamma$

Example 2 *For any $E \in S$, $\{S, \varnothing, E, E^c\}$ is a σ-algebra.*

- A random experiment can be described by a probability space (S, Γ, P), where S is the sample space, Γ is a σ-algebra on S, and P is a probability measure on Γ.

Definition 20 (Probability Set Function) *A set function $P : \Gamma \to \mathbb{R}$ on a σ-algebra is a probability measure if*

1. *For all $A \in \Gamma$, $P(A) \geq 0$*
2. *$P(S) = 1$*
3. *If a sequence $\{A_i\}_{i=1}^{\infty}$ in Γ is such that for all $i \neq j$, $A_i \cap A_j = \varnothing$, then*

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

Theorem 16 *Probability set function satisfies the following:*

1. $P(\varnothing) = 0$
2. $P(A) = 1 - P(A^c)$
3. $P(B) = P(A \cap B) + P(A^c \cap B)$
4. $A \subset B \Rightarrow [P(A) \leq P(B)]$
5. $0 \leq P(A) \leq 1$
6. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

The result "$P(A \cup B) = P(A) + P(B) - P(A \cap B)$" has at least three implications:

Theorem 17 (Inclusion-Exclusion Formula)

$$P(C_1 \cup C_2 \cup C_3) = p_1 - p_2 + p_3$$

where

$$\begin{aligned} p_1 &= P(C_1) + P(C_2) + P(C_3) \\ p_2 &= P(C_1 \cap C_2) + P(C_1 \cap C_3) + P(C_2 \cap C_3) \\ p_3 &= P(C_1 \cap C_2 \cap C_3) \end{aligned}$$

More generally,

$$P\left(\bigcup_{i=1}^{k} C_i\right) = p_1 - p_2 + p_3 - \cdots + (-1)^{k+1} p_k$$

where p_i is the sum of all possible intersections involving i sets.

Theorem 18 (Boole's inequality)

$$P\left(\bigcup_{i=1}^{n} C_i\right) \leq \sum_{i=1}^{n} P\left(C_i\right)$$

Theorem 19 (Bonferroni's Inequality)

$$P\left(C_1 \cap C_2\right) \geq P\left(C_1\right) + P\left(C_2\right) - 1$$

which can be deduced from

$$1 \geq P\left(C_1 \cup C_2\right) = P\left(C_1\right) + P\left(C_2\right) - P\left(C_1 \cap C_2\right)$$

Theorem 20 *If* $C_1 \subset C_2 \subset C_3 \subset \cdots$, *define* $\lim_{i\to\infty} C_i \equiv \cup_{i=1}^{\infty} C_i$. *Then,*

$$P\left(\lim_{i\to\infty} C_i\right) = P\left(\bigcup_{i=1}^{\infty} C_i\right) = \lim_{i\to\infty} P\left(C_i\right)$$

Example 3 *Let* $C_i \equiv \left\{X \leq x - \frac{1}{i}\right\}$, *and convince yourself that* $\cup_{i=1}^{\infty} C_i = \{X < x\}$. *The result above then translates to*

$$P\left(X < x\right) = \lim_{i\to\infty} P\left(X \leq x - \frac{1}{i}\right).$$

Letting $F\left(x\right) \equiv P\left(X \leq x\right)$, *we can see that it can be related to the earlier lesson that* $F\left(x\right)$ *has a limit from the left, and the limit (from the left) is equal to* $P\left(X < x\right)$.

Theorem 21 *If* $C_1 \supset C_2 \supset C_3 \supset \cdots$, *define* $\lim_{i\to\infty} C_i \equiv \cap_{i=1}^{\infty} C_i$. *Then, then*

$$P\left(\lim_{i\to\infty} C_i\right) = P\left(\bigcap_{i=1}^{\infty} C_i\right) = \lim_{i\to\infty} P(C_i)$$

Example 4 *Let* $C_i \equiv \left\{X \leq x + \frac{1}{i}\right\}$, *and convince yourself that* $\cap_{i=1}^{\infty} C_i = \{X \leq x\}$. *The result above then translates to*

$$P\left(X \leq x\right) = \lim_{i\to\infty} P\left(X \leq x + \frac{1}{i}\right).$$

Again, letting $F\left(x\right) \equiv P\left(X \leq x\right)$, *we can see that it can be related to the earlier lesson that* $F\left(x\right)$ *is continuous from the right.*

2.3 Integrals

- It seems intuitive to have a unified notation and write

$$\int 1_A\left(\omega\right) dP\left(\omega\right) = P\left(A\right)$$

  where $1_A\left(\omega\right)$ is the indicator function such that $1_A\left(\omega\right) = 1$ if $\omega \in A$ and $1_A\left(\omega\right) = 0$ otherwise.

- In general, we want to define the integral $\int g(\omega)\, dP(\omega)$. This definition requires a few steps:
  - First we consider the case that g is positive valued. We then consider a class of functions $\varphi(\omega)$ that can be written

$$\varphi(\omega) = \sum_{i=1}^{n} c_i 1_{A_i}(\omega)$$

    i.e., a linear combination of indicator functions. We sometimes call such $\varphi(\omega)$ a simple function. It would be sensible to define

$$\int \varphi(\omega)\, dP(\omega) = \int \left(\sum_{i=1}^{n} c_i 1_{A_i}(\omega) \right) dP(\omega) = \sum_{i=1}^{n} c_i P(X \in A_i).$$

    We then define

$$\int g(\omega)\, dP(\omega) = \sup \left\{ \int \varphi(\omega)\, dP(\omega) : 0 \leq \varphi \leq g,\ \varphi \text{ simple} \right\}$$

  - If g is not positively valued, we write $g(\omega) = g^+(\omega) - g^-(\omega)$, and

$$\int g(\omega)\, dP(\omega) = \int g^+(\omega)\, dP(\omega) - \int g^-(\omega)\, dP(\omega)$$

- This definition of integral preserves your usual intuition:
  - $\int (c_1 g_1(\omega) + c_2 g_2(\omega))\, dP(\omega) = c_1 \int g_1(\omega)\, dP(\omega) + c_2 \int g_2(\omega)\, dP(\omega)$
  - $\left| \int g(\omega)\, dP(\omega) \right| \leq \int |g(\omega)|\, dP(\omega)$

2.4 Random Variables

- Suppose (S, Γ, P) is a probability space. A random variable is a function $X : S \to \mathbb{R}$, which induces a probability space (S_X, Γ_X, P_X) on $\mathbb{R}$, by
  - $S_X = \{ X(s) \in \mathbb{R} \mid s \in S \}$
  - Γ_X is such that for all $B \in \Gamma_X$, $X^{-1}(B) = \{ s \in S \mid X(s) \in B \} \in \Gamma$
  - $\forall B \in \Gamma_X$, $P_X(B) = P(X^{-1}(B))$
- For example, if $B = (a, b)$, we have $X^{-1}(B) = \{\omega : a < X(\omega) < b\}$, and $P_X(B) = P(X^{-1}(B))$.

- In Example 3, you were asked to convince yourself that $\left\{X \leq x - \frac{1}{i}\right\}$, and convince yourself that $\cup_{i=1}^{\infty} \left\{X \leq x - \frac{1}{i}\right\} = \{X < x\}$. Perhaps you could not intuitively convince yourself, and want to relate it to the notations above. It then translates to $\cup_{i=1}^{\infty} \left\{\omega : X(\omega) \leq x - \frac{1}{i}\right\} = \{\omega : X(\omega) < x\}$. This is mathematically correct because we know that $X(\omega) \leq x - \frac{1}{i}$ for some positive integer i if and only if $X(\omega) < x$. Similarly in Example 1.1, we can understand it to mean $\cap_{i=1}^{\infty} \left\{\omega : X(\omega) \leq x + \frac{1}{i}\right\} = \{\omega : X(\omega) \leq x\}$, which is mathematically correct because we know that $X(\omega) \leq x + \frac{1}{i}$ for all positive integer i if and only if $X(\omega) \leq x$.

- We can define
$$E[X] \equiv \int x dP_X(x)$$
starting from
$$\int 1_A(x) dP_X(x) = P(X \in A).$$
On the other hand, we may wonder about the integral
$$\int X(\omega) dP(\omega)$$
understanding $X(\omega)$ as a function of ω. It turns out that they are equal to each other.

- In general,
$$\int g(x) dP_X(x) = \int g(X(\omega)) dP(\omega).$$
We may therefore write with some abuse of notation
$$E[g(X)] = \int g(x) dP(x).$$

2.5 Some Limits

- Given the triple (S, Γ, P), a statement about the elements of S is said to be true a.e. (almost everywhere) or a.s. (almost surely), if $P(T) = 0$ where T is the collection of elements of S on which the statement is not true.

- Let f_n denote a sequence of functions on (S, Γ, P). (A random variable was defined to be such a function, so you can imagine a sequence of random variables.) Assume that $f_n \to f$ almost surely, i.e., $P(\{\omega : f_n(\omega) \nrightarrow f(\omega)\}) = 0$. If there is a function $g \geq 0$ such that (i) $\int g(x) dP(x) < \infty$, and (ii) $|f_n(x)| \leq g(x)$, then $\lim_{n\to\infty} \int f_n(x) dP(x) = \int f(x) dP(x)$. This result is often called the Dominated Convergence Theorem.

- Let f_n denote a sequence of nonnegative-valued functions on (S, Γ, P). Assume that (i) $f_n \leq f_{n+1}$ for all n; and (ii) $f_n \to f$. Then $\lim_{n\to\infty} \int f_n(x) dP(x) = \int f(x) dP(x)$. This result is often called the Monotone Convergence Theorem.

Remark 6 *The Dominated Convergence Theorem protects us from some mathematical pathologies. For this purpose, consider the following sequence of functions defined on* $[0,1]$*:*

$$f_n(x) = \begin{cases} n - n^2 x & 0 \leq x \leq \frac{1}{n} \\ 0 & \frac{1}{n} \leq x \leq 1 \end{cases}$$

It is easy to verify that $f_n(x) \to 0$ *for all* $x \in (0,1]$*, but* $\int_0^1 f_n(x)\,dx = \frac{1}{2} \neq 0 = \int_0^1 0 dx$*.*

Remark 7 *We can understand Theorems 20 and 21 using the Dominated Convergence Theorem. Suppose that the* P *refers to the probability distribution of* X*, i.e., let* $P(C) = E[1_C(X)]$*, e.g. We have*

$$P\left(\lim_{i\to\infty} C_i\right) = E\left[1_{\lim_{i\to\infty} C_i}(X)\right]$$

and

$$\lim_{i\to\infty} E[1_{C_i}(X)] = \lim_{i\to\infty} P(C_i)$$

Because $1_{C_i}(\cdot)$ *is dominated by the constant 1, we would have*

$$E\left[\lim_{i\to\infty} 1_{C_i}(X)\right] = \lim_{i\to\infty} E[1_{C_i}(X)]$$

by Dominated Convergence. Therefore, if $1_{\lim_{i\to\infty} C_i}(X) = \lim_{i\to\infty} 1_{C_i}(X)$*,*[1] *then we obtain*

$$P\left(\lim_{i\to\infty} C_i\right) = \lim_{i\to\infty} P(C_i)$$

Remark 8 *The Dominated Convergence Theorem is often invoked when it is desired to prove continuity an object such as* $E[f(X,t)]$ *seen as a function of* t*. Suppose that* $|f(X,t)| \leq g(X)$ *such that* $E[g(X)] < \infty$*. Also suppose that* $f(X,t) \to f(X,t_0)$ *as* $t \to t_0$ *almost surely. Then, for every sequence* $t_n \to t_0$*, we have*

$$E[f(X,t_n)] = \int f(x,t_n)\,dP(x) \to \int f(x,t_0)\,dP(x) = E[f(X,t_0)].$$

[1] Here's a proof. The condition $C_1 \subset C_2 \subset C_3 \subset \cdots$ or $C_1 \supset C_2 \supset C_3 \supset \cdots$ ensures that the sequence of functions $1_{C_i}(\cdot)$ has a limit; the sequence $1_{C_i}(x)$ for fixed x is a monotone sequence, and because it is bounded by 1 and 0, it has a limit, by the monotone convergence theorem of real numbers. It remains to show that $\lim_{i\to\infty} 1_{C_i}(x) = 1_{\lim_{i\to\infty} C_i}(x)$. Consider first the case that $C_1 \subset C_2 \subset C_3 \subset \cdots$. If x is such that$1_{\lim_{i\to\infty} C_i}(x) = 1$, then $x \in \lim_{i\to\infty} C_i = \cup_{i=1}^{\infty} C_i$, and there is some j such that $x \in C_j$, and as a consequence, $x \in C_i$ for all $i \geq j$. We then have $1_{C_i}(x) = 1$ for all $i \geq j$, and as a consequence, we have $\lim_{i\to\infty} 1_{C_i}(x) = 1$. On the other hand, if x is such that$1_{\lim_{i\to\infty} C_i}(x) = 0$, then $x \in (\lim_{i\to\infty} C_i)^c = (\cup_{i=1}^{\infty} C_i)^c = \cap_{i=1}^{\infty} C_i^c$, so $1_{C_i}(x) = 0$ for all i. As a consequence, we should have $\lim_{i\to\infty} 1_{C_i}(x) = 0$. We therefore have $1_{\lim_{i\to\infty} C_i}(X) = \lim_{i\to\infty} 1_{C_i}(X)$ when $C_1 \subset C_2 \subset C_3 \subset \cdots$. Now consider the case that $C_1 \supset C_2 \supset C_3 \supset \cdots$. If x is such that$1_{\lim_{i\to\infty} C_i}(x) = 1$, then $x \in \lim_{i\to\infty} C_i = \cap_{i=1}^{\infty} C_i$, and as a consequence, $x \in C_i$ for all i. It follows that $1_{C_i}(x) = 1$ for all i, and as a consequence, we have $\lim_{i\to\infty} 1_{C_i}(x) = 1$. If x is such that$1_{\lim_{i\to\infty} C_i}(x) = 0$, then $x \in (\lim_{i\to\infty} C_i)^c = (\cap_{i=1}^{\infty} C_i)^c = \cup_{i=1}^{\infty} C_i^c$, so there is some j such that $x \in C_j^c$, and as a consequence, we have $x \in C_i^c$ for all $i \geq j$. We then have $1_{C_i}(x) = 0$ for all $i \geq j$, and as a consequence, we have $\lim_{i\to\infty} 1_{C_i}(x) = 0$. We therefore have $1_{\lim_{i\to\infty} C_i}(X) = \lim_{i\to\infty} 1_{C_i}(X)$ when $C_1 \supset C_2 \supset C_3 \supset \cdots$.

Remark 9 *Likewise, the Dominated Convergence Theorem is often invoked when it is desired to justify interchanging expectation and differentaition. Suppose that $f(X,t)$ is differentiable almost surely in a neighborhood (a,b) containing t_0, and that $|\partial f(X,t)/\partial t| \leq g(X)$ for all $t \in (a,b)$ with $E[g(X)] < \infty$. We note that*

$$\begin{aligned}\frac{E[f(X,t_n)] - E[f(X,t_0)]}{t_n - t_0} &= \frac{\int f(x,t_n)\,dP(x) - \int f(x,t_0)\,dP(x)}{t_n - t_0} \\ &= \int \frac{f(x,t_n) - f(x,t_0)}{t_n - t_0} dP(x).\end{aligned}$$

We also note that

$$\frac{f(X,t) - f(X,t_0)}{t - t_0} = \left.\frac{\partial f(X,t)}{\partial t}\right|_{t=s}$$

by the mean value theorem, so

$$\left|\frac{f(X,t) - f(X,t_0)}{t - t_0}\right| \leq g(X).$$

By the Dominated Convergence, we have

$$\lim \frac{E[f(X,t_n)] - E[f(X,t_0)]}{t_n - t_0} = \int \lim \frac{f(x,t_n) - f(x,t_0)}{t_n - t_0} dP(x) = \int \frac{\partial f(x,t_0)}{\partial t} dP(x)$$

or

$$\left.\frac{\partial E[f(X,t)]}{\partial t}\right|_{t=t_0} = E\left[\left.\frac{\partial f(X,t)}{\partial t}\right|_{t=t_0}\right]$$

2.6 Inequalities

Theorem 22 *If $E[X^m]$ exists and $k \leq m$, then $E[X^k]$ exists (m, k are integers)*

Proof. We have

$$\begin{aligned}\int |x|^k\,dP(x) &= \int_{|x|\leq 1} |x|^k\,dP(x) + \int_{|x|>1} |x|^k\,dP(x) \\ &\leq \int_{|x|\leq 1} dP(x) + \int_{|x|>1} |x|^m\,dP(x) \\ &\leq \int dP(x) + \int |x|^m\,dP(x) \\ &= 1 + E[|X|^m] < \infty\end{aligned}$$

■

Theorem 23 (Markov's Inequality) *If $E[|X|]$ exists and $a > 0$*

$$\Pr[|X| > a] \leq \frac{E[|X|]}{a}$$

Proof.

$$\begin{aligned} E[|X|] &= \int |x| \, dP(x) \\ &= \int_{-\infty}^{-a} |x| \, dP(x) + \int_{-a}^{a} |x| \, dP(x) + \int_{a}^{+\infty} |x| \, dP(x) \end{aligned}$$

since the first and third terms are large than or equal to a, and the middle term is nonnegative,

$$E[|X|] \geq a \left[\int_{-\infty}^{-a} dP(x) + \int_{a}^{+\infty} dP(x) \right] = a \Pr(|x| \geq a)$$

■

Theorem 24 (Chebyshev's Inequality)

$$\Pr\left[\left| \frac{X - \mu_X}{\sigma_x} \right| > b \right] \leq \frac{1}{b^2}$$

Proof. By Markov's Inequality

$$\Pr\left[\left| \frac{X - \mu_X}{\sigma_x} \right| > b \right] = \Pr\left[\left(\frac{X - \mu_X}{\sigma_x} \right)^2 > b^2 \right] \leq \frac{E[\left(\frac{X - \mu_X}{\sigma_x} \right)^2]}{b^2} = \frac{1}{b^2}$$

■

Theorem 25 (Jensen's Inequality) *If ϕ is a convex function on an open interval I and X is a random variable whose support is contained in I*

$$\phi(E[X]) \leq E[\phi(X)]$$

If ϕ is strictly convex, then the inequality is strict, unless X is a constant random variable.

Proof. Intuitive proof: We only prove the inequality assuming ϕ is differentiable. Let $Y = \phi(X)$ and let μ_X denote the expectation of X. If ϕ is convex,

$$\phi(X) \geq \phi(\mu_X) + \phi'(\mu_X)(X - \mu_X).$$

Taking expectation on both sides, we get

$$E[\phi(X)] \geq \phi(\mu_X) + E[\phi'(\mu_X)(X - \mu_X)] = \phi(\mu_X) = \phi(E[X])$$

■

Remark 10 *A little more rigorous proof, but with the same intuition:*
We first note that the set $S = \{(x, y) : y \geq \phi(x), \quad x \in I\}$ is convex. Let $(x\prime, y'), (x'', y'') \in S$. Then $ty' + (1-t)y'' \geq t\phi(x') + (1-t)\phi(x'') \geq \phi(tx' + (1-t)x'')$, and therefore,
$(tx' + (1-t)x'', ty' + (1-t)y'') \in S$. We now note that $(\mu_X, \phi(\mu_X))$ is a point on the boundary of S. By the supporting hyperplane theorem, we know that there is (α, β) such that $y \geq \alpha + \beta x$ for all $(x, y) \in S$, and $\phi(\mu_X) = \alpha + \beta\mu_X$. Because $(X, \phi(X)) \in S$, we get $\phi(X) \geq \alpha + \beta X$. Taking expectations on both sides, we obtain $E[\phi(X)] \geq \alpha + \beta E[X] = \alpha + \beta\mu_X = \phi(\mu_X)$.

2.7 Homework

1. Let X have the exponential distribution such that $E[X] = 1$. (See Section 1.9.1.) Let $P(A) = E[1_A(X)]$. Let $A_k = \{x : 2 - 1/k < x \leq 3\}$ for $k = 1, 2, 3, \ldots$ Find $\lim_{k\to\infty} A_k$ and $P(\lim_{k\to\infty} A_k)$. Find $P(A_k)$ and $\lim_{k\to\infty} P(A_k)$.

2. Let X be a random variable with mean μ and let $E\left[(X-\mu)^{2k}\right]$ exist, for a positive integer k. Show, with $d > 0$, that

$$\Pr(|X - \mu| \geq d) \leq \frac{E\left[(X-\mu)^{2k}\right]}{d^{2k}}$$

   Hint: Note that $\Pr(|X - \mu| \geq d) = \Pr\left(|X-\mu|^{2k} \geq d^{2k}\right)$.

3. If X is a random variable such that $E[X] = 3$ and $E[X^2] = 13$, use Chebyshev's inequality to determine a lower bound for the probability $\Pr(-2 < X < 8)$.

4. (This question may look lengthy, but you will see that it is solved for you for all practical purpose. It forces you to prove Scheffe's lemma, more or less. In terms of implication, it justifies the undergraduate practice of using the standard normal distribution table if the degrees of freedom of t-distribution is sufficiently large. Textbook proves this in Example 4.3.3, which seems a little less intuitive than the argument here.) Let P denote the distribution of some random variable X. Let $f_n(\cdot)$ denote a sequence of non-negative valued functions such that $\int f_n(x)\, dP(x) = 1$. We will define a sequence of probability distributions by $P_n(A) = \int_A f_n(x)\, dP(x)$. (In other words, we define $P_n(A) = E[1(X \in A) f_n(X)]$.) Suppose that $f_n(\cdot) \to 1$ almost surely.

   (a) Let $g_n = 1 - f_n$. Prove that $\int g_n^+(x)\, dP(x) = \int g_n^-(x)\, dP(x)$. Hint: We have

$$0 = \int f_n(x)\, dP(x) - \int 1 dP(x) = \int g_n(x)\, dP(x) = \int \left(g_n^+(x) - g_n^-(x)\right) dP(x).$$

   (b) Prove that $\int |g_n(x)|\, dP(x) = 2\int g_n^+(x)\, dP(x)$. Hint: We have

$$\int |g_n(x)|\, dP(x) = \int \left(g_n^+(x) + g_n^-(x)\right) dP(x).$$

   Use the result from (a).

   (c) Let $B_n = \{x : g_n(x) \geq 0\}$. Prove that $|P_n(B_n) - P(B_n)| = \frac{1}{2}\int |g_n(x)|\, dP(x)$. Hint:

We have

$$\begin{aligned}
|P_n(B_n) - P(B_n)| &= \left| \int_{g_n(x)\geq 0} f_n(x)\, dP(x) - \int_{g_n(x)\geq 0} dP(x) \right| \\
&= \left| \int_{g_n(x)\geq 0} (f_n(x) - 1)\, dP(x) \right| \\
&= \left| \int_{g_n(x)\geq 0} (-g_n(x))\, dP(x) \right| \\
&= \left| \int_{g_n(x)\geq 0} g_n(x)\, dP(x) \right| \\
&= \left| \int g_n(x)\, 1(g_n(x)\geq 0)\, dP(x) \right| \\
&= \left| \int g_n^+(x)\, dP(x) \right|.
\end{aligned}$$

Use $\int g_n^+(x)\, dP(x) = \frac{1}{2}\int |g_n(x)|\, dP(x)$, which was established in (b).

(d) Prove that $\sup_A |P_n(A) - P(A)| \geq \frac{1}{2}\int |g_n(x)|\, dP(x)$. Hint: We should have

$$\sup_A |P_n(A) - P(A)| \geq |P_n(B_n) - P(B_n)|.$$

Use the result from (c).

(e) Prove that $|P_n(A) - P(A)| \leq \frac{1}{2}\int |g_n(x)|\, dP(x)$. Conclude that $\sup_A |P_n(A) - P(A)| \leq \frac{1}{2}\int |g_n(x)|\, dP(x)$. Hint: We have

$$\begin{aligned}
|P_n(A) - P(A)| &= \left| \int_A f_n(x)\, dP(x) - \int_A dP(x) \right| \\
&= \left| \int_A (f_n(x) - 1)\, dP(x) \right| \\
&= \left| \int_A (-g_n(x))\, dP(x) \right| \\
&= \left| \int_A g_n(x)\, dP(x) \right| \\
&= \left| \int_{A\cap B_n} g_n(x)\, dP(x) + \int_{A\cap B_n^c} g_n(x)\, dP(x) \right|.
\end{aligned}$$

Because $g_n(x) \geq 0$ on $A \cap B_n$, we have

$$\begin{aligned}
\int_{A\cap B_n} g_n(x)\, dP(x) + \int_{A\cap B_n^c} g_n(x)\, dP(x) &\leq \int_{A\cap B_n} g_n(x)\, dP(x) = \int_{A\cap B_n} g_n^+(x)\, dP(x) \\
&\leq \int g_n^+(x)\, dP(x)
\end{aligned}$$

Likewise, because $g_n(x) < 0$ on $A \cap B_n^c$, we have

$$\int_{A\cap B_n} g_n(x)\,dP(x) + \int_{A\cap B_n^c} g_n(x)\,dP(x) \geq \int_{A\cap B_n^c} g_n(x)\,dP(x) = -\int_{A\cap B_n^c} g_n^-(x)\,dP(x)$$
$$\geq -\int g_n^-(x)\,dP(x) = -\int g_n^+(x)\,dP(x)$$

where we used $\int g_n^-(x)\,dP(x) = \int g_n^+(x)\,dP(x)$, which was established in (a). Now recall that $\int g_n^+(x)\,dP(x) = \frac{1}{2}\int |g_n^-(x)|\,dP(x)$, which was established in (b).

(f) Use (d) and (e), and conclude that $\sup_A |P_n(A) - P(A)| = \frac{1}{2}\int |g_n(x)|\,dP(x)$.

(g) Prove that $\int g_n^+(x)\,dP(x) \to 0$. Hint: Because $f_n(\cdot) \to 1$ almost surely, we have $g_n(\cdot) \to 0$ almost surely. As a consequence, we have $|g_n(\cdot)| \to 0$, $g_n^+(\cdot) \to 0$, and $g_n^-(\cdot) \to 0$ almost surely. We also have $0 \leq g_n^+(x) \leq 1$ by definition, so $g_n^+(\cdot)$ is dominated by 1. Use Dominated Convergence.

(h) Prove that $\int g_n^-(x)\,dP(x) \to 0$. Hint: Use the results in (a) and (g).

(i) Prove that $\sup_A |P_n(A) - P(A)| \to 0$. Hint: Use the results in (f), (g), and (h).

(j) Now, let

$$h_n(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \frac{1}{(1+x^2/n)^{(n+1)/2}}$$
$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Prove that

$$\sup_A \left| \int_A h_n(x)\,dx - \int_A \varphi(x)\,dx \right| \to 0.$$

Hint: Let $X \sim N(0,1)$, and

$$f_n(x) = \frac{h_n(x)}{\varphi(x)}.$$

Note that $h_n(x)$ is equal to the density of the t-distribution with n degrees of freedom. See (3.5). By (3.6), we should have $f_n(x) \to 1$ for all x. Using the results in (i), we should get

$$\sup_A |P_n(A) - P(A)| \to 0$$

Now use the fact that

$$P(A) = E[1(X \in A)] = \int_A \varphi(x)\,dx,$$
$$P_n(A) = E[1(X \in A) f_n(X)] = \int 1(x \in A) f_n(x) \varphi(x)\,dx$$
$$= \int_A \frac{h_n(x)}{\varphi(x)} \varphi(x)\,dx = \int_A h_n(x)\,dx.$$

(k) Use (j) and conclude that

$$\int_{-\infty}^{t} h_n(x)\,dx \to \int_{-\infty}^{t} \varphi(x)\,dx$$

for any t. Interpret the result, recalling that h_n and φ are densities of t-distributions and standard normal distribution.

Chapter 3

Random Vectors, Multivariate distributions

3.1 Conditional Probability - Intuitive Approach

Definition 21 (Conditional probability) $\forall A, B \in \Gamma$ *such that* $P(B) > 0$, *the conditional probability of* A *given* B

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

- $P(\cdot|B)$ is a probability measure on S.
- $P(A)$ is called the prior probability of A
- $P(A|B)$ is called the posterior probability of A

Definition 22 *Events* A *and* B *in* Γ *are independent if*

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

Theorem 26 *If* $A, B \in \Gamma$ *are independent, then* A^c *and* B^c *are independent,* A^c *and* B *are independent, and* A *and* B^c *are independent.*

Theorem 27 (Law of Total Probability) *If* $C_1, C_2, \ldots, C_n \in \Gamma$ *is a partition of* S $(n \leq \infty)$ *and* $A \in \Gamma$, *then*

$$P(A) = \sum_{i=1}^{n} P(A|C_i)P(C_i)$$

Theorem 28 (Bayes' Theorem) *If* $C_1, C_2, \ldots, C_n \in \Gamma$ *is a partition of* S $(n \leq \infty)$ *and* $A \in \Gamma$, *then for all* $k = 1, \ldots, n$

$$P(C_k|A) = \frac{P(A|C_k)P(C_k)}{\sum_{i=1}^{n} P(A|C_i)P(C_i)}$$

3.2 Conditional Expectations and Distributions

- Let's start with something simple. For a given random variable Y, consider the minimization problem
$$\min_t E\left[(Y-t)^2\right]$$
It is straightforward to show that the solution is given by $E[Y]$.

- Now, let (X,Y) denote a random vector, and consider the following minimization problem:
$$\min_{\varphi} E\left[(Y-\varphi(X))^2\right]$$
The solution will be called $E[Y|X]$.

- Let $\varphi_*(X)$ denote the solution, and let $\tilde{\varphi}(X)$ be an arbitrary function. We note that
$$\min_t E\left[(Y-(\varphi_*(X)+t\tilde{\varphi}(X)))^2\right]$$
should be minimized at $t=0$, and the FOC should be
$$0 = E[\tilde{\varphi}(X)(Y-\varphi_*(X))]$$
Because this holds for arbitrary[1] $\tilde{\varphi}(X)$, we conclude that we should have
$$0 = E[\tilde{\varphi}(X)(Y-E[Y|X])] \tag{3.1}$$
for arbitrary $\tilde{\varphi}(X)$.

- As for the conditional probability, recall that $P(Y\in A) = E[1_A(Y)]$. We can therefore define $P(Y\in A|X)$ to be $E[1_A(Y)|X]$.

Theorem 29 *We have*
$$E[E[Y|X]] = E[Y]$$
$$\mathrm{Var}(E[Y|X]) \leq \mathrm{Var}(Y)$$

Proof. Taking $\tilde{\varphi}(X)=1$ in (3.1), we obtain the law of iterated expectations:
$$E[E[Y|X]] = E[Y]$$
Taking $\tilde{\varphi}(X) = E[Y|X]-E[Y]$ in (3.1),we can also see that
$$E[(E[Y|X]-E[Y])(Y-E[Y|X])] = 0.$$

[1]It is better to impose some regularity conditions here, but I am avoiding them for simplicity.

It follows that

$$\begin{aligned}
\operatorname{Var}(Y) &= E\left[(Y - E[Y])^2\right] \\
&= E\left[((Y - E[Y|X]) + (E[Y|X] - E[Y]))^2\right] \\
&= E\left[(Y - E[Y|X])^2\right] \\
&+ 2E\left[(Y - E[Y|X])(E[Y|X] - E[Y])\right] \\
&+ E\left[(E[Y|X] - E[Y])^2\right] \\
&= E\left[(Y - E[Y|X])^2\right] + E\left[(E[Y|X] - E[Y])^2\right] \\
&= E\left[(Y - E[Y|X])^2\right] + \operatorname{Var}(E[Y|X]) \\
&\geq \operatorname{Var}(E[Y|X])
\end{aligned}$$

■

Theorem 30 *Let h denote an arbitrary function of X. We have*

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

Proof. Let $E[Y|X] \equiv g(X)$, and note that

$$\begin{aligned}
(h(X)Y - \varphi(X))^2 &= (h(X)Y - h(X)g(X) + h(X)g(X) - \varphi(X))^2 \\
&= (h(X)Y - h(X)g(X))^2 + 2(h(X)g(X) - \varphi(X))(h(X)Y - h(X)g(X)) \\
&+ (h(X)g(X) - \varphi(X))^2.
\end{aligned}$$

Let's consider the expectation of the second term on the right:

$$\begin{aligned}
E[(h(X)g(X) - \varphi(X))(h(X)Y - h(X)g(X))] &= E[(h(X)g(X) - \varphi(X))h(X)(Y - g(X))] \\
&= 0
\end{aligned}$$

where we use (3.1) and the fact that $(h(X)g(X) - \varphi(X))h(X)$ is a function of X. As a consequence, we have

$$\begin{aligned}
E\left[(h(X)Y - \varphi(X))^2\right] &= E\left[(h(X)Y - h(X)g(X))^2\right] + E\left[(h(X)g(X) - \varphi(X))^2\right] \\
&\geq E\left[(h(X)Y - h(X)g(X))^2\right]
\end{aligned}$$

and the minimum on the right can be achieved by setting $\varphi(X) = h(X)g(X)$. ■

Corollary 1 *For an arbitrary event A, we have*

$$E[1(X \in A)Y] = E[1(X \in A)E[Y|X]].$$

Proof. We have

$$E[1(X \in A)Y] = E\{E[1(X \in A)Y|X]\} = E\{1(X \in A)E[Y|X]\}$$

where the first equality is by the law of iterated expectations, and the second equality is by Theorem 30. ■

Theorem 31 *We have*

$$E\left[Y_1+Y_2|\,X\right]=E\left[Y_1|\,X\right]+E\left[Y_2|\,X\right]$$

Proof. Let $E\left[Y_1|\,X\right]\equiv g_1\left(X\right)$,and $E\left[Y_2|\,X\right]\equiv g_2\left(X\right)$. Note that

$$\begin{aligned}\left(Y_1+Y_2-\varphi\left(X\right)\right)^2&=\left(Y_1-g_1\left(X\right)+Y_2-g_2\left(X\right)+\left(g_1\left(X\right)+g_2\left(X\right)\right)-\varphi\left(X\right)\right)^2\\&=\left(Y_1-g_1\left(X\right)+Y_2-g_2\left(X\right)\right)^2\\&+2\left(g_1\left(X\right)+g_2\left(X\right)-\varphi\left(X\right)\right)\left(Y_1-g_1\left(X\right)\right)\\&+2\left(g_1\left(X\right)+g_2\left(X\right)-\varphi\left(X\right)\right)\left(Y_2-g_2\left(X\right)\right)\\&+\left(g_1\left(X\right)+g_2\left(X\right)-\varphi\left(X\right)\right)^2.\end{aligned}$$

Let's consider the expectation of the second term on the right:

$$E\left[\left(g_1\left(X\right)+g_2\left(X\right)-\varphi\left(X\right)\right)\left(Y_1-g_1\left(X\right)\right)\right]=0,$$

where we use (3.1) and the fact that $g_1\left(X\right)+g_2\left(X\right)-\varphi\left(X\right)$ is a function of X. Likewise, we should have

$$E\left[\left(g_1\left(X\right)+g_2\left(X\right)-\varphi\left(X\right)\right)\left(Y_2-g_2\left(X\right)\right)\right]=0.$$

It follows that

$$\begin{aligned}E\left[\left(Y_1+Y_2-\varphi\left(X\right)\right)^2\right]&=E\left[\left(Y_1-g_1\left(X\right)+Y_2-g_2\left(X\right)\right)^2\right]+E\left[\left(g_1\left(X\right)+g_2\left(X\right)-\varphi\left(X\right)\right)^2\right]\\&\geq E\left[\left(Y_1-g_1\left(X\right)+Y_2-g_2\left(X\right)\right)^2\right]\end{aligned}$$

and the minimum on the right can be achieved by setting $\varphi\left(X\right)=g_1\left(X\right)+g_2\left(X\right)$. ■

Theorem 32 *We have*

$$E\left[E\left[Y|\,X,Z\right]|\,Z\right]=E\left[Y|\,Z\right]$$

Proof. Write $E\left[Y|\,X,Z\right]\equiv g\left(X,Z\right)$ and $E\left[Y|\,Z\right]\equiv h\left(Z\right)$. For arbitrary $\tilde{\varphi}\left(Z\right)$, we can use (3.1) and conclude that

$$\begin{aligned}0&=E\left[\tilde{\varphi}\left(Z\right)\left(Y-h\left(Z\right)\right)\right]\\&=E\left[\tilde{\varphi}\left(Z\right)\left(Y-g\left(X,Z\right)\right)\right]+E\left[\tilde{\varphi}\left(Z\right)\left(g\left(X,Z\right)-h\left(Z\right)\right)\right]\end{aligned}$$

Because $\tilde{\varphi}\left(Z\right)$ can be veiwed as a function of (X,Z), we can use (3.1) and conclude that

$$E\left[\tilde{\varphi}\left(Z\right)\left(Y-g\left(X,Z\right)\right)\right]=0.$$

To summarize, we have

$$E\left[\tilde{\varphi}\left(Z\right)\left(g\left(X,Z\right)-h\left(Z\right)\right)\right]=0 \tag{3.2}$$

for arbitrary $\tilde{\varphi}\left(Z\right)$.

Now, consider the minimization problem

$$\min_{\varphi} E\left[\left(g\left(X,Z\right)-\varphi\left(Z\right)\right)^{2}\right]. \tag{3.3}$$

For any fixed φ, we have

$$\begin{aligned}
&E\left[\left(g\left(X,Z\right)-\varphi\left(Z\right)\right)^{2}\right]\\
&=E\left[\left(g\left(X,Z\right)-h\left(Z\right)+h\left(Z\right)-\varphi\left(Z\right)\right)^{2}\right]\\
&=E\left[\left(g\left(X,Z\right)-h\left(Z\right)\right)^{2}\right]+2E\left[\left(h\left(Z\right)-\varphi\left(Z\right)\right)\left(g\left(X,Z\right)-h\left(Z\right)\right)\right]\\
&+E\left[\left(h\left(Z\right)-\varphi\left(Z\right)\right)^{2}\right]\\
&=E\left[\left(g\left(X,Z\right)-h\left(Z\right)\right)^{2}\right]+E\left[\left(h\left(Z\right)-\varphi\left(Z\right)\right)^{2}\right],
\end{aligned}$$

where the last equality is based on

$$E\left[\left(h\left(Z\right)-\varphi\left(Z\right)\right)\left(g\left(X,Z\right)-\varphi\left(Z\right)\right)\right]=0$$

which can be obtained by taking $\tilde{\varphi}\left(Z\right)=h\left(Z\right)-\varphi\left(Z\right)$ in (3.2). It follows that

$$\begin{aligned}
E\left[\left(g\left(X,Z\right)-\varphi\left(Z\right)\right)^{2}\right] &= E\left[\left(g\left(X,Z\right)-h\left(Z\right)\right)^{2}\right]+E\left[\left(h\left(Z\right)-\varphi\left(Z\right)\right)^{2}\right]\\
&\geq E\left[\left(g\left(X,Z\right)-h\left(Z\right)\right)^{2}\right]
\end{aligned}$$

and the minimum on the RHS can be achieved by taking $\varphi\left(Z\right)=h\left(Z\right)$. We conclude that the solution to the minimization problem (3.3) is $h\left(Z\right)$, which means that $E\left[g\left(X,Z\right)|\,Z\right]=h\left(Z\right)$. Recalling the definitions of $g\left(X,Z\right)$ and $h\left(Z\right)$, we get the desired conclusion. ■

3.3 Intuitive Approach

- When we confine our attention to the two elementary cases, an intuitive connection to the conditional probability can be made. Recall that the conditional probability of A given B is defined to be
$$\Pr\left(A|\,B\right)=\frac{\Pr\left(A\cap B\right)}{\Pr\left(B\right)}.$$

- Let X_1 and X_2 be two discrete random variables with joint pmf $p_{X_1,X_2}\left(x_1,x_2\right)$. Let $p_{X_1}\left(x_1\right)$ and $p_{X_2}\left(x_2\right)$ denote the marginal probability density functions of X_1, and X_2, and x_1 be a point in the support of X_1 (i.e., $p_{X_1}\left(x_1\right)>0$). Then, for all $x_2\in S_{X_2}$,
$$P\left(X_2=x_2\mid\ X_1=x_1\right)=\frac{P(X_1=x_1,X_2=x_2)}{P(X_1=x_1)}=\frac{p_{X_1,X_2}\left(x_1,x_2\right)}{p_{X_1}\left(x_1\right)}$$

- Define this function as
$$p_{X_2|X_1}\left(x_2|\,x_1\right)\equiv\frac{p_{X_1,X_2}\left(x_1,x_2\right)}{p_{X_1}\left(x_1\right)},\quad x_2\in S_{X_2}.$$

- For any fixed x_1 such that $p_{X_1}(x_1) > 0$, $p_{X_2|X_1}(x_2|\, x_1)$ is a pmf.
- $p_{X_2|X_1}(x_2|\, x_1)$ is the *conditional pmf* of X_2 given $X_1 = x_1$.
- In the continuous case, the *conditional pdf* of X_2 given $X_1 = x_1$, $f_{X_2|X_1}(x_2|\, x_1)$, is defined by:

$$f_{X_2|X_1}(x_2|\, x_1) \equiv \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)} \tag{3.4}$$

  where $f_{X_1}(x_1) > 0$. Clearly, $f_{X_2|X_1}(x_2|\, x_1)$ satisfies the properties of a pdf.
- Since each $f_{X_1|X_2}(x_1|\, x_2)$ and $f_{X_2|X_1}(x_2|\, x_1)$ is a pdf of one random variables, we can compute probabilities and mathematical expectations. For example,

$$P(a < X_1 < b \mid X_2 = x_2) = \int_a^b f_{X_1|X_2}(x_1|\, x_2)\ dx_1$$

$$P(c < X_2 < d \mid X_1 = x_1) = \int_c^d f_{X_2|X_1}(x_2|\, x_1)\ dx_2$$

- If $u(X_2)$ is a function of X_2, the conditional expectation of $u(X_2)$ given $X_1 = x_1$, if it exists, is given by:

$$E[u(X_2) \mid x_1] = \int_{-\infty}^{\infty} u(x_2)\ f_{X_2|X_1}(x_2|\, x_1)\ dx_2$$

  In particular, if they do exist,

$$E[X_2 \mid x_1] = \int_{-\infty}^{\infty} x_2\ f_{X_2|X_1}(x_2|\, x_1)\ dx_2$$

  is the *conditional mean* of X_2 given $X_1 = x_1$, whereas

$$\begin{aligned} \mathrm{Var}(X_2 \mid x_1) &= E\left[(X_2 - E[X_2|\, x_1])^2 \mid x_1\right] \\ &= E\left[X_2^2|\, x_1\right] - (E[X_2|\, x_1])^2 \end{aligned}$$

  is the *conditional variance* of X_2 given $X_1 = x_1$.
- With discrete random variables, these conditional probabilities and conditional expectations are computed by using summation instead of integration.
- The formal definition in the previous section is general enough to accommodate cases that go beyond the discrete and/or continuous distributions, although in many cases, your intuition would work just fine.
  - See Questions #6 and #7. In these questions, we have a vector (Y, X) with Y continuous and X discrete, more or less.

3.4 Digression - Connection between Intuitive and Formal Approaches

- We now adopt an undergraduate framework, and focus on purely discrete or continuous random variables.

- Suppose that (X_1, X_2) is discrete. Recall that for any function $u(X_2)$ of X_2, we formally define $E[u(X_2)|X_1]$ to be the solution to the problem

$$\min_{\varphi} E\left[\left(u(X_2) - \varphi(X_1)\right)^2\right].$$

Let $\tilde{x}_1$ denote a element in the support of X_1, and consider $\tilde{\varphi}(x) \equiv 1(x = \tilde{x}_1)/p_{X_1}(\tilde{x}_1)$. Then the counterpart of (3.1) is

$$\begin{aligned}
0 &= E\left[\tilde{\varphi}(X_1)\left(u(X_2) - E[u(X_2)|X_1]\right)\right] \\
&= \sum_{x_1}\sum_{x_2} p_{X_1,X_2}(x_1, x_2)\,\tilde{\varphi}(x_1)\left(u(x_2) - E[u(X_2)|x_1]\right) \\
&= \sum_{x_2} \frac{p_{X_1,X_2}(\tilde{x}_1, x_2)}{p_{X_1}(\tilde{x}_1)}\left(u(x_2) - E[u(X_2)|\tilde{x}_1]\right) \\
&= \sum_{x_2} \frac{p_{X_1,X_2}(\tilde{x}_1, x_2)}{p_{X_1}(\tilde{x}_1)} u(x_2) - \frac{\sum_{x_2} p_{X_1,X_2}(\tilde{x}_1, x_2)}{p_{X_1}(\tilde{x}_1)} E[u(X_2)|\tilde{x}_1]
\end{aligned}$$

Because

$$\sum_{x_2} p_{X_1,X_2}(\tilde{x}_1, x_2) = p_{X_1}(\tilde{x}_1),$$

we get

$$0 = \sum_{x_2} p_{X_2|X_1}(x_2|\tilde{x}_1)\,u(x_2) - E[u(X_2)|\tilde{x}_1]$$

- Now suppose that (X_1, X_2) is continuous. Let $\tilde{x}_1$ denote a element in the support of X_1, and consider

$$\tilde{\varphi}_t(X_1) = 1(X_1 \leq t).$$

Using (3.4), we can write (3.1) as

$$\begin{aligned}
0 &= E\left[\tilde{\varphi}_t(X_1)\left(u(X_2) - E[u(X_2)|X_1]\right)\right] \\
&= \int_{-\infty}^{t}\int f_{X_1,X_2}(x_1, x_2)\left(u(x_2) - E[u(X_2)|x_1]\right)dx_2 dx_1 \\
&= \int_{-\infty}^{t} f_{X_1}(x_1)\int \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_1}(x_1)}\left(u(x_2) - E[u(X_2)|x_1]\right)dx_2 dx_1 \\
&= \int_{-\infty}^{t} f_{X_1}(x_1)\left(\left(\int f_{X_2|X_1}(x_2|x_1)\,u(x_2)\,dx_2\right) - \left(\int f_{X_2|X_1}(x_2|x_1)\,dx_2\right)E[u(X_2)|x_1]\right)dx_1.
\end{aligned}$$

Using

$$\int f_{X_2|X_1}(x_2|\,x_1)\,dx_2 = 1,$$

we obtain

$$0 = \int_{-\infty}^{t} f_{X_1}(x_1)\left(\left(\int f_{X_2|X_1}(x_2|\,x_1)\,u(x_2)\,dx_2\right) - E\left[u(X_2)|\,x_1\right]\right)dx_1,$$

which holds for all t. Differentiating this expression with respect to t, we obtain

$$0 = \left(\int f_{X_2|X_1}(x_2|\,x_1)\,u(x_2)\,dx_2\right) - E\left[u(X_2)|\,x_1\right]$$

3.5 Correlation Coefficient

- Let X and Y be two random variables with:

$$\begin{aligned} E[X] &= \mu_X, \qquad & E[Y] &= \mu_Y, \\ \mathrm{Var}(X) &= \sigma_X^2, \qquad & \mathrm{Var}(Y) &= \sigma_Y^2. \end{aligned}$$

  The *covariance* of X and Y, $\mathrm{Cov}(X,Y)$, is given by:

$$\begin{aligned} \mathrm{Cov}(X,Y) &\equiv E[(X-\mu_X)(Y-\mu_Y)] \\ &= E[XY - \mu_Y X - \mu_x Y + \mu_X\mu_Y] \\ &= E[XY] - \mu_Y E[X] - \mu_x E[Y] + \mu_X\mu_Y \\ &= E[XY] - \mu_X\mu_Y. \end{aligned}$$

- If σ_X and σ_Y are strictly positive, the *correlation coefficient* of X and Y, ρ, is defined by:

$$\rho = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\ \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\ \sigma_Y}.$$

  - $-1 \le \rho \le 1$.
  - If $\rho = 1$, there is a perfect linear relationship between X and Y, i.e., there is a line with equation: $y = a + bx$ with $b > 0$ such that $P(Y = a + bX) = 1$
  - If $\rho = -1$, we also have a perfect linear relationship, now with $b < 0$.

- From the above equations,

$$E[XY] = \mu_X\mu_Y + \rho\sigma_X\sigma_Y = \mu_X\ \mu_Y\ + \mathrm{Cov}(X,Y).$$

- Suppose (X,Y) have a joint distribution with σ_X^2 and σ_Y^2 finite and positive. If $E[Y|\,X]$ is linear in X, then

$$E[Y|\,X] = \mu_Y + \frac{\rho\sigma_Y}{\sigma_X}(X-\mu_X)$$

  and

$$E[\mathrm{Var}(Y|\,X)] = \sigma_Y^2(1-\rho^2).$$

– If $E[Y|X] = a + bX$, we have

$$\begin{aligned} E[Y] &= E[E[Y|X]] = E[a + bX] = a + bE[X] \\ E[XY] &= E[E[XY|X]] = E[XE[Y|X]] \\ &= E[X(a + bX)] = aE[X] + bE\left[X^2\right] \end{aligned}$$

Because

$$\begin{aligned} E[XY] &= \mathrm{Cov}(X, Y) + E[X]E[Y] \\ E\left[X^2\right] &= \mathrm{Var}(X) + (E[X])^2, \end{aligned}$$

we obtain

$$\begin{aligned} \mu_Y &= a + \mu_X b \\ \rho\sigma_X\sigma_Y + \mu_X\mu_Y &= \mu_X a + \left(\sigma_X^2 + \mu_X^2\right) b \end{aligned}$$

which can be solved for a and b to yield

$$\begin{aligned} a &= \mu_Y - \frac{\rho\sigma_Y}{\sigma_X}\mu_X \\ b &= \frac{\rho\sigma_Y}{\sigma_X} \end{aligned}$$

- Recall the definition of the mgf for the random vector (X, Y). As for random variables, the joint mgf also gives explicit formulas for certain moments:

$$\frac{\partial^{k+m} M(t_1, t_2)}{\partial t_1^k \, \partial t_2^m} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k \; y^m \; e^{t_1 x + t_2 y} \; f(x, y) \; dx \; dy$$

so that

$$\left.\frac{\partial^{k+m} M(t_1, t_2)}{\partial t_1^k \, \partial t_2^m}\right|_{t_1 = t_2 = 0} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^k \; y^m \; f(x, y) \; dx \; dy = E\left[X^k \; Y^m\right].$$

- For instance,

$$\begin{aligned} \mu_X &= E[X] = \frac{\partial M(0,0)}{\partial t_1} \\ \mu_Y &= E[Y] = \frac{\partial M(0,0)}{\partial t_2} \\ \sigma_X^2 &= E\left[X^2\right] - \mu_X^2 = \frac{\partial^2 M(0,0)}{\partial t_1^2} - \mu_X^2 \\ \sigma_Y^2 &= E\left[Y^2\right] - \mu_Y^2 = \frac{\partial^2 M(0,0)}{\partial t_2^2} - \mu_Y^2 \\ \mathrm{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] = \frac{\partial^2 M(0,0)}{\partial t_1 \partial t_2} - \mu_X \; \mu_Y. \end{aligned}$$

Hence, ρ may be computed from the mgf of the joint distribution.

3.6 Independent random variables

Definition 23 *Let the random variables X_1 and X_2 have the joint pdf f_{X_1,X_2} (joint pmf p_{X_1,X_2}) and the marginal pdf's (pmf's) f_{X_1} (p_{X_1}) and f_{X_2} (p_{X_2}), respectively. X_1 and X_2 are independent if and only if $f_{X_1,X_2}(x_1,x_2) = f_{X_1}(x_1) f_{X_2}(x_2)$ ($p_{X_1,X_2}(x_1,x_2) = p_{X_1}(x_1) p_{X_2}(x_2)$). Random variables that are not independent are said to be dependent.*

Theorem 33 *Let X_1 and X_2 have supports S_1 and S_2, respectively, and have the joint pdf $f(x_1,x_2)$. Then, X_1 and X_2 are independent if and only if*

$$f(x_1,x_2) = g(x_1)\, h(x_2),$$

where $g(x_1) > 0$, $x_1 \in S_1$, zero elsewhere, and $h(x_2) > 0$ for $x_2 \in S_2$, zero elsewhere.

Theorem 34 *Suppose that X_1 and X_2 are independent and that $E[u(X_1)]$ and $E[u(X_2)]$ exist. Then,*

$$E[u(X_1)\ v(X_2)] = E[u(X_1)]\, E[v(X_2)].$$

Proof. *We have*

$$\begin{aligned} E[u(X_1)\ v(X_2)] &= \int\int u(x_1)\, v(x_2)\, f_{X_1,X_2}(x_1,x_2)\, dx_1 dx_2 \\ &= \int\int u(x_1)\, v(x_2)\, f_{X_1}(x_1)\, f_{X_2}(x_2)\, dx_1 dx_2 \\ &= \left(\int\int u(x_1)\, f_{X_1}(x_1)\, dx_1\right)\left(\int v(x_2)\, f_{X_2}(x_2)\, dx_2\right) \\ &= E[u(X_1)]\, E[v(X_2)] \end{aligned}$$

■

Theorem 35 *Let (X_1,X_2) have joint cdf $F_{X_1,X_2}(x_1,x_2)$, and let X_1 and X_2 have marginal cdf's $F_{X_1}(x_1)$ and $F_{X_2}(x_2)$, respectively. Then, X_1 and X_2 are independent if and only if*

$$F_{X_1,X_2}(x_1,x_2) = F_{X_1}(x_1)\, F_{X_2}(x_2) \quad \forall\ (x_1,x_2) \in \mathbb{R}^2.$$

Proof. *($\Leftarrow$) We get*

$$f_{X_1,X_2}(x_1,x_2) = \frac{\partial^2 F_{X_1,X_2}(x_1,x_2)}{\partial x_1 \partial x_2} = \frac{\partial F_{X_1}(x_1)}{\partial x_1}\frac{\partial F_{X_2}(x_2)}{\partial x_2} = f_{X_1}(x_1)\, f_{X_2}(x_2)$$

($\Rightarrow$) Let $u(X_1) = 1(X_1 \le x_1)$ and $v(X_2) = 1(X_2 \le x_2)$ and apply the previous result. ■

Theorem 36 *X_1 and X_2 are independent random variables iff for every $a < b$, $c < d$, a,b,c,d constants,*

$$P(a < X_1 \le b, c < X_2 \le d) = P(a < X_1 \le b) P(c < X_2 \le d).$$

Proof. *($\Leftarrow$) Let $a = -\infty$ and $c = -\infty$, and we get $F_{X_1,X_2}(b,d) = F_{X_1}(b)\, F_{X_2}(d)$, so by the previous result, we get independence.*

($\Rightarrow$) Let $u(X_1) = 1(a < X_1 \le b)$ and $v(X_2) = 1(c < X_2 \le d)$ and apply the previous result.

■

Theorem 37 *Suppose that the joint mgf, $M(t_1, t_2)$ exists for X_1 and X_2. Then, X_1 and X_2 are independent if and only if*

$$M(t_1, t_2) = M(t_1, 0)\ M(0, t_2).$$

Theorem 38 (H, McC, C, p. 137) *Let $X_1, X_2, \ldots, X_m$ be independent random variables such that X_i has binomial $b(n_i, p)$ distribution, for $i = 1, 2, \ldots, m$. Let $Y = \sum_{i=1}^{m} X_i$. Then Y has a binomial $b\left(\sum_{i=1}^{m} n_i, p\right)$ distribution.*

Proof. We have

$$\begin{aligned} M_Y(t) &= E[\exp(tY)] = E\left[\exp\left(t\sum_{i=1}^{m} X_i\right)\right] = E\left[\prod_{i=1}^{m} \exp(tX_i)\right] \\ &= \prod_{i=1}^{m} E[\exp(tX_i)] = \prod_{i=1}^{m} [(1-p) + pe^t]^{n_i} = [(1-p) + pe^t]^{\sum_{i=1}^{m} n_i} \end{aligned}$$

■

Theorem 39 (H, McK, C, p. 146) *Suppose $X_1, X_2, \ldots, X_n$ are independent random variables and suppose X_i has a Poisson distribution with parameter m_i. Then $Y = \sum_{i=1}^{n} X_i$ has a Poisson distribution with parameter $\sum_{i=1}^{n} m_i$.*

Proof. We have

$$\begin{aligned} M_Y(t) &= E[\exp(tY)] = E\left[\exp\left(t\sum_{i=1}^{n} X_i\right)\right] = E\left[\prod_{i=1}^{n} \exp(tX_i)\right] \\ &= \prod_{i=1}^{n} E[\exp(tX_i)] = \prod_{i=1}^{n} e^{m_i(e^t-1)} = e^{\sum_{i=1}^{n} m_i(e^t-1)} \end{aligned}$$

■

Theorem 40 (H, McK, C, p. 154) *Let $X_1, X_2, \ldots, X_n$ be independent random variables. Suppose, for $i = 1, \ldots, n$, that X_i has a $\Gamma(\alpha_i, \beta)$ distribution. Let $Y = \sum_{i=1}^{n} X_i$. Then Y has $\Gamma\left(\sum_{i=1}^{n} \alpha_i, \beta\right)$ distribution.*

Proof. Straightforward once you recall that the MGF of $\Gamma(\alpha, \beta)$ is

$$\left(\frac{1}{1-\beta t}\right)^{\alpha}.$$

■

3.7 $t-$ and F-Distributions

Definition 24 *If $W \sim N(0,1)$, $V \sim \chi^2(r)$, and W and V are independently distributed, then*

$$T = \frac{W}{\sqrt{V/r}}$$

has a t-distribution with r degrees of freedom, and with pdf

$$f_T(x) = \frac{\Gamma\left(\frac{r+1}{2}\right)}{\sqrt{\pi r}\Gamma\left(\frac{r}{2}\right)} \frac{1}{(1+x^2/r)^{(r+1)/2}}, \quad -\infty < t < \infty. \tag{3.5}$$

Remark 11 *I know that $N(0,1)$ is defined only later in Section 4.1, but you must have taken undergraduate statistics.*

Remark 12 *It is unfortunate that the symbol $f_T(x)$ above looks like it does not depend on r, even though it clearly does. In any case, it can be shown that*

$$\lim_{r\to\infty} f_T(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right). \tag{3.6}$$

Note that the RHS is the PDF of $N(0,1)$.

Theorem 41 *Let $X_1, \ldots, X_n$ be iid random variables each having a normal distribution with mean μ and variance σ^2. Define the random variables:*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

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

Then:

1. $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$.
2. $\overline{X}$ *and* S^2 *are independent.*
3. $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$ *distribution.*
4. *The random variable*

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a Student t-distribution with $n-1$ degrees of freedom.

Definition 25 *Let U and V be two independent chi-square random variables having r_1 and r_2 degrees of freedom, respectively. Then*

$$W = \frac{U/r_1}{V/r_2}$$

has an F-distribution, with pdf

$$f_W(w) = \begin{cases} \dfrac{\Gamma\left[(r_1+r_2)/2\right]}{\Gamma(r_1/2)\,\Gamma(r_2/2)} \left(r_1/r_2\right)^{r_1/2} \dfrac{w^{(r_1/2)-1}}{\left(1+r_1 w/r_2\right)^{(r_1+r_2)/2}} & 0 < w < \infty \\ 0 & \textit{elsewhere} \end{cases}$$

3.8 Homework

1. Let $C_1, \ldots, C_4$ be sets in a sample space, S. Prove that

$$P\left(C_1 \cap C_2 \cap C_3 \cap C_4\right) = P\left(C_1\right) P\left(C_2 \middle| C_1\right) P\left(C_3 \middle| C_1 \cap C_2\right) P\left(C_4 \middle| C_1 \cap C_2 \cap C_3\right)$$

2. Let A, B, and C be three events in Γ. Suppose you know the probabilities of all intersections.

   (a) What is the probability that exactly one of the events will occur, in terms of the known probabilities.

   (b) What is the probability that only A will occur, given that B does not occur?

3. Events A and B are said to be independent conditional on even C if

$$P\left(A \cap B \middle| C\right) = P\left(A \middle| C\right) \cdot P\left(B \middle| C\right)$$

   (a) Are events A^c and B^c independent conditional on C? Explain.

   (b) Are events A and B independent conditional on C^c? Explain.

   (c) Suppose that A and B are independent conditional on C and that A and B are also independent conditional on D. Are events A and B independent conditional on $(C \cap D)$? Explain.

4. The joint pdf of random variables X and Y is given by

$$f_{Y,X}\left(y,x\right) = \begin{cases} c\, e^{-x-y} & \text{if } 0 < y < 1,\ y > x > 0 \\ 0 & \text{otherwise} \end{cases}$$

   (a) Determine the value of c.

   (b) Calculate the joint cdf $F_{Y,X}$.

(c) Calculate F_Y

(d) Calculate the conditional density of Y given X. (Use the intuitive approach.)

(e) Calculate the Moment Generating Function of (Y, X).

(f) Are Y and X independent?

(g) Calculate the expectation of the vector (Y, X).

5. Let (X, Y) be a two dimensional random vector with joint pdf

$$f_{X,Y}(x, y) = \begin{cases} e^{-y} & y > x > 0 \\ 0 & \text{otherwise} \end{cases}$$

(a) Calculate the joint cumulative distribution functions of (X, Y).

(b) Calculate the marginal cumulative distribution of X and the marginal cumulative distribution of Y.

(c) Calculate the expectation of (X, Y).

(d) Calculate $\text{Var}(X)$, $\text{Var}(Y)$, and $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$

(e) Calculate $E[Y|X]$ and $\text{Var}[Y|X]$

(f) Calculate $E[X|Y]$ and $\text{Var}[X|Y]$

(g) Let $Z = X + Y$. Calculate the pdf of Z.

6. Suppose that X is Bernoulli(p);that is $X = 1$ with probability p and $X = 0$ with probability $(1 - p)$. Suppose also that the distribution of Y conditional on $X = 0$ is $\exp(\lambda_0)$ and the distribution of Y conditional on $X = 1$ is $\exp(\lambda_1)$; that is, when $X = 0$,

$$f_{Y|X=0} = \begin{cases} \lambda_0 \, e^{-\lambda_0 \, y} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

and

$$f_{Y|X=1} = \begin{cases} \lambda_1 \, e^{-\lambda_1 \, y} & \text{if } y > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\lambda_0 > 0$ and $\lambda_1 > 0$.

(a) Calculate $E[Y|X]$ and $\text{Var}(Y|X)$ when $X = 0, 1$

(b) Calculate $\text{Cov}(Y, X) = E[(X - E[X])(Y - E[Y])]$

(c) Calculate $E[Y]$.

(d) Can X and Y be independent? Explain.

(e) What is the moment generating function of (Y, X)?

7. Suppose that $Z \sim N(0,1)$, i.e., its PDF is equal to $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) = \phi(z)$. What is $E[Z|\, Z \geq t]$? Hint: We can see that the conditional PDF of Z given $Z \geq t$ should be

$$\frac{\phi(z)\, 1(z \geq t)}{\int_t^\infty \phi(z)\, dz}$$

It follows that

$$E[Z|\, Z \geq t] = \frac{\int_t^\infty z\phi(z)\, dz}{\int_t^\infty \phi(z)\, dz}.$$

Finish the integral calculus using the fact that

$$\frac{\partial(-\phi(z))}{\partial z} = z\phi(z).$$

8. (Optional) Suppose that X and Y are independent and identically distributed random variables, with an everywhere positive density f. Consider the random vector (X, Y). Let $Z_1 = \min(X, Y)$ and $Z_2 = \max(X, Y)$.

   (a) Calculate the joint density of (Z_1, Z_2).

   (b) Calculate the conditional densities and conditional expectations of Z_i given Z_j $(i \neq j; i, j = 1, 2)$.

   (c) Calculate the marginal densities of Z_1 and of Z_2.

   (d) Derive the conditional density of Z_1 given Z_2.

9. (Optional. The hint is so thorough that the question is virtually solved for you, though.)

   (a) We define the median of a random variable X as the solution to the problem $\min_t E[|X - t|]$. Let m denote the solution. Assuming that X has a continuous distribution, prove that the FOC of this minimization problem is given by

$$0 = E[1(X < m) - 1(X \geq m)] = \Pr(X < m) - \Pr(X \geq m).$$

   (b) We define the conditional median of Y given X as the solution to the problem $\min_\varphi E[|Y - \varphi(X)|]$. Let $m(X)$ denote the solution. Consider a smaller problem $\min_t E[|Y - (m(X) + t\tilde{\varphi}(X))|]$, where $\tilde{\varphi}(X)$ is arbitrary. Prove that the FOC of this minimization problem is given by

$$0 = E[\tilde{\varphi}(X)(1(Y < m(X)) - 1(Y \geq m(X)))].$$

   under reasonable conditions.

- Hint for (a): Let $\tilde{X} = X - m$. If $t \geq 0$, we have

$$\begin{aligned}
&\left|\tilde{X} - t\right| - \left|\tilde{X}\right| \\
&= t1\left(\tilde{X} < 0\right) + \left(t - 2\tilde{X}\right) 1\left(0 \leq \tilde{X} \leq t\right) - t1\left(\tilde{X} > t\right) \\
&= t1\left(\tilde{X} < 0\right) + \left(2t - 2\tilde{X}\right) 1\left(0 \leq \tilde{X} \leq t\right) - t1\left(0 \leq \tilde{X} \leq t\right) - t1\left(\tilde{X} > t\right) \\
&= t1\left(\tilde{X} < 0\right) + 2\left(t - \tilde{X}\right) 1\left(0 \leq \tilde{X} \leq t\right) - t1\left(\tilde{X} \geq 0\right)
\end{aligned}$$

and if $t < 0$, we have

$$\begin{aligned}
\left|\tilde{X} - t\right| - \left|\tilde{X}\right| &= t1\left(\tilde{X} < t\right) + \left(2\tilde{X} - t\right) 1\left(t \leq \tilde{X} \leq 0\right) - t1\left(\tilde{X} > 0\right) \\
&= t1\left(\tilde{X} < t\right) + t1\left(t \leq \tilde{X} \leq 0\right) + \left(2\tilde{X} - 2t\right) 1\left(t \leq \tilde{X} \leq 0\right) - t1\left(\tilde{X} > 0\right) \\
&= t1\left(\tilde{X} < 0\right) + 2\left(\tilde{X} - t\right) 1\left(t \leq \tilde{X} \leq 0\right) - t1\left(\tilde{X} > 0\right)
\end{aligned}$$

To summarize, we have

$$\left|\tilde{X} - t\right| - \left|\tilde{X}\right| = t\left(1\left(\tilde{X} < 0\right) - 1\left(\tilde{X} \geq 0\right)\right) + 2\left|t - \tilde{X}\right| 1\left(\min\left(0, t\right) \leq \tilde{X} \leq \max\left(0, t\right)\right)$$

or

$$\begin{aligned}
|X - (t + m)| - |X - m| &= t\left(1\left(X < m\right) - 1\left(X \geq m\right)\right) \\
&\quad + 2\left|t + m - X\right| 1\left(\min\left(0, t\right) \leq X - m \leq \max\left(0, t\right)\right) \quad (3.7)
\end{aligned}$$

We claim that

$$E\left[\frac{\left|t - \tilde{X}\right| 1\left(\min\left(0, t\right) \leq \tilde{X} \leq \max\left(0, t\right)\right)}{t}\right] \to 0 \quad (3.8)$$

if X is continuously distributed. For simplicity, consider $t_n > 0$. Note that the sequence of functions

$$g_n\left(x\right) = \frac{\left|t_n + m - x\right| 1\left(m \leq x \leq m + t_n\right)}{t_n}$$

is nonnegative and bounded by 1. Also note that as $t_n \to 0$, we have

$$g_n\left(x\right) \to 1\left(x = m\right).$$

If X has a continuous distribution, then $1\left(x = m\right)$ is equal to zero almost surely. Therefore, by Dominated Convergence, we obtain (3.8). Therefore, the derivative of the expectation of (3.7) can be taken care of by focusing on the first term on the RHS.

- Hint for (b): Using the same derivation leading up to (3.7), we obtain

$$\begin{aligned}&|Y-(m(X)+t\tilde{\varphi}(X))|-|Y-m(X)|\\&=t\tilde{\varphi}(X)(1(Y<m(X))-1(Y\geq m(X)))\\&+2|t\tilde{\varphi}(X)+m(X)-Y|\,1(\min(0,t\tilde{\varphi}(X))\leq Y-m(X)\leq\max(0,t\tilde{\varphi}(X)))\end{aligned}$$

Letting $\tilde{Y}=Y-m(X)$, we examine

$$g_n(x,y)=\frac{(t_n\tilde{\varphi}(x)-y)\,1(\min(0,t_n\tilde{\varphi}(x))\leq y\leq\max(0,t_n\tilde{\varphi}(x)))}{t_n}$$

We consider several cases. If $\tilde{\varphi}(x)=0$, then $g_n(x,y)=0$, so $g_n(x,y)\to 0$. If $\tilde{\varphi}(x)\neq 0$, we see that $|g_n(x,y)|$ is bounded by $|\tilde{\varphi}(x)|$, and it converges to 0 except when $y=0$, in which case it converges to $|\tilde{\varphi}(x)|$. We can summarize it by writing $g_n(x,y)\to|\tilde{\varphi}(x)|\cdot 1(y=0)$. Assuming that $E[1(Y=0)|X]=0$ almost everywhere, we can conclude that

$$\frac{E[|Y-(m(X)+t\tilde{\varphi}(X))|-|Y-m(X)|]}{t}\to E[\tilde{\varphi}(X)(1(Y<m(X))-1(Y\geq m(X)))]$$

as $t\to 0$.

3.9 Technical Details - Derivation of (3.6)

Using Stirling's Formula on p.211, we have

$$\frac{\Gamma(k+1)}{\sqrt{2\pi}k^{k+1/2}\exp(-k)}\to 1.$$

This implies that

$$\begin{aligned}\lim_{r\to\infty}\frac{\Gamma\left(\frac{r}{2}+\frac{1}{2}\right)}{\sqrt{\pi r}\Gamma\left(\frac{r}{2}\right)}&=\lim_{r\to\infty}\frac{1}{\sqrt{\pi r}}\frac{\sqrt{2\pi}\left(\frac{r}{2}-\frac{1}{2}\right)^{\frac{r}{2}}\exp\left(-\frac{r}{2}+\frac{1}{2}\right)}{\sqrt{2\pi}\left(\frac{r}{2}-1\right)^{\frac{r}{2}-\frac{1}{2}}\exp\left(-\frac{r}{2}+1\right)}\\&=\frac{1}{\sqrt{\pi}}\lim_{r\to\infty}\frac{\left(\frac{r}{2}-1\right)^{\frac{1}{2}}}{\sqrt{r}}\left(\frac{\frac{r}{2}-\frac{1}{2}}{\frac{r}{2}-1}\right)^{\frac{r}{2}}\exp\left(-\frac{1}{2}\right)\end{aligned}$$

Because

$$\lim_{r\to\infty}\frac{\left(\frac{r}{2}-1\right)^{\frac{1}{2}}}{\sqrt{r}}=\lim_{r\to\infty}\left(\frac{\frac{r}{2}-1}{r}\right)^{\frac{1}{2}}=\sqrt{\lim_{r\to\infty}\left(\frac{\frac{r}{2}-1}{r}\right)}=\sqrt{\frac{1}{2}},$$

and

$$\begin{aligned}\lim_{r\to\infty}\left(\frac{\frac{r}{2}-\frac{1}{2}}{\frac{r}{2}-1}\right)^{\frac{r}{2}}&=\lim_{r\to\infty}\left(1+\frac{1}{r-2}\right)^{\frac{r-2}{2}}\left(1+\frac{1}{r-2}\right)\\&=\sqrt{\lim_{r\to\infty}\left(1+\frac{1}{r-2}\right)^{r-2}}\cdot\lim_{r\to\infty}\left(1+\frac{1}{r-2}\right)=\sqrt{e},\end{aligned}$$

we conclude that

$$\begin{aligned}\lim_{r\to\infty}\frac{\Gamma\left(\frac{r}{2}+\frac{1}{2}\right)}{\sqrt{\pi r}\Gamma\left(\frac{r}{2}\right)} &= \frac{1}{\sqrt{\pi}}\lim_{r\to\infty}\frac{\left(\frac{r}{2}-1\right)^{\frac{1}{2}}}{\sqrt{r}}\cdot\lim_{r\to\infty}\left(\frac{\frac{r}{2}-\frac{1}{2}}{\frac{r}{2}-1}\right)^{\frac{r}{2}}\cdot\exp\left(-\frac{1}{2}\right)\\ &= \frac{1}{\sqrt{2\pi}}.\end{aligned}$$

We also have

$$\begin{aligned}\lim_{r\to\infty}\left(1+x^2/r\right)^{-(r+1)/2} &= \lim_{r\to\infty}\left(1+x^2/r\right)^{-1/2}\left(\left(1+x^2/r\right)^{-r}\right)^{1/2}\\ &= \left(\lim_{r\to\infty}\left(1+x^2/r\right)\right)^{-1/2}\cdot\left(\lim_{r\to\infty}\left(1+x^2/r\right)^{-r}\right)^{1/2}\\ &= 1\cdot\sqrt{\exp\left(-x^2\right)} = \exp\left(-\frac{x^2}{2}\right).\end{aligned}$$

Chapter 4

Normal Distribution

4.1 Univariate Normal Distribution

- A random variable X has a *normal distribution* if its pdf is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad \text{for } -\infty < x < \infty.$$

- The parameters μ and σ^2 are the mean and variance of X, respectively. We will write that X has a $N(\mu, \sigma^2)$ distribution.

- If $X \sim N(\mu, \sigma^2)$, then the random variable $Z = \dfrac{X-\mu}{\sigma}$ has a $N(0,1)$ distribution, also known as *standard normal*.

- For $t \in \mathbb{R}$, the mgf of Z is given by:

$$\begin{aligned} E[\exp(tZ)] &= \int_{-\infty}^{\infty} \exp(tz) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \, dz \\ &= \exp\left(\frac{1}{2}t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(z-t)^2\right) \, dz \\ &= \exp\left(\frac{1}{2}t^2\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}w^2\right) \, dw. \end{aligned}$$

- The integrand above is the pdf of a standard normal variable, and hence the integral has value 1. Therefore:

$$M_Z(t) = \exp\left(\frac{1}{2}t^2\right)$$

- For the mgf of X, we use the relationship $X = \sigma Z + \mu$ and the mgf of Z to obtain:

$$M_X(t) = \exp\left(\mu t + \frac{1}{2}\sigma^2 t^2\right) \quad \text{for} -\infty < t < \infty$$

- Because of the relationship between normal and standard normal variables, we need only compute probabilities for standard normal random variables. Let $X \sim N(\mu, \sigma^2)$. Suppose we want to compute $F_X(x) = P(X \leq x)$ for a specified x. Then:

$$F_X(x) = P(X \leq x) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

where

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right) dw.$$

- Thus, we only need numerical integration computations for $\Phi(z)$. Note that, because the pdf of Z is symmetric about 0, we have

$$\Phi(-z) = 1 - \Phi(z).$$

Theorem 42 (H, McK, C, p. 154) *Let $X_1, X_2, \ldots, X_n$ be independent random variables. Suppose, for $i = 1, \ldots, n$, that X_i has a $\chi^2(r_i)$ distribution. Let $Y = \sum_{i=1}^n X_i$. Then Y has $\chi^2\left(\sum_{i=1}^n r_i\right)$ distribution.*

Proof. Recall that the MGF is given by

$$M(t) = (1-2t)^{-r/2}, \quad t < 1/2$$

■

Theorem 43 (H, McK, C., p. 166) *If the random variable X is $N(\mu, \sigma^2)$, $\sigma^2 > 0$, then the random variable $V = (X-\mu)^2/\sigma^2$ is $\chi^2(1)$.*

Proof. (To be skipped) Note that $W \equiv \dfrac{X-\mu}{\sigma} \sim N(0,1)$. We have

$$\begin{aligned} G(v) &= \Pr(V \leq v) = \Pr\left(W^2 \leq v\right) = \Pr\left(-\sqrt{v} \leq W \leq \sqrt{v}\right) \\ &= \int_{-\sqrt{v}}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw = 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw, \end{aligned}$$

so

$$\begin{aligned} g(v) &= \frac{dG(v)}{dv} = \frac{2}{\sqrt{2\pi}} e^{-\frac{(\sqrt{v})^2}{2}} \frac{d\sqrt{v}}{dv} \\ &= \frac{1}{\sqrt{2\pi}} v^{\frac{1}{2}-1} e^{-\frac{v}{2}}, \end{aligned}$$

which is the density of $\chi^2(1)$ ■

Theorem 44 (H, McK, C., p. 166) *Let $X_1, X_2, \ldots, X_n$ be independent random variables such that, for $i = 1, 2, \ldots, n$, X_i has $N(\mu_i, \sigma_i^2)$ distribution. Let $Y = \sum_{i=1}^n a_i X_i$, where $a_1, \ldots, a_n$ are constants. Then, $Y \sim N\left(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n a_i^2\ \sigma_i^2\right)$.*

Theorem 45 (H, McK, C., p. 167) *Let $X_1, X_2, \ldots, X_n$ be iid random variables with a common $N(\mu, \sigma^2)$ distribution. Let $\overline{X} = n^{-1}\sum_{i=1}^n X_i$. Then $\overline{X} \sim N(\mu, \sigma^2/n)$ distribution.*

4.2 Expectation and variance of a random vector

- For a random vector

$$\underset{n\times 1}{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_n \end{pmatrix},$$

we define

$$E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_n] \end{pmatrix}$$

and

$$\begin{aligned}
&\operatorname{Var}(X) \\
&\equiv E\left[(X - E[X])(X - E[X])'\right] \\
&= \begin{bmatrix} E[(X_1 - E[X_1])(X_1 - E[X_1])] & \cdots & E[(X_1 - E[X_1])(X_n - E[X_n])] \\ \vdots & \ddots & \vdots \\ E[(X_n - E[X_n])(X_1 - E[X_1])] & \cdots & E[(X_n - E[X_n])(X_n - E[X_n])] \end{bmatrix} \\
&= \begin{bmatrix} \operatorname{Var}(X_1) & \operatorname{Cov}(X_1, X_2) & \cdots & \operatorname{Cov}(X_1, X_n) \\ \operatorname{Cov}(X_2, X_1) & \operatorname{Var}(X_2) & \cdots & \operatorname{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \operatorname{Cov}(X_n, X_1) & \operatorname{Cov}(X_n, X_2) & \cdots & \operatorname{Var}(X_n) \end{bmatrix}
\end{aligned}$$

- We often write $\operatorname{Var}(X) = \operatorname{Cov}(X)$ when X is a vector
- It is useful to note that

$$E\left[(X - E[X])(X - E[X])'\right] = E[XX'] - E[X]E[X]'$$

- If A and B are nonrandom matrices, we have

$$\begin{aligned}
E[AX] &= AE[X] \\
E[XB] &= E[X]B \\
\operatorname{Var}(AX) &= A\operatorname{Var}(X)A'
\end{aligned}$$

The last equality follows from the fact

$$\begin{aligned}
\operatorname{Var}(AX) &= E\left\{(AX - E[AX])(AX - E[AX])'\right\} \\
&= E\left\{(AX - AE[X])(AX - AE[X])'\right\} \\
&= E\left\{A(X - E[X])(A(X - E[X]))'\right\} \\
&= E\left\{A(X - E[X])(X - E[X])'A'\right\} \\
&= AE\left\{(X - E[X])(X - E[X])'\right\}A' \\
&= A\operatorname{Var}(X)A'
\end{aligned}$$

- $\mathrm{Var}(X)$ is always non-negative definite, because $t' \mathrm{Var}(X) t = \mathrm{Var}(t'X) \geq 0$ for all t.

4.3 Multivariate Normal Distribution

- Consider the random vector $Z = (Z_1, \ldots . Z_n)'$,where $Z_1, \ldots Z_n$, are iid $N(0,1)$ random variables. Then the density, Z is

$$\begin{aligned} f_Z(z) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} z_i^2\right) \\ &= \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} z_i^2\right) \\ &= \left(\frac{1}{2\pi}\right)^{n/2} \exp\left(-\frac{1}{2} z'z\right) \end{aligned}$$

  for $z \in \mathbb{R}^n$.

- Because the Z_i 's have mean 0, variance 1, and are uncorrelated, the mean and covariance matrix of Z are

$$E[Z] = 0 \text{ and } \mathrm{Cov}(Z) = I_n.$$

- The mgf of Z is

$$\begin{aligned} M_Z(t) = E[\exp(t'Z)] = E\left[\prod_{i=1}^{n} \exp(t_i Z_i)\right] = \prod_{i=1}^{n} E[\exp(t_i Z_i)] \\ = \exp\left(\frac{1}{2} \sum_{i=1}^{n} t_i^2\right) = \exp\left(\frac{1}{2} t't\right) \end{aligned}$$

  for all $t \in \mathbb{R}^n$. We say that Z has a *multivariate normal distribution* with mean vector 0 and covariance matrix I_n. We write that $Z \sim N(0, I_n)$.

- Let $Z \sim N(0, I_n)$ distribution. Let Σ be a positive semi-definite, symmetric matrix and let μ be an $n \times 1$ vector of constants. Define the random vector X by

$$X = \Sigma^{1/2} Z + \mu.$$

We then have that $E[X] = \mu$ and $\text{Cov}(X) = \Sigma^{1/2}\Sigma^{1/2} = \Sigma$.[1] The mgf of X is given by:

$$\begin{aligned} M_X(t) &= E[\exp(t'X)] = E\left[\exp\left(t'\Sigma^{1/2}Z + t'\mu\right)\right] \\ &= \exp(t'\mu) E\left[\exp\left(\left(\Sigma^{1/2}t\right)' Z\right)\right] \\ &= \exp(t'\mu) \exp\left[(1/2)\left(\Sigma^{1/2}t\right)'\left(\Sigma^{1/2}t\right)\right] \\ &= \exp(t'\mu) \exp\left(\frac{t'\Sigma t}{2}\right) \end{aligned}$$

- *Definition (Multivariate Normal)*: We say an n-dimensional random vector X has a *multivariate normal distribution* if its mgf is:

$$M_X(t) = \exp\left(t'\mu + \frac{t'\Sigma t}{2}\right)$$

for all $t \in \mathbb{R}^n$ and where Σ is a symmetric, positive semi-definite matrix and $\mu \in \mathbb{R}^n$.[2] We abbreviate this by saying that $X \sim N_n(\mu, \Sigma)$ distribution.

- If Σ is positive definite, the transformation between X and Z is one-to-one with the inverse transformation:

$$Z = \Sigma^{-1/2}(X - \mu)$$

[1] The square root of a positive definite matrix is not uniquely defined. See Section 4.8 for one possible definition. Here, we understand $\Sigma^{1/2}$ to mean any matrix C such that $C'C = \Sigma$. (Obviously there is an abuse of notation when I wrote $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.)

[2] If you want to avoid the abuse of notation above, you can note that

$$X = C'Z + \mu,$$

and

$$\begin{aligned} M_X(t) &= E[\exp(t'X)] = E[\exp(t'C'Z + t'\mu)] \\ &= \exp(t'\mu) E\left[\exp\left((Ct)' Z\right)\right] \\ &= \exp(t'\mu) \exp\left[(1/2)(Ct)'(Ct)\right] \\ &= \exp(t'\mu) \exp\left(\frac{t'C'Ct}{2}\right) = \exp(t'\mu) \exp\left(\frac{t'\Sigma t}{2}\right). \end{aligned}$$

We should also understand that

$$Z = (C')^{-1}(X - \mu) = \left(C^{-1}\right)'(X - \mu)$$

and

$$\begin{aligned} \left(\left(C^{-1}\right)'(x - \mu)\right)'\left(\left(C^{-1}\right)'(x - \mu)\right) &= (x - \mu)' C^{-1}\left(C^{-1}\right)'(x - \mu) \\ &= (x - \mu)'(C'C)^{-1}(x - \mu) = (x - \mu)'\Sigma^{-1}(x - \mu). \end{aligned}$$

with Jacobian $\left|\Sigma^{-1/2}\right| = |\Sigma|^{-1/2}$**.** The pdf of X is then given by:

$$f_X(x) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-\mathbf{1}}(x-\mu)\right) \quad \text{for } x \in \mathbb{R}^n.$$

4.4 Properties of Multivariate Normal Distribution

Theorem 46 (H, McK, C., p. 173) *Suppose $X \sim N(\mu, \Sigma)$ distribution. Let $Y = AX + b$. Then $Y \sim N(A\mu + b, A\Sigma A')$.*

Proof. We have

$$\begin{aligned} M_Y(t) &= E[\exp(t'Y)] \\ &= E[\exp(t'AX + t'b)] \\ &= E\left[\exp\left((A't)'X\right)\exp(t'b)\right] \\ &= M_X(A't)\exp(t'b) \\ &= \exp\left((A't)'\mu + \frac{(A't)'\Sigma(A't)}{2}\right)\exp(t'b) \\ &= \exp\left(t'(A\mu+b) + \frac{t'(A\Sigma A')t}{2}\right) \end{aligned}$$

■

Corollary 2 (H, McK, C., p. 174) *Suppose $X \sim N(\mu, \Sigma)$, and the partitions*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \ \mu = \begin{bmatrix} \mu_{\mathbf{1}} \\ \mu_{\mathbf{2}} \end{bmatrix}, \ \Sigma = \begin{bmatrix} \Sigma_{\mathbf{11}} & \Sigma_{\mathbf{12}} \\ \Sigma_{\mathbf{21}} & \Sigma_{\mathbf{22}} \end{bmatrix}.$$

Then $X_1 \sim N(\mu_{\mathbf{1}}, \Sigma_{\mathbf{11}})$.

Theorem 47 (H, McK, C., p. 175) *Suppose $X \sim N_n(\mu, \Sigma)$ distribution, partitioned as before. Then X_1 and X_2 are independent if and only if $\Sigma_{12} = 0$.*

Theorem 48 (H, McK, C., p. 175) *Suppose $X \sim N_n(\mu, \Sigma)$ distribution, which is partitioned as before. Assume that Σ is positive definite. Then, the conditional distribution of $X_1|X_2$ is normal with mean $\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2)$ and variance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.*

Proof. For simplicity of intuition and notation, let $\mu = 0$. Let

$$\beta = \Sigma_{22}^{-1}\Sigma_{21}, \quad \varepsilon = X_1 - \beta' X_2.$$

Then

$$\begin{aligned} E[\varepsilon X_2'] &= E\left[X_1X_2' - \Sigma_{12}\Sigma_{22}^{-1}X_2X_2'\right] = E[X_1X_2'] - \Sigma_{12}\Sigma_{22}^{-1}E[X_2X_2'] \\ &= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0 \end{aligned}$$

so ε and X_2 are independent by Theorem 47. Therefore, we have

$$X_1 = \beta' X_2 + \varepsilon$$

and

$$\begin{aligned} E\left[\varepsilon\varepsilon'\right] &= E\left[\left(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2\right)\left(X_1 - \Sigma_{12}\Sigma_{22}^{-1}X_2\right)'\right] \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \end{aligned}$$

In general, you want to work with

$$\begin{aligned} \beta &= \Sigma_{22}^{-1}\Sigma_{21}, \\ \alpha &= \mu_1 - \beta'\mu_2, \\ \varepsilon &= X_1 - \alpha - \beta' X_2 = (X_1 - \mu_1) - \beta'\left(X_2 - \mu_2\right). \end{aligned}$$

■

Theorem 49 (H, McK, C., p. 177) *Suppose $X \sim N(\mu, \Sigma)$ where Σ is $n \times n$ and positive definite. Then the random variable $W = (X-\mu)'\Sigma^{-1}(X-\mu)$ has a $\chi^2(n)$ distribution.*

4.5 A Few Useful Applications

4.5.1 Why Are $\overline{X}$ and S^2 independent?

This is about Theorem 41. Let ℓ denote an $n \times 1$ matrix consisting of ones. Let

$$P \equiv \frac{1}{n}\ell\ell', \quad M \equiv I_n - P$$

and note that

$$\begin{aligned} MP = PM &= 0, \\ P\ell &= \ell, \\ M\ell &= 0, \\ P' &= P, \\ M' &= M, \\ PP &= P, \\ MM &= M \end{aligned}$$

Now, write

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \\ \vdots \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & & 0 \\ 0 & \sigma^2 & & 0 \\ & & \ddots & \\ 0 & 0 & & \sigma^2 \end{bmatrix}\right) = N\left(\mu\ell, \sigma^2 I_n\right)$$

and note that

$$Y \equiv \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \equiv \begin{bmatrix} \overline{X} \\ X_1 - \overline{X} \\ X_2 - \overline{X} \\ \vdots \\ X_n - \overline{X} \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\ell' X \\ X - PX \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\ell' X \\ MX \end{bmatrix} = \begin{bmatrix} \frac{1}{n}\ell' \\ M \end{bmatrix} X$$

so the variance covariance matrix of Y is

$$\begin{bmatrix} \frac{1}{n}\ell' \\ M \end{bmatrix} \left(\sigma^2 I_n\right) \begin{bmatrix} \frac{1}{n}\ell & M' \end{bmatrix} = \sigma^2 \begin{bmatrix} \frac{1}{n^2}\ell'\ell & \frac{1}{n}\ell' M \\ \frac{1}{n}M\ell & MM' \end{bmatrix} = \sigma^2 \begin{bmatrix} \frac{1}{n} & 0 \\ 0 & M \end{bmatrix}$$

In particular, we can see that

$$E\left[\left(Y_1 - E\left[Y_1\right]\right)\left(Y_2 - E\left[Y_2\right]\right)'\right] = 0$$

so Y_1 and Y_2 are independent by Theorem 47. In other words, $\overline{X}$ and $X - \overline{X}\ell$ are independent, and as such $\overline{X}$ and

$$S^2 = \frac{1}{n-1}\left(X - \overline{X}\ell\right)'\left(X - \overline{X}\ell\right)$$

should be independent.

4.5.2 Distribution of $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

This is about Theorem 41 again. We would like to investigate the distribution of

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \left(X - \overline{X}\ell\right)'\left(X - \overline{X}\ell\right) = Y_2' Y_2$$

We know from the preceding analysis that the variance of Y_2 is $\sigma^2 M$. We can easily see that

$$E\left[Y_2\right] = E\left[MX\right] = M\left(\mu\ell\right) = 0.$$

For simplicity, we will start by considering the case with $\sigma^2 = 1$.

We will first prove that if an $n \times n$ matrix A is such that $A' = A$ and $AA = A$, and if $Z \sim N\left(0, I_n\right)$, then, $Z'AZ \sim \chi^2\left(\text{trace}\left(A\right)\right)$. For this purpose, we will use the eigenvalue decomposition in Section 4.8, and write $A = \Gamma\Lambda\Gamma'$ with $\Gamma\Gamma' = I_n$ and Λ is diagonal. Because $AA = A$, it is trivial to see that the diagonal elements of Λ are either 1 or 0. Without loss of generality, we will write that m diagonal elements of Λ are 1, and they are the first m diagonal elements. Now, we see that

$$\widetilde{Z} \equiv \Gamma' Z \sim N\left(0, \Gamma' I_n \Gamma\right) = N\left(0, I_n\right),$$

and therefore,

$$Z'AZ = \left(\Gamma' Z\right)'\Lambda\left(\Gamma' Z\right) = \widetilde{Z}'\Lambda\widetilde{Z} = \sum_{i=1}^{m}\widetilde{Z}_i^2 \sim \chi^2\left(m\right)$$

We then note that

$$\text{trace}\,(A) = \text{trace}\,(\Gamma\Lambda\Gamma') = \text{trace}\,(\Lambda\Gamma'\Gamma) = \text{trace}\,(\Lambda) = m,$$

which yields the conclusion.

Getting back to $Y_2'Y_2$ with $\underline{\sigma^2 = 1}$, we can write $Y_2'Y_2 = Z'MZ$ so

$$Y_2'Y_2 = Z'MZ \sim \chi^2\,(\text{trace}\,(M)) = \chi^2\,(n-1)\,.$$

4.6 Noncentral χ^2 Distribution

- Let's consider a slightly simplified version of the discussion on pp.475-476 of the textbook. Let $X_1, \ldots, X_n$ be independent $N\,(\mu_i, 1)$ variables. We consider the distribution of $Y = \sum_{i=1}^n X_i^2$. (Obviously it is $\chi^2\,(n)$ if μs are all zeros.)

- We will let $X = (X_1, X_2, \ldots, X_n)'$, and note $X \sim N\,(\mu, I_n)$ for $\mu = (\mu_1, \mu_2, \ldots, \mu_n)'$. We then note that $Y = X'X$.

- It turns out that the distribution of Y depends only on two "parameters" n and $\theta^2 \equiv \mu'\mu$, and we call the distribution noncentral χ^2 distribution. We use the symbol $\chi^2\,(n, \theta^2)$. See Section 4.9.

- The distribution "increases" as a function of the noncentrality parameter.[3] For intuition, consider $n = 1$, and $Z \sim N\,(0, 1)$. We see that for $\theta, t > 0$,

$$\begin{aligned}\Pr\left(\chi^2\left(1, \theta^2\right) \geq t^2\right) &= \Pr\left((Z+\theta)^2 \geq t^2\right) \\ &= \Pr\,(Z + \theta \geq t) + \Pr\,(Z + \theta \leq -t) \\ &= 1 - \Phi\,(t - \theta) + \Phi\,(-t - \theta)\end{aligned}$$

  and hence,

$$\frac{\partial \Pr\,(\chi^2\,(1, \theta^2) \geq t^2)}{\partial \theta} = \phi\,(t-\theta) - \phi\,(-t-\theta) = \phi\,(|t-\theta|) - \phi\,(t+\theta)\,.$$

  Because $\theta, t > 0$, we have $|t - \theta| \leq t + \theta$, and as a consequence, we have $\phi\,(|t-\theta|) - \phi\,(t+\theta) \geq 0$. Also note that

$$\Pr\left(\chi^2\left(1, \theta^2\right) \geq t^2\right) \to 1$$

  as $\theta \to \infty$.

[3] You may want to Google "first order stochastic dominance".

4.7 Homework

1. Suppose that $Z \sim N(0, I_n)$. Prove that $Z'Z \sim \chi^2(n)$ using Theorems 42 and 43.

2. Suppose that an $n \times n$ matrix Σ is positive definite, and $X \sim N(\mu, \Sigma)$. Prove that $(X-\mu)' \Sigma^{-1} (X-\mu) \sim \chi^2(n)$. Hint: Recall that $\Sigma^{-1/2}(X-\mu) \sim N(0, I_n)$.

4.8 Technical Details - Eigenvalue and Eigenvector

- If $\Sigma_{n\times n}$ is postive-definite and symmetric, then there exist n vectors $x_1, x_2, \ldots x_n$ and n scalars $\lambda_1, \lambda_2, \ldots \lambda_n$ such that

$$\Sigma x_j = \lambda_j x_j$$

and

$$x_j' x_j = 1, \quad x_j' x_i = 0 \qquad \forall i \neq j$$

- Let

$$\Gamma = [x_1, x_2, \ldots x_n],$$

and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \ldots & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & \lambda_n \end{bmatrix}$$

then

$$\Sigma\Gamma = \Gamma\Lambda$$

and

$$\Sigma = \Gamma\Lambda\Gamma'$$

- If we define

$$\Sigma^{1/2} = \Gamma\Lambda^{1/2}\Gamma'$$

since $\Gamma\Gamma' = I$, then $\Sigma^{1/2} \cdot \Sigma^{1/2} = \Sigma$.

- Since $\left|\Sigma^{1/2}\right| \cdot \left|\Sigma^{-1/2}\right| = 1$, $\left|\Sigma^{-1/2}\right| = |\Sigma^{1/2}|^{-1}$

4.9 Technical Details - Orthogonal Matrix

- Let $X \sim N(\mu, I_n)$. We argue that the distribution of $Y = X'X$ depends only on two "parameters" n and $\theta \equiv \mu'\mu$, and we call the distribution noncentral χ^2 distribution. We use the symbol $\chi^2(n, \theta)$.

- Note that $Y = (HX)'(HX)$, where H is an arbitrary orthgonal matrix satisfying $H'H = I_k$.

  - An intuition about H can be gained by thinking of 2-dimensional space. Any point u on the unit circle in $\mathbb{R}^2$ can be represented by $(\cos A, \sin A)$ for some A. Also, any such matrix H should satisfy

$$H = \begin{bmatrix} \cos B & -\sin B \\ \sin B & \cos B \end{bmatrix}$$

    for some B. Therefore,

$$\begin{aligned} Hu &= \begin{bmatrix} \cos B & -\sin B \\ \sin B & \cos B \end{bmatrix} \begin{bmatrix} \cos A \\ \sin A \end{bmatrix} = \begin{bmatrix} \cos A \cos B - \sin A \sin B \\ \cos A \sin B + \cos B \sin A \end{bmatrix} \\ &= \begin{bmatrix} \cos (A+B) \\ \sin (A+B) \end{bmatrix} \end{aligned}$$

    and H can be argued to "rotate" u by the angle B.

- Because $HX \sim N(H\mu, HI_nH') = N(H\mu, I_n)$, we can see that the distribution of Y should depend on μ only through $H\mu$ for H arbitrary.

- Now, here's a useful linear algebraic fact. Suppose that x is a vector. Then there exists an orthogonal matrix such that $Hx = |x| e_1$, where $|x| \equiv \sqrt{x'x}$ and e_1 is the unit vector where every component except the first one is zero.

  - Here's one way of proving it. It can be shown that there exists an orthogonal matrix Q_x such that the first row of Q_x is equal to $x/|x|$. (If you are familiar with Gram-Schmidt orthogonalization, you know how to construct such a matrix.) For such Q_x, we have $Q_xQ_x' = I_k$, so $Q_xx = |x| e_1$. Now, let $H = Q_x$ so that $Hx = |x| e_1$.

- Getting back to $Y = (HX)'(HX)$, we can find an orthogonal matrix H such that $H\mu = \sqrt{\mu'\mu}e_1$. This suggests that the distribution of Y depends only on two "parameters" n and $\theta \equiv \mu'\mu$, and we call the distribution noncentral χ^2 distribution. We use the symbol $\chi^2(n, \theta)$.

Chapter 5

Unbiasedness, Consistency, and Limiting Distributions

5.1 Definitions: Random Sample, Statistic, and Point Estimator

Definition 26 *The random variables $X_1, X_2, \ldots X_n$ are called a random sample if $X_1, X_2, \ldots X_n$ are mutually independent random variables, and have the same PDF. We will abbreviate this by saying that $X_1, X_2, \ldots X_n$ are independent and identically distributed, i.e., i.i.d.*

Definition 27 *A statistic T is a function of the sample: $T = T(X_1, X_2, \ldots X_n)$.*

- *Point estimator*: When we use T to estimate θ, we say that T is a *point estimator* of θ. Particularly important point statistics are the *sample mean* $\overline{X}$ and the *sample variance* S^2 of a random sample.

Definition 28 (Unbiased estimator) *Let X be a random variable with pdf $f(x;\theta)$ (pmf $p(x;\theta)$), with $\theta \in \Omega$. Let $X_1, X_2, \ldots X_n$ be a random sample from the distribution of X and let T denote a statistic. T is an unbiased estimator of θ if*

$$E[T] = \theta \quad \text{for all } \theta \in \Omega.$$

If T is not unbiased, i.e., $E[T] \neq \theta$, we say that T is a biased estimator of θ.

Example 5 *Let $X_1, X_2, \ldots X_n$ be a random sample from the distribution of a random variable X with mean μ and variance σ^2. Then, the sample mean*

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an unbiased estimator of μ, and the sample variance

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

is an unbiased estimator of σ^2.

Proof.

$$E\left[\overline{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

$$\text{Var}\left(\overline{X}\right) = \frac{\sigma^2}{n}$$

$$E\left[\overline{X}^2\right] = \left(E\left[\overline{X}\right]\right)^2 + \text{Var}\left(\overline{X}\right) = \mu^2 + \frac{\sigma^2}{n}$$

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

and

$$E\left[S^2\right] = \frac{1}{n-1}\left(\sum_{i=1}^{n} E\left[X_i^2\right] - nE\left[\overline{X}^2\right]\right)$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\left(\mu^2 + \sigma^2\right) - n\left(\mu^2 + \frac{\sigma^2}{n}\right)\right)$$
$$= \sigma^2$$

■

5.2 Convergence in Probability

Definition 29 *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable defined on a sample space. X_n converges in probability to X if for all $\epsilon > 0$*

$$\lim_{n\to\infty} P\left[|X_n - X| \geq \epsilon\right] = 0,$$

or equivalently,

$$\lim_{n\to\infty} P\left[|X_n - X| < \epsilon\right] = 1.$$

If so, we write

$$X_n \xrightarrow{P} X.$$

Lemma 1 *Let $\{Y_n\}$ be a sequence of random variables that converges to Y in mean square, i.e., $E\left[(Y_n - Y)^2\right] \to 0$. Then $Y_n \xrightarrow{P} Y$.*

Proof. Pick any $\epsilon > 0$, and the desired result follows immediately from Markov inequality

$$P\left[|Y_n - Y| \geq \epsilon\right] = P\left[(Y_n - Y)^2 \geq \epsilon^2\right] \leq \frac{E\left[(Y_n - Y)^2\right]}{\epsilon^2}.$$

■

Theorem 50 (Weak Law of Large Numbers) *Let $\{X_n\}$ be a sequence of independent random variables having a common mean μ and variance $\sigma^2 < \infty$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$, then*

$$\overline{X}_n \xrightarrow{P} \mu.$$

Proof. Note that

$$E\left[\left(\overline{X}_n - \mu\right)^2\right] = \text{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n} \to 0$$

and apply the previous lemma. ■

Remark 13 *WLLN is an implication the following two observations: (1) "Let $\{X_n\}$ be a sequence of independent random variables having a common mean μ. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$, then $\overline{X}_n \to \mu$ a.s. (Strong Law of Large Numbers)"; and (2) "If $X_n \to X$ a.s., then $X_n \to X$ in probability".*

Theorem 51 *Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_n + Y_n \xrightarrow{P} X + Y$.*

Proof. Pick $\epsilon > 0$, and note that $|(X_n + Y_n) - (X + Y)| \leq |X_n - X| + |Y_n - Y|$. Then the desired result follows from

$$\begin{aligned} P\left[|(X_n + Y_n) - (X + Y)| \geq \epsilon\right] &\leq P\left[|X_n - X| + |Y_n - Y| \geq \epsilon\right] \\ &\leq P\left[|X_n - X| \geq \epsilon/2\right] + P\left[|Y_n - Y| \geq \epsilon/2\right]. \end{aligned}$$

■

Theorem 52 *Suppose $X_n \xrightarrow{P} X$ and a is a constant. Then $aX_n \xrightarrow{P} aX$.*

Proof. When $a = 0$ the result is clear. When $a \neq 0$, the desired result follows almost immediately from definition. ■

Theorem 53 *Suppose $X_n \xrightarrow{P} a$ and the real function g is continuous at a. Then $g(X_n) \xrightarrow{P} g(a)$.*

Proof. Let $\epsilon > 0$. Then, since g is continuous at a, there exists $\delta > 0$ such that $|x - a| < \delta$ implies $|g(x) - g(a)| < \epsilon$, or equivalently, $|g(x) - g(a)| \geq \epsilon \Rightarrow |x - a| \geq \delta$. And the result follows immediately after setting $x = X_n$. ■

Theorem 54 *Suppose $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$. Then $X_n Y_n \xrightarrow{P} XY$.*

Definition 30 *Consistent estimator: Let X be a random variable with cdf $F(x, \theta)$, $\theta \in \Omega$. Let $X_1, X_2, \ldots X_n$ be a sample from the distribution of X and let T_n denote a statistic. T_n is a consistent estimator of θ if*

$$T_n \xrightarrow{P} \theta.$$

5.3 Convergence in Distribution

Definition 31 *Convergence in Distribution: Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. Let F_{X_n} and F_X be the cdfs of X_n and X, respectively. Let $C(F_X)$ denote the set of all points where F_X is continuous. X_n converges in distribution to X if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ for all } x \in C(F_X).$$

We denote this convergence by $X_n \xrightarrow{D} X$.

Theorem 55 *If X_n converges to X in probability, then X_n converges to X in distribution.*

Proof. *We have*

$$P[X_n \le t] \le P[X \le t+\varepsilon] + P[|X_n - X| > \varepsilon],$$
$$P[X \le t-\varepsilon] \le P[X_n \le t] + P[|X_n - X| > \varepsilon],$$

so

$$\limsup P[X_n \le t] \le P[X \le t+\varepsilon] + \lim_{n\to\infty} P[|X_n - X| > \varepsilon] = P[X \le t+\varepsilon],$$
$$P[X \le t-\varepsilon] \le \liminf P[X_n \le t] + \lim_{n\to\infty} P[|X_n - X| > \varepsilon] = \liminf P[X_n \le t],$$

from which we obtain

$$P[X \le t-\varepsilon] \le \liminf P[X_n \le t] \le \limsup P[X_n \le t] \le P[X \le t+\varepsilon]$$

Now assume that the CDF of X is continuous at t, and let $\varepsilon \to 0$. We would then obtain

$$P[X \le t] \le \liminf P[X_n \le t] \le \limsup P[X_n \le t] \le P[X \le t]$$

from which we conclude that

$$\lim P[X_n \le t] = P[X \le t].$$

■

Theorem 56 *If X_n converges to the constant b in distribution, then X_n converges to b in probability.*

Proof. Since X_n converges to b in distribution, then $\lim_n F_{X_n}(y) = 0$ when $y < b$, whereas $\lim_n F_{X_n}(y) = 1$ when $y \ge b$. Pick now any $\epsilon > 0$, and note that

$$\begin{aligned} P[|X_n - b| \ge \epsilon] &\le P[X_n \le b-\epsilon] + P[X_n > b+\epsilon/2] \\ &= F_{X_n}(b-\epsilon) + [1 - F_{X_n}(b+\epsilon/2)]. \end{aligned}$$

Finally, by taking the limit as $n \to \infty$ we obtain the desired result. ■

Theorem 57 *Suppose X_n converges to X in distribution and Y_n converges in probability to 0. Then $X_n + Y_n$ converges to X in distribution.*

Theorem 58 (Continuous Mapping Theorem) *Suppose X_n converges to X in distribution and g is a continuous function on the support of X. Then $g(X_n)$ converges to $g(X)$ in distribution.*

Theorem 59 *Slutsky's Theorem : Let X_n, X, A_n, B_n be random variables and let a and b be constants. If $X_n \xrightarrow{D} X$, $A_n \xrightarrow{P} a$, and $B_n \xrightarrow{P} b$, then*

$$A_n + B_n X_n \xrightarrow{D} a + bX.$$

5.4 Bounded in Probability

Definition 32 *The sequence of random variables $\{X_n\}$ is bounded in probability if for all $\epsilon > 0$ there exists a constant $B_\epsilon > 0$ and an integer N_ϵ such that*

$$n \geq N_\epsilon \Rightarrow P\left[|X_n| \leq B_\epsilon\right] \geq 1 - \epsilon.$$

Theorem 60 *Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. If $X_n \to X$ in distribution, then $\{X_n\}$ is bounded in probability.*

Theorem 61 *Let $\{X_n\}$ be a sequence of random variables bounded in probability and let $\{Y_n\}$ be a sequence of random variables which converge to 0 in probability. Then $X_n Y_n \xrightarrow{P} 0$.*

Definition 33 *$X_n = o_p(Y_n)$ iff $\frac{X_n}{Y_n} \xrightarrow{p} 0$ as $n \to \infty$*

Definition 34 *$X_n = O_p(Y_n)$ iff $\frac{X_n}{Y_n}$ is bounded in probability as $n \to \infty$*

Theorem 62 *Suppose $\{Y_n\}$ is a sequence of random variables which is bounded in probability. Suppose $X_n = o_p(Y_n)$. Then $X_n \xrightarrow{P} 0$ as $n \to \infty$.*

5.5 $O(1)$, $o(1)$, $O_p(1)$, $o_p(1)$

Definition 35 *If a nonstochastic sequence x_n of numbers is bounded, i.e., if there exists some $B < \infty$ such that $|x_n| \leq B$, we write $x_n = O(1)$*

Definition 36 *If a nonstochastic sequence x_n of numbers converges to zero, we write $x_n = o(1)$*

- Note that

$$\begin{aligned} O(1) + O(1) &= O(1) \\ O(1) \cdot O(1) &= O(1) \\ O(1) \cdot o(1) &= o(1) \\ o(1) + o(1) &= o(1) \\ o(1) \cdot o(1) &= o(1) \end{aligned}$$

  We also have

$$g(a + o(1)) = g(a) + o(1)$$

  if g is continuous at a.

- We similarly have

$$\begin{aligned} O_p(1) + O_p(1) &= O_p(1) \\ O_p(1) \cdot O_p(1) &= O_p(1) \\ O_p(1) \cdot o_p(1) &= o_p(1) \\ o_p(1) + o_p(1) &= o_p(1) \\ o_p(1) \cdot o_p(1) &= o_p(1) \end{aligned}$$

- Some results in previous sections may be rewritten

$$\begin{aligned} (X + o_p(1)) + (Y + o_p(1)) &= X + Y + o_p(1) \\ (X + o_p(1)) \cdot (Y + o_p(1)) &= X \cdot Y + o_p(1) \\ a(X + o_p(1)) &= aX + o_p(1) \end{aligned}$$

  and

$$g(a + o_p(1)) = g(a) + o_p(1)$$

  if g is continuous at a. If $X_n \xrightarrow{D} X$, we have

$$\begin{aligned} &X_n + o_p(1) \xrightarrow{D} X \\ &(a + o_p(1)) + (b + o_p(1)) X_n \xrightarrow{D} a + bX \end{aligned}$$

5.6 Delta Method

Theorem 63 *Let $\{X_n\}$ be a sequence of random variables such that $\sqrt{n}(X_n - \theta) \xrightarrow{D} N(0, \sigma^2)$. Suppose the function $g(x)$ is continuously differentiable at θ. Then,*

$$\sqrt{n}[g(X_n) - g(\theta)] \xrightarrow{D} N\left(0, \sigma^2 [g'(\theta)]^2\right).$$

Proof. (Sketch) since $g(X_n) = g(\theta) + g'(\theta)(X_n - \theta) + o_p(|X_n - \theta|)$, then we have

$$\sqrt{n}[g(X_n) - g(\theta)] = g'(\theta)\sqrt{n}(X_n - \theta) + o_p(\sqrt{n}|X_n - \theta|) \xrightarrow{D} g'(\theta)N\left(0, \sigma^2\right) + 0$$

■

Theorem 64 *Let $\{\mathbf{X}_n\}$ be a sequence of $k \times 1$ random vectors such that $\sqrt{n}(\mathbf{X}_n - \theta) \xrightarrow{D} N(0, \Sigma)$, where Σ is a $k \times k$ positive semi-definite matrix. Suppose the function $h : \mathbb{R}^k \to \mathbb{R}^s$ is differentiable at θ and $\nabla h(\theta)_{s\times k}$ denotes the matrix of partial derivatives. Then,*

$$\sqrt{n}[h(\mathbf{X}_n) - h(\theta)] \xrightarrow{D} N(0, \nabla h(\theta)\ \Sigma\ \nabla h(\theta)').$$

5.7 Central Limit Theorem

Theorem 65 *Let $\{X_n\}$ be a sequence of random variables with mgf $M_{X_n}(t)$ that exists for $-h < t < h$ for all n. Let X be a random variable with mgf $M(t)$, which exists for $|t| \leq h_1 \leq h$. If $\lim_{n\to\infty} M_{X_n}(t) = M(t)$ for $|t| \leq h_1$, then $X_n \xrightarrow{D} X$.*

Theorem 66 (Central Limit Theorem) *Let $X_1, X_2, \ldots X_n$ denote the observations of an i.i.d. sample from a distribution that has mean μ and (positive) variance σ^2. Then,*

$$\sqrt{n}\left(\overline{X}_n - \mu\right) \xrightarrow{D} N\left(0, \sigma^2\right).$$

Remark 14 *Hence, when n is a large, fixed positive integer, $\overline{X}$ has an approximate normal distribution with mean μ and variance σ^2/n.*

Remark 15 *Note that $\sqrt{n}\left(\overline{X}_n - \mu\right) = O_p(1)$, so $\overline{X}_n = \mu + O_p\left(n^{-1/2}\right)$*

Proof. (Sketch) Because

$$\sqrt{n}\left(\overline{X}_n - \mu\right) = \sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sqrt{n}}\right),$$

the mgf of the LHS is equal to

$$M\left(\frac{t}{\sqrt{n}}\right)^n,$$

where M denotes the mgf of $X_i - \mu$. We note that $M^{(1)}(0) = 0$ and $M^{(2)}(0) = \sigma^2$, so we may write

$$\begin{aligned} M\left(\frac{t}{\sqrt{n}}\right) &= M(0) + M^{(1)}(0)\frac{t}{\sqrt{n}} + \frac{M^{(2)}(\xi)}{2}\left(\frac{t}{\sqrt{n}}\right)^2 \\ &= 1 + \frac{\sigma^2 t^2}{2n} + \frac{M^{(2)}(\xi) - \sigma^2}{2}\frac{t^2}{n} \end{aligned}$$

where ξ is between 0 and $\frac{t}{\sqrt{n}}$. Therefore, the mgf of $\sqrt{n}\left(\overline{X}_n - \mu\right)$ is

$$\left(1 + \frac{\sigma^2 t^2}{2n} + \frac{M^{(2)}(\xi) - \sigma^2}{2}\frac{t^2}{n}\right)^n \to \exp\left(\frac{\sigma^2 t^2}{2}\right)$$

because

$$\lim_{n\to\infty}\left(1 + \frac{b}{n} + \frac{\phi(n)}{n}\right)^{cn} = e^{bc} \quad \text{if} \quad \lim_{n\to\infty}\phi(n) = 0$$

■

5.8 Homework

1. Suppose that $\{X_i\}_{i=1,2,\ldots}$ is a sequence of independent random variables such that $E[X_i] = \mu_i$ and $\mathrm{Var}(X_i) = \sigma_i^2$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ and $\overline{\mu}_n = n^{-1}\sum_{i=1}^n \mu_i$. Show that if $\sup_{i=1,2,\ldots}\sigma_i^2 < \infty$ then

$$\overline{X}_n - \overline{\mu}_n \xrightarrow{p} 0$$

Hint: Use

$$E\left[\overline{X}_n\right] = \overline{\mu}_n,$$
$$\mathrm{Var}\left(\overline{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^n \sigma_i^2,$$
$$\sum_{i=1}^n \sigma_i^2 \le n \sup_i \sigma_i^2,$$

along with Chebyshev.

2. Let the random variable Y_n have the distribution $b(n,p)$.

   (a) Prove that Y_n/n converges in probability to p. Hint: Calculate the mean and variance of Y_n/n.

   (b) Prove that $1 - Y_n/n$ converges in probability to $1-p$.

   (c) Prove that $(Y_n/n)(1 - Y_n/n)$ converges in probability to $p(1-p)$.

3. Let $X_1, X_2, X_3, \ldots$ denote an i.i.d. sample from the p.d.f $f_X(\cdot;\lambda)$ where $\lambda > 0$ and

$$\begin{aligned} f_X(x) &= \lambda\, e^{-\lambda x} && \text{if } x > 0 \\ &= 0 && \text{otherwise} \end{aligned}$$

Define $\widehat{\lambda}_n = \left(\overline{X}_n\right)^{-1}$. Show that $\widehat{\lambda}_n$ is a consistent estimator for λ and derive the limiting distribution of $\sqrt{n}\left(\widehat{\lambda}_n - \lambda\right)$.

4. Let $X_1, X_2, \ldots, X_n$ be a sequence of i.i.d. random vectors in $\mathbb{R}^q$ from a distribution with mean μ and variance Σ. By the Central Limit Theorem, $Y_n = \sqrt{n}\left(\overline{X}_n - \mu\right)$ converges in distribution to a multivariate normal random vector with mean μ and variance Σ. Let $\widehat{\Sigma}$ be a consistent estimator of Σ. What is the limiting distribution of $W_n \equiv n\left(\overline{X}_n - \mu\right)' \widehat{\Sigma}^{-1}\left(\overline{X}_n - \mu\right)$? Explain. Hint: We have $\sqrt{n}\left(\overline{X}_n - \mu\right) = O_p(1)$ and $\widehat{\Sigma}^{-1} = \Sigma^{-1} + o_p(1)$. Therefore,

$$\begin{aligned}
&n\left(\overline{X}_n - \mu\right)' \widehat{\Sigma}^{-1}\left(\overline{X}_n - \mu\right) \\
&= \left(\sqrt{n}\left(\overline{X}_n - \mu\right)\right)' \Sigma^{-1} \sqrt{n}\left(\overline{X}_n - \mu\right) + \sqrt{n}\left(\overline{X}_n - \mu\right)' \left(\widehat{\Sigma}^{-1} - \Sigma^{-1}\right) \sqrt{n}\left(\overline{X}_n - \mu\right) \\
&= Y_n' \Sigma^{-1} Y_n + O_p(1)\, o_p(1)\, O_p(1) \\
&= Y_n' \Sigma^{-1} Y_n
\end{aligned}$$

Now, let $g(y) = y'\Sigma^{-1}y$ and apply the Continuous Mapping Theorem. Finally use Theorem 49.

5. (Optional Question) (page 218 in textbook) Let Y_n denote the maximum of a random sample of size n from a distribution of the continuous type that has cdf $F(x)$ and pdf $f(x) = F'(x)$. Find the limiting distribution of $Z_n = n[1 - F(Y_n)]$. (Note: $F(Y_n)$ is the random variable defined as the value of the function F when $x = Y_n$.) Hint: Let $X_1, \ldots, X_n$ denote the underlying random sample, i.e., let $Y_n = \max(X_1, \ldots, X_n)$. Note that $n[1 - F(Y_n)] \geq z$ if and only if $1 - F(Y_n) \geq \frac{z}{n}$ if and only if $F(Y_n) \leq 1 - \frac{z}{n}$ if and only if $Y_n \leq F^{-1}\left(1 - \frac{z}{n}\right)$ if and only if $X_k \leq F^{-1}\left(1 - \frac{z}{n}\right)$ for all $k = 1, \ldots, n$. It follows that

$$\begin{aligned}
\Pr(Z_n \geq z) &= \Pr\left[X_k \leq F^{-1}\left(1 - \frac{z}{n}\right) \text{ for all } k = 1, \ldots, n\right] \\
&= \prod_{k=1}^{n} \Pr\left[X_k \leq F^{-1}\left(1 - \frac{z}{n}\right)\right]
\end{aligned}$$

6. (Optional Question) (page 218 in textbook) Let the pmf of Y_n be $p_n(y) = 1$ if $y = n$ and $p_n(y) = 0$ otherwise. Show that Y_n does not have a limiting distribution. Hint: If Y_n had a limiting distribution, we should have $Y_n = O_p(1)$.

7. (Optional Question) (page 219 in textbook) Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution that is $N(\mu, \sigma^2)$. Show that the sum $Z_n = \sum_{i=1}^{n} X_i$ does not have a limiting distribution. Hint: If Z_n had a limiting distribution, we should have $Y_n = O_p(1)$.

8. (Very Optional Question. This question may look long, but you will see that it is solved for you for all practical purpose. While "solving" this question, you will be forced to use many seemingly unrelated results. This is not the standard method to prove the asymptotic distribution of the sample median. Also, don't be too scared because you will

NOT be asked to reproduce this line of argument in the final or comp exam.) Derive the asymptotic distribution of the sample median.

(a) Suppose that X has a continuous and strictly increasing CDF F. Suppose that U is a uniform$(0,1)$ random variable. Let $Y = F^{-1}(U)$. Prove that Y has the same CDF as X.

(b) Let $U_1, \ldots, U_n$ be i.i.d. uniform$(0,1)$. Let $U_{(k)}$ denote the k-th smallest element of the set $\{U_1, \ldots, U_n\}$. Prove that the CDF of $U_{(k)}$ is equal to

$$F_{(k)}(x) = \sum_{j=k}^{n} \binom{n}{j} x^j (1-x)^{n-j}$$

Hint: $U_{(k)} \leq x$ if and only if k or more of the U_i's are $\leq x$.

(c) Prove that the PDF of $U_{(k)}$ is equal to

$$f_{(k)}(x) = \frac{n!}{(k-1)!\,(n-k)!} x^{k-1} (1-x)^{n-k}$$

Compare with the PDF of $beta(\alpha, \beta)$, and conclude that $U_{(k)}$ has $beta(k, n+1-k)$ distribution. For the PDF of $beta(\alpha, \beta)$, look up Question #3 in the homework immediately following Section 1.15. Use the result of the same question ("If X_1 and X_2 are independent and have $\Gamma(\alpha, 1)$ and $\Gamma(\beta, 1)$ distributions, then $X_1/(X_1 + X_2)$ has the beta(α, β) distribution.") to conclude that $U_{(k)}$ has the same distribution as $Y_1/(Y_1 + Y_2)$, where $Y_1 \sim \Gamma(k, 1)$ and $Y_2 \sim \Gamma(n+1-k, 1)$.

(d) Prove that $U_{(k)}$ has the same distribution as

$$\frac{\sum_{i=1}^{k} V_i}{\sum_{i=1}^{k} V_i + \sum_{i=k+1}^{n+1} V_i},$$

where V_i are i.i.d. with common density equal to $\exp(-x)$ for $x > 0$, and 0 elsewhere. Hint: Use Theorem 40, and conlcude that Y_1 has the same distribution as $\sum_{i=1}^{k} V_i$ and Y_2 has the same distribution as $\sum_{i=k+1}^{n+1} V_i$. Convince yourself that $\sum_{i=1}^{k} V_i$ should be independent of $\sum_{i=k+1}^{n+1} V_i$.

(e) Using the above result and the central limit theorem, derive the asymptotic distribution of

$$\sqrt{n}\left(U_{(\lfloor \frac{n}{2} \rfloor)} - \frac{1}{2}\right).$$

Here, the $\lfloor \frac{n}{2} \rfloor$ denotes the smallest integer larger than or equal to $\frac{n}{2}$. You probably want to start by analyzing the mean and variance of V_i in the previous question. You may ignore the possibility that $\frac{n}{2}$ may not be an integer in your answer, i.e.,

you may assume that $\left\lfloor \frac{n}{2} \right\rfloor = \frac{n}{2}$. You may do so by letting $n = 2m$ for some integer m. Hint: Let $n = 2m$, and work with

$$\begin{bmatrix} A_m \\ B_m \end{bmatrix} = \begin{bmatrix} \frac{1}{m} \sum_{i=1}^{m} V_i \\ \frac{1}{m} \sum_{i=m+1}^{2m+1} V_i \end{bmatrix}.$$

Recal properties of V_i from Section 1.9.1, i.e., $E[V_i] = \text{Var}(V_i) = 1$, and use the CLT to conclude that

$$\sqrt{m} \left(\begin{bmatrix} A_m \\ B_m \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \right) = \begin{bmatrix} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} (V_i - 1) \\ \frac{1}{\sqrt{m}} \sum_{i=m+1}^{2m+1} (V_i - 1) \end{bmatrix} \to N(0, I_2)$$

Use the Delta Method to

$$\frac{\sum_{i=1}^{k} V_i}{\sum_{i=1}^{k} V_i + \sum_{i=k+1}^{n+1} V_i} = \frac{A_m}{A_m + B_m}$$

and conclude that

$$\sqrt{m} \left(\frac{A_m}{A_m + B_m} - \frac{1}{1+1} \right) \to N(0, \Delta' I_2 \Delta)$$

for

$$\Delta = \begin{bmatrix} \frac{1}{4} \\ -\frac{1}{4} \end{bmatrix}$$

Conclude that

$$\sqrt{m} \left(U_{(m)} - \frac{1}{2} \right) \to N\left(0, \frac{1}{8}\right)$$

Conclude that

$$\sqrt{n} \left(U_{\left(\frac{n}{2}\right)} - \frac{1}{2} \right) = \sqrt{2m} \left(U_{(m)} - \frac{1}{2} \right) \to N\left(0, \frac{1}{4}\right).$$

(f) Suppose that $X_1, \ldots, X_n$ be i.i.d. with a continuous and strictly increasing CDF F. Let $X_{(k)}$ denote the k-th smallest element of the set $\{X_1, \ldots, X_n\}$. Prove that $X_{(k)}$ has the same distribution as $F^{-1}\left(U_{(k)}\right)$, where $U_{(k)}$ is the k-th smallest element of the i.i.d. uniform$(0,1)$ $U_1, \ldots, U_n$. Hint: Use Theorem 9, and recognize that we can write $X_k = F^{-1}(U_k)$. Note that F^{-1} is a monotonically increasing function, so the orders are preserved.

(g) Derive the asymptotic distribution of $\sqrt{n}\left(X_{\left(\lfloor \frac{n}{2} \rfloor\right)} - F^{-1}\left(\frac{1}{2}\right)\right)$. Hint: Using the delta method, derive the asymptotic distribution of $\sqrt{n}\left(F^{-1}\left(U_{\left(\lfloor \frac{n}{2} \rfloor\right)}\right) - F^{-1}\left(\frac{1}{2}\right)\right)$. We may want to note that the derivative of F^{-1} at $\frac{1}{2}$ is equal to

$$\frac{1}{f\left(F^{-1}\left(\frac{1}{2}\right)\right)}.$$

Chapter 6

Hypothesis Testing

6.1 Introduction to Hypothesis Testing

- $X \sim f(x;\theta)$ with $\theta \in \Omega$. (When we deal with a random sample $X_1, \ldots, X_n$, we just understand $X = (X_1, \ldots, X_n)'$.)
- $\theta \in \omega_0$ or $\theta \in \omega_1$, where $\omega_0 \cup \omega_1 = \Omega$
- $H_0 : \theta \in \omega_0$ vs. $H_1 : \theta \in \omega_1$
- Let $\mathcal{D}$ denote the sample space of the random sample X
- A test is based on a subset $\mathcal{C}$ of $\mathcal{D}$ such that we reject H_0 iff $X \in \mathcal{C}$
- Such a set is called the critical region
- We usually distinguish Type I error and Type II error:

	True State of Nature	
Decision	H_0 is true	H_1 is true
Reject H_0	Type I Error	Correct Decision
Accept H_0	Correct Decision	Type II Error

- We often consider Type I error to be the worse of the two errors, and try to select critical regions which bound the probability of Type I error and minimize the probability of Type II error
- We say a critical region C is of size α if

$$\alpha = \sup_{\theta \in \omega_0} P_\theta [X \in C].$$

Frequently α is also called the significance level of the test.

- We say that the power of the test at θ is

$$1 - P_\theta\left[\text{Type II error}\right] = P_\theta\left[X \in C\right].$$

  Therefore, minimization of the probability of Type II error is equivalent to maximization of power.

- We define the power function of a critical region to be

$$\gamma_C(\theta) = P_\theta\left[X \in C\right]; \quad \theta \in \omega_1$$

- We say that a test C is unbiased if

$$\sup_{\theta \in \omega_0} P_\theta\left[X \in C\right] \leq \inf_{\theta \in \omega_1} P_\theta\left[X \in C\right]$$

Example 6 *Suppose that $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$, where σ^2 is known to be equal to 1. We are given $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$. The test you learned as an undergraduate takes the form: Reject H_0 if*

$$\frac{\overline{X}}{1/\sqrt{n}} = \sqrt{n}\overline{X} > c$$

for some c. If $\alpha = 5\%$, we need to choose c such that

$$P_\mu\left[\sqrt{n}\overline{X} > c\right] = 5\%$$

when $\mu = 0$. If $\mu = 0$, we know that $\sqrt{n}\overline{X} \sim N(0,1)$, so using the usual Z notation, we can write

$$P_0\left[\sqrt{n}\overline{X} > c\right] = P\left[Z > c\right]$$

Therefore, if we choose $c = 1.645$, we can guarantee that $P_\mu\left[\sqrt{n}\overline{X} > c\right] = 5\%$. What is the power of this test? For some arbitrary $\mu > 0$, we see that the power is $P_\mu\left[\sqrt{n}\overline{X} > 1.645\right]$. But $\overline{X} \sim N\left(\mu, \frac{1}{n}\right)$ so $\sqrt{n}\overline{X} \sim N\left(\sqrt{n}\mu, 1\right)$. With some abuse of notation, we can therefore write $\sqrt{n}\overline{X} = \sqrt{n}\mu + Z$, and hence,

$$\begin{aligned} P_\mu\left[\sqrt{n}\overline{X} > 1.645\right] &= P\left[\sqrt{n}\mu + Z > 1.645\right] \\ &= P\left[Z > 1.645 - \sqrt{n}\mu\right] \\ &= 1 - \Phi\left(1.645 - \sqrt{n}\mu\right) \\ &= \Phi\left(\sqrt{n}\mu - 1.645\right) \end{aligned}$$

6.2 Most Powerful Tests

- We begin with testing a simple hypothesis against a simple alternative: $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$.

- For simplicity of notations, we will understand by $f(x;\theta)$ the joint density of $X \equiv (X_1, \ldots, X_n)$
- We would like to find the best critical region by solving the constrained maximization problem

$$\max_C P_{\theta_1}(X \in C) \text{ s.t. } P_{\theta_0}(X \in C) = \alpha$$

In effect, the best critical region maximizes the power of the test while keeping the significance level of the test equal to α.

Theorem 67 (Neyman-Pearson Theorem) *Let X denote a random vector with p.d.f. $f(x;\theta) = L(\theta;x)$. Assume that $H_0 : \theta = \theta_0$ and $H_1 : \theta = \theta_1$. Let*

$$C \equiv \left\{ x : \frac{L(\theta_1;x)}{L(\theta_0;x)} \geq k \right\}, \tag{6.1}$$

where k is chosen in such a way that

$$P[X \in C; H_0] = \alpha.$$

Then, C is the best critical region of size α for testing H_0 against the alternative H_1.

Proof. Assume that A is another critical region of size α. We want to show that

$$\int_C f(x;\theta_1)\,dx \geq \int_A f(x;\theta_1)\,dx.$$

By using the familiar indicator function notation, we can rewrite the inequality as

$$\int (1(x \in C) - 1(x \in A)) \cdot f(x;\theta_1)\,dx \geq 0. \tag{6.2}$$

For this purpose, it suffices to show that

$$(1(x \in C) - 1(x \in A)) \cdot f(x;\theta_1) \geq k \cdot (1(x \in C) - 1(x \in A)) \cdot f(x;\theta_0). \tag{6.3}$$

(This is because of the following reason. If it (6.3) holds, it follows that

$$\int (1(x \in C) - 1(x \in A)) \cdot f(x;\theta_1)\,dx \geq k \cdot \int (1(x \in C) - 1(x \in A)) \cdot f(x;\theta_0)\,dx.$$

But since

$$\int (1(x \in C) - 1(x \in A)) \cdot f(x;\theta_0)\,dx = \int_C f(x;\theta_0)\,dx - \int_A f(x;\theta_0)\,dx = \alpha - \alpha = 0,$$

we obtain (6.2).)

Notice that $1(x \in C) - 1(x \in A)$ equals 1, 0, or -1. We show below that (6.3) holds for all three cases:

(1) When $1(x \in C) - 1(x \in A) = 0$, (6.3) holds with both sides of the inequality equal to 0.
(2) When $1(x \in C) - 1(x \in A) = 1$, we have $1(x \in C) = 1$ and $1(x \in A) = 0$, so (6.3) is equivalent to $f(x;\theta_1) \geq kf(x;\theta_0)$. But because $1(x \in C) = 1$, we have $f(x;\theta_1) \geq kf(x;\theta_0)$ by (6.1). Therefore, (6.3) holds.
(3) When $1(x \in C) - 1(x \in A) = -1$, we have $1(x \in C) = 0$ and $1(x \in A) = 1$, so (6.3) is equivalent to $f(x;\theta_1) \leq kf(x;\theta_0)$. But because $1(x \in C) = 0$, we have $f(x;\theta_1) < kf(x;\theta_0)$ by (6.1). Therefore, (6.3) holds. ■

Example 7 *Consider* $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$. *We assume that* σ^2 *is known. We have* $H_0 : \mu = \mu_0$ *vs.* $H_1 : \mu = \mu_1$ *with* $\mu_0 < \mu_1$. *Now,*

$$\frac{L(\mu_1; x_1, \ldots, x_n)}{L(\mu_0; x_1, \ldots, x_n)} = \frac{\left(1/\sqrt{2\pi\sigma^2}\right)^n \exp[-\left(\sum_i (x_i - \mu_1)^2\right)/2\sigma^2]}{\left(1/\sqrt{2\pi\sigma^2}\right)^n \exp[-\left(\sum_i (x_i - \mu_0)^2\right)/2\sigma^2]}$$
$$= \exp\left[\left(\sum_i x_i\right)(\mu_1 - \mu_0)/\sigma^2 - n\left(\mu_0^2 - \mu_1^2\right)/2\sigma^2\right].$$

The best critical region C *takes the form*

$$\exp\left[\left(\sum_i x_i\right)(\mu_1 - \mu_0)/\sigma^2 - n\left(\mu_0^2 - \mu_1^2\right)/2\sigma^2\right] \geq k$$

for some k. *In other words,* C *takes the form*

$$\overline{x} \geq c$$

for some c. *We can find the value of* c *easily from the standard normal distribution table.*

6.3 Uniformly Most Powerful Tests

Definition 37 *The critical region* C *is a uniformly most powerful critical region of size* α *for testing a simple* H_0 *against a composite* H_1 *if the set* C *is a best critical region of size* α *for testing* H_0 *against each simple hypothesis in* H_1. *A Test defined by this critical region is called a uniformly most powerful test with significance level* α.

Example 8 *Assume that* $X_1, \ldots, X_n$ *are* i.i.d. $N(0, \theta)$ *random variables. We want to test* $H_0 : \theta = \theta'$ *against* $H_1 : \theta > \theta'$. *We first consider a simple alternative* $H_1 : \theta = \theta''$ *where* $\theta'' > \theta'$. *The best critical region takes the form*

$$k \leq \frac{\left(1/\sqrt{2\pi\theta''}\right)^n \exp\left[-\sum_i x_i^2/2\theta''\right]}{\left(1/\sqrt{2\pi\theta'}\right)^n \exp\left[-\sum_i x_i^2/2\theta'\right]} = \left(\frac{\theta'}{\theta''}\right)^{n/2} \exp\left[\frac{\theta'' - \theta'}{2\theta''\theta'} \sum_i x_i^2\right]$$

In other words, the best critical region takes the form $\sum_i x_i^2 \geq c$ *for some* c, *which is determined by the size of the test. Notice that the same argument holds for any* $\theta'' > \theta'$. *It thus follows that* $\sum_i x_i^2 \geq c$ *is the uniformly most powerful test of* H_0 *against* H_1*!*

Example 9 *Let $X_1, \ldots, X_n$* i.i.d. *$N(\theta, 1)$. There exists no uniformly most powerful test of $H_0 : \theta = \theta'$ against $H_1 : \theta \neq \theta'$. Consider $\theta'' \neq \theta'$. The best critical region for testing $\theta = \theta'$ against $\theta = \theta''$ takes the form*

$$\frac{\left(1/\sqrt{2\pi}\right)^n \exp\left[-\sum_i \left(x_i - \theta''\right)^2 /2\right]}{\left(1/\sqrt{2\pi}\right)^n \exp\left[-\sum_i \left(x_i - \theta'\right)^2 /2\right]} \geq k$$

or

$$\exp\left[\left(\theta'' - \theta'\right)\sum_i x_i - n\left(\left(\theta''\right)^2 - \left(\theta'\right)^2\right)/2\right] \geq k$$

Thus, when $\theta'' > \theta'$, the best critical region takes the form

$$\sum_i x_i \geq c,$$

and when $\theta'' < \theta'$, it takes the form

$$\sum_i x_i \leq c.$$

It thus follows that there exists no uniformly most powerful test.

6.4 Beyond UMP

6.4.1 Two-Sided Test

- We can try to 'save' the usual two-sided test from the problem discussed in Example 9.
- Suppose that $X \sim N(0, \sigma^2)$ and we want to test $H_0 : \mu = 0$ vs $H_1 : \mu = \pm\beta$.
- The alternative is no longer simple, so we consider giving equal 'weights' to $\pm\beta$.
- The Neyman-Pearson test would then look at

$$\begin{aligned}
&\frac{\frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\beta)^2}{2\sigma^2}\right) + \frac{1}{2}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x+\beta)^2}{2\sigma^2}\right)}{\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{x^2}{2\sigma^2}\right)} \\
&= \frac{1}{2}\exp\left(-\frac{-2\beta x + \beta^2}{2\sigma^2}\right) + \frac{1}{2}\exp\left(-\frac{2\beta x + \beta^2}{2\sigma^2}\right) \\
&= \frac{1}{2}\exp\left(-\frac{\beta^2}{2\sigma^2}\right)\left[\exp\left(\frac{\beta x}{\sigma^2}\right) + \exp\left(-\frac{\beta x}{\sigma^2}\right)\right] \\
&= \frac{1}{2}\exp\left(-\frac{\beta^2}{2\sigma^2}\right)\left[\exp\left(\left|\frac{\beta}{\sigma^2}\right| x\right) + \exp\left(-\left|\frac{\beta}{\sigma^2}\right| x\right)\right] \\
&= \exp\left(-\frac{\beta^2}{2\sigma^2}\right)\cosh\left(\left|\frac{\beta}{\sigma^2}\right| x\right), \qquad (6.4)
\end{aligned}$$

where we recall that

$$\cosh t \equiv \frac{1}{2}\exp(t) + \frac{1}{2}\exp(-t).$$

We also recall that (i) $\cosh t$ is an even function in t, and (ii) it is monotonically increasing in $|t|$.

- Therefore, the likelihood ratio is monotonically increasing in $|x|$, so the Neyman-Pearson test rejects for large values of $|X|$, from which the UMP test properties follow.[1]

6.4.2 Multivariate Generalization

- This topic will not be covered in class.
- Suppose that $X \sim N(\mu, \sigma^2 I_k)$ and we want to test $H_0 : \mu = 0$ vs $H_1 : |\mu| = \beta > 0$.
- In order to go through the similar analysis, we would like to attach equal weights to the sphere in $\mathbb{R}^k$ with radius equal to β.
  - If $k = 2$, we can imagine a 'uniform' distribution on the circle $(\beta\cos\theta, \beta\sin\theta)$ by thinking that θ has a uniform distribution over $0 < \theta < 2\pi$.
  - In general[2], we can imagine a uniform random vector U on a sphere with radius β which would have the property that the distribution of U is identical to the distribution of HU, where H is an arbitrary orthgonal matrix satisfying $H'H = I_k$.
  - It can be shown that if U has such a uniform distribution, then the distribution of $x'U$ is identical to that of $|x|\,U_1$. This can be done by choosing an orthogonal matrix H satisfying $Hx = |x|\,e_1$, and noting that $x'U = x'H'HU = (Hx)'HU = |x|\,e_1'HU$ should have the same distribution as $|x|\,e_1'U = |x|\,U_1$.
- Now, we would like to go through the same analysis as (6.4). We note that the counterpart of the ratio on the far LHS should be

$$\frac{\int \frac{1}{\left(\sigma\sqrt{2\pi}\right)^k}\exp\left(-\frac{(x-\mu)'(x-\mu)}{2\sigma^2}\right)\Gamma(d\mu)}{\frac{1}{\left(\sigma\sqrt{2\pi}\right)^k}\exp\left(-\frac{x'x}{2\sigma^2}\right)}$$

where Γ denotes the uniform distribution over the sphere with radius β.

[1] Note that the test only depends on $|X|$. Also note that the power of the test

$$P[|X| > c] = P[|\sigma Z + \mu| > c]$$

at $\mu = \beta$, i.e., $P[|\sigma Z + \beta| > c]$ is identical to the power $P[|\sigma Z - \beta| > c]$ at $\mu = -\beta$. It is because $|\sigma Z + \beta|$ has the same distribution as $|\sigma(-Z) + \beta| = |\sigma Z - \beta|$, due to the symmetry of Z.

[2] This kind of random variable for $\beta = 1$, e.g., can be generated from your computer by letting $\mu \sim Z/|Z|$ where $Z \sim N(0, I_k)$, and $|Z| = \sqrt{Z'Z}$.

- It is shown in Lemma 2 that this ratio is equal to

$$\exp\left(-\frac{\beta^2}{2\sigma^2}\right)\int_0^{\beta}\cosh\left(\frac{|x|}{\sigma^2}u\right)2g\left(u\right)du$$

for some positive valued function $g\left(u\right)$. Now, recall again that cosh is monotonically increasing if its argument is positive.

- We should therefore conclude that is monotonically increasing in $|x|$, so the Neyman-Pearson test rejects for large values of $|X|$, from which the UMP test properties follow.[3]

- We can generalize the analysis to the case where the variance-covariance matrix takes a general form Σ. Not surprisingly, the test rejects for large values of $X'\Sigma^{-1}X$. For notational simplicity, the proof is not given.

6.4.3 Weighted Average Power

- In case the alternative hypothesis is no longer simple, i.e., when ω_1 consists of more than one element, perhaps we can be a little less obsessed about UMP. Perhaps we can try to maximize the weighted average power:

$$\max_C \int_{\omega_1} P_\theta\left(X\in C\right)w\left(\theta\right)d\theta \text{ s.t. } P_{\theta_0}\left(X\in C\right)=\alpha$$

for some weight $w\left(\theta\right)$ over ω_1. Without loss of generality, we will assume that $\int w\left(\theta\right)d\theta = 1$. Because

$$P_\theta\left(X\in C\right)=\int 1\left(x\in C\right)f\left(x;\theta\right)dx$$

we can write the objective alternatively as

$$\begin{aligned}\int_{\omega_1} P_\theta\left(X\in C\right)w\left(\theta\right)d\theta &= \int_{\omega_1}\left(\int 1\left(x\in C\right)f\left(x;\theta\right)dx\right)w\left(\theta\right)d\theta\\ &= \int 1\left(x\in C\right)\left(\int_{\omega_1} f\left(x;\theta\right)w\left(\theta\right)d\theta\right)dx\end{aligned}$$

where $\int_{\omega_1} f\left(x;\theta\right)w\left(\theta\right)d\theta$ is the weighted average of models under H_1.

- This suggests that we can apply the Neyman-Pearson yet again.

[3] Note that the test only depends on $|X|$. Also note that the power of the test

$$P\left[|X|>c\right]=P\left[|\sigma Z+\mu|>c\right]$$

is constant on the sphere with radius β. Let μ^* be such that $|\mu^*|=\beta$. We can find an orthogonal H such that $\mu=H\mu^*$. Note that $|\sigma Z+\mu|$ has the same distribution as $|\sigma H'Z+\mu|=|\sigma H'Z+H'H\mu|=|H'\left(\sigma Z+H\mu\right)|=|H'\left(\sigma Z+\mu^*\right)|=|\sigma Z+\mu^*|$. As a consequence, $P\left[|\sigma Z+\mu|>c\right]$ is constant on the sphere with radius β.

Example 10 *Suppose that* $X \sim N(\mu, \Sigma)$*, and that* $H_0 : \mu = 0$ *vs* $H_1 : \mu \neq 0$*. We can put a 'weight' on* μ *proportional to the density of* $N(0, V)$*. Then the Neyman-Pearson strategy would examine the ratio*

$$\frac{\int \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-\mathbf{1}}(x-\mu)\right) \frac{1}{(2\pi)^{k/2}|V|^{1/2}} \exp\left(-\frac{1}{2}\mu' V^{-\mathbf{1}}\mu\right) d\mu}{\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}x'\Sigma^{-\mathbf{1}}x\right)} \tag{6.5}$$

The numerator looks a little too complicated, so let's see if we can simplify it a bit by imposing the weight with $V = \Sigma$*. Lemma 3 establishes that the Neyman-Pearson test boils down to rejecting* $H_0 : \mu = 0$ *for large values of* $X'\Sigma^{-\mathbf{1}}X$*.*

6.4.4 Comments

- In this section, we worked with the one-sample scenario. This is just for notational simplicity, but you may be worried that we are losing substance. We are not.
- Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(0, \sigma^2)$ and we want to test $H_0 : \mu = 0$ vs $H_1 : \mu = \pm\beta$.
- The alternative is no longer simple, so we consider giving equal 'weights' to $\pm\beta$.
- The Neyman-Pearson test would then look at

$$\begin{aligned}
&\frac{\frac{1}{2}\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i-\beta)^2}{2\sigma^2}\right) + \frac{1}{2}\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x_i+\beta)^2}{2\sigma^2}\right)}{\prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2\sigma^2}\right)} \\
&= \frac{\frac{1}{2}\prod_{i=1}^n \exp\left(-\frac{(x_i-\beta)^2}{2\sigma^2}\right) + \frac{1}{2}\prod_{i=1}^n \exp\left(-\frac{(x_i+\beta)^2}{2\sigma^2}\right)}{\prod_{i=1}^n \exp\left(-\frac{x_i^2}{2\sigma^2}\right)} \\
&= \frac{\frac{1}{2}\exp\left(-\frac{\sum_{i=1}^n x_i^2 - 2\beta\sum_{i=1}^n x_i + n\beta^2}{2\sigma^2}\right) + \frac{1}{2}\exp\left(-\frac{\sum_{i=1}^n x_i^2 + 2\beta\sum_{i=1}^n x_i + n\beta^2}{2\sigma^2}\right)}{\exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\right)}
\end{aligned}$$

  so with $x = \frac{1}{n}\sum_{i=1}^n x_i$, the above expression can be written as

$$\begin{aligned}
&\frac{1}{2}\exp\left(-\frac{-2\beta x + \beta^2}{2\sigma^2/n}\right) + \frac{1}{2}\exp\left(-\frac{2\beta x + \beta^2}{2\sigma^2/n}\right) \\
&= \frac{1}{2}\exp\left(-\frac{\beta^2}{2\sigma^2/n}\right)\left[\exp\left(\frac{\beta x}{\sigma^2/n}\right) + \exp\left(-\frac{\beta x}{\sigma^2/n\sigma^2}\right)\right] \\
&= \frac{1}{2}\exp\left(-\frac{\beta^2}{2\sigma^2/n}\right)\left[\exp\left(\left|\frac{\beta}{\sigma^2/n}\right| x\right) + \exp\left(-\left|\frac{\beta}{\sigma^2/n}\right| x\right)\right] \\
&= \exp\left(-\frac{\beta^2}{2\sigma^2/n}\right)\cosh\left(\left|\frac{\beta}{\sigma^2/n}\right| x\right).
\end{aligned}$$

- So the situation becomes mathematical identical to the one-sample case, where $X \sim N(0, \sigma^2/n)$, and you did not miss anything substantive.
- The same comment applies to the multivariate cases.

6.5 Some Asymptotics

- If we consider different sample sizes, we should acknowledge that the test/critical region C should really be indexed by n. The test should be written C_n, and the probability of rejection should be written $P_\theta[(X_1, \ldots, X_n) \in C_n]$. For simplicity of notation, we will omit the subscript n.
- We say that a test (really a sequence of tests) is asymptotically of level α if

$$\lim_{n\to\infty} \sup_{\theta\in\omega_0} P_\theta[(X_1, \ldots, X_n) \in C] \leq \alpha \tag{6.6}$$

- We say that a test C with power function $\gamma_C(\theta)$ is consistent at level α against the alternative θ if it is asymptotically of level α and $\gamma_C(\theta) \to 1$. If it is consistent against every alternative, we simply say that the test is consistent.

Example 11 *We continue Example 6, except that the null is now $H_0 : \mu \leq 0$. As before, we reject H_0 if $\sqrt{n}\overline{X} > 1.645$. Recall that*

$$P_\mu\left[\sqrt{n}\overline{X} > 1.645\right] = \Phi\left(\sqrt{n}\mu - 1.645\right)$$

which is monotonically increasing in μ. Therefore,

$$\sup_{\mu\leq 0} P_\mu\left[\sqrt{n}\overline{X} > 1.645\right] = \Phi(0 - 1.645) = 5\%$$

from which we obtain

$$\lim_{n\to\infty} \sup_{\mu\leq 0} P_\mu\left[\sqrt{n}\overline{X} > 1.645\right] = \lim_{n\to\infty} 5\% = 5\%$$

Therefore, the test is asymptotically of level 5%. As for the power of the test, we still have

$$P_\mu\left[\sqrt{n}\overline{X} > 1.645\right] = \Phi\left(\sqrt{n}\mu - 1.645\right)$$

For each fixed $\mu > 0$, we see that $\sqrt{n}\mu \to \infty$ as $n \to \infty$, so for every μ satisfying the alternative, we have

$$\lim_{n\to\infty} P_\mu\left[\sqrt{n}\overline{X} > 1.645\right] = \lim_{n\to\infty} \Phi\left(\sqrt{n}\mu - 1.645\right) = 1$$

i.e., the test is consistent.

Remark 16 *It is tempting to use*

$$\sup_{\theta \in \omega_0} \lim_{n \to \infty} P_\theta \left[(X_1, \ldots, X_n) \in C \right] \tag{6.7}$$

as the asymptotic size, but it may not be the best idea. Here's the reason, based on the previous example with the twist $H_0 : \mu < 0$*. We see that*

$$\sup_{\mu<0} P_\mu \left[\sqrt{n}\overline{X} > 1.645 \right] = \sup_{\mu<0} \Phi \left(\sqrt{n}\mu - 1.645 \right) = 5\%$$

for each fixed n*, so it seems like a good idea to define the asymptotic size of the test to be 5%. On the other hand, we can see for each* $\mu < 0$*, we should have*

$$\lim_{n \to \infty} P_\mu \left[\sqrt{n}\overline{X} > 1.645 \right] = \lim_{n \to \infty} \Phi \left(\sqrt{n}\mu - 1.645 \right) = 0$$

so we would end up calling $\sup_{\mu<0} \lim_{n\to\infty} P_\mu \left[\sqrt{n}\overline{X} > 1.645 \right] = 0$ *the asymptotic size of the test if the problematic definition (6.7) were to be used.*

6.6 Efficiency Aspect - Warm Up

- Here, we learn why moving alternatives are a useful mathematical device.
- Most interesting tests are consistent anyway, and as such, consistency is not such a useful concept in practice. One way to assess the performance of a test is to examine its power in problems that becomes harder as more observations become available.
- It may be useful to have some concrete example. Let $X_1, \ldots, X_n$ *i.i.d.* $N(\theta, 1)$. Suppose that we test $H_0 : \theta = 0$ against $H_1 : \theta \neq 0$. Suppose that we use the usual test of rejecting the null when $\left| \sqrt{n}\overline{X} \right| \geq 1.96$.
  - It is trivial to calculate the size of the test, and confirm that it is 5% regardless of the sample size.
  - For $\theta \neq 0$, we note that $\overline{X} \sim N\left(\theta, \frac{1}{n}\right)$, so

    $$\sqrt{n}\overline{X} \sim N\left(\sqrt{n}\theta, 1\right)$$

    and the power of the test is given by

    $$\Pr\left(\left| \sqrt{n}\overline{X} \right| \geq 1.96 \right).$$

    Using the discussion in Section 4.6, we can see that (with some abuse of notations),

    $$\Pr\left(\left| \sqrt{n}\overline{X} \right| \geq 1.96 \right) = \Pr\left(\left(\sqrt{n}\overline{X} \right)^2 \geq 1.96^2 \right) = \Pr\left(\chi^2\left(1, n\theta^2\right) \geq 1.96^2 \right). \tag{6.8}$$

- Because the noncentrality parameter $n\theta^2 \to \infty$, the probability of rejection converges to 1, i.e., the test is consistent.
- Consider the test based on the sample average of the first half of the sample, discarding the second half. For simplicity, assume that $n = 2m$, and let $\overline{X}_{(m)}$ denote the sample average based on the first m observations. The same analysis as above will go through, and we can see that the power of the test must be equal to

$$\Pr\left(\chi^2\left(1, m\theta^2\right) \geq 1.96^2\right) = \Pr\left(\chi^2\left(1, \frac{n\theta^2}{2}\right) \geq 1.96^2\right)$$

  Because for fixed n the noncentrality parameter in this silly test is exactly half of the noncentrality parameter in (6.8), we can see that this test has less power than the common sense based test. Even then, this silly test is still consistent, and from the consistency point of view, there is no reason to prefer the common sense test to this silly test.
- If we take $\theta = \frac{h}{\sqrt{n}}$, then the power of the original test is now

$$\Pr\left(\chi^2\left(1, h^2\right) \geq 1.96^2\right)$$

  which does not converge to 1 as $n \to \infty$. (This example is special because the probability of rejecting the null at $\frac{h}{\sqrt{n}}$ does not change as $n \to \infty$.)
- On the other hand, the power of the silly test is

$$\Pr\left(\chi^2\left(1, \frac{h^2}{2}\right) \geq 1.96^2\right)$$

  which should be dominated by $\Pr\left(\chi^2\left(1, h^2\right) \geq 1.96^2\right)$. In other words, the device $\theta = \frac{h}{\sqrt{n}}$ makes it possible to appreciate the loss of power even in asymptotic analysis.

6.7 Efficiency Aspect - Bottom Line

- Here's a more abstract example. Suppose that we are given $f(x;\theta)$ with scalar θ. We would like to test $H_0 : \theta = 0$ against $H_1 : \theta > 0$. Instead of evaluating the power of the test at a fixed θ, we can evaluate it when $\theta = \frac{h}{\sqrt{n}}$.
- Often, the test takes the form of rejecting for large values of a test statistic T_n. Assume that for all the sequences of the form $\theta_n = \frac{h}{\sqrt{n}}$

$$\frac{\sqrt{n}\left(T_n - \mu\left(\theta_n\right)\right)}{\sigma\left(\theta_n\right)} \to N(0,1) \text{ under } \theta_n$$

  This would imply that $\frac{\sqrt{n}(T_n - \mu(0))}{\sigma(0)} \to N(0,1)$ under $\theta = 0$, so the test would reject if $\sqrt{n}\left(T_n - \mu(0)\right)$ exceeds $\sigma(0) z_\alpha$.

- We will assume (as is often the case) that under $\theta = 0$,

$$\sqrt{n}\left(T_n - \mu\left(0\right)\right) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p\left(1\right)$$

where $\psi_i = \psi\left(X_i\right)$ has a zero expectation. By the CLT, we should have $\sqrt{n}\left(T_n - \mu\left(0\right)\right) \rightarrow N\left(0, E\left[\psi_i^2\right]\right)$ under $\theta = 0$.

- We want to choose T_n such that the power at $\theta_n = \frac{h}{\sqrt{n}}$ is maximized.

- Bottom line is that it can be done by choosing T_n such that

$$\psi_i \propto s_i, \tag{6.9}$$

where

$$s_i \equiv \left.\frac{\partial \log f\left(X_i;\theta\right)}{\partial \theta}\right|_{\theta=0}$$

is the "score".

- Another bottom line requires definition of MLE:

  - Suppose that $X_1, \ldots, X_n$ are iid with PDF $f\left(x;\theta_0\right)$.
  - We estimate θ_0 by

$$\arg\max \prod_{i=1}^{n} f\left(X_i;\theta\right) = \arg\max \sum_{i=1}^{n} \log f\left(X_i;\theta\right)$$

  - We then have

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) = I\left(\theta_0\right)^{-1} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i\right) + o_p\left(1\right) \tag{6.10}$$

    where

$$I\left(\theta\right) \equiv -E\left[\frac{\partial^2 \log f\left(X;\theta\right)}{\partial \theta^2}\right].$$

  - Digression: It can be shown that

$$I\left(\theta\right) = E\left[\left(\frac{\partial \log f\left(X;\theta\right)}{\partial \theta}\right)^2\right].$$

    Combined with (6.10), we can see that

$$\sqrt{n}\left(\widehat{\theta} - \theta_0\right) \rightarrow N\left(0, I^{-1}\left(\theta_0\right)\right).$$

6.8 Maximum Likelihood Tests

- Suppose that we want to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. There are three tests associated with the MLE $\widehat{\theta}$.
- For simplicity, we assume that θ is a scalar.
- Likelihood ratio test rejects H_0 when

$$\Lambda = \frac{L(\theta_0)}{L\left(\widehat{\theta}\right)}$$

is small, where $L(\theta) = \prod_{i=1}^{n} f(X_i; \theta)$. More precisely, the null is rejected if $-2\log\Lambda \geq \chi^2_\alpha(1)$.

– Under the null, we have

$$\begin{aligned}
-2\log\Lambda &= 2\log L\left(\widehat{\theta}\right) - 2\log L(\theta_0) \\
&= 2\sum_{i=1}^{n} \log f\left(X_i; \widehat{\theta}\right) - 2\sum_{i=1}^{n} \log f(X_i; \theta_0) \\
&= 2\left(\sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta_0)}{\partial\theta}\right)\left(\widehat{\theta} - \theta_0\right) + \left(\sum_{i=1}^{n} \frac{\partial^2 \log f\left(X_i; \widetilde{\theta}\right)}{\partial\theta^2}\right)\left(\widehat{\theta} - \theta_0\right)^2
\end{aligned}$$

for some $\widetilde{\theta}$ in between $\widehat{\theta}$ and θ_0. Using (6.10), we write

$$\begin{aligned}
2\left(\sum_{i=1}^{n} \frac{\partial \log f(X_i; \theta_0)}{\partial\theta}\right)\left(\widehat{\theta} - \theta_0\right) &= 2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right)\sqrt{n}\left(\widehat{\theta} - \theta_0\right) + o_p(1) \\
&= 2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right) I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right) + o_p(1) \\
&= 2I(\theta_0)\left(I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right)\right)^2 + o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
&\left(\sum_{i=1}^{n} \frac{\partial^2 \log f\left(X_i; \widetilde{\theta}\right)}{\partial\theta^2}\right)\left(\widehat{\theta} - \theta_0\right)^2 \\
&= -I(\theta_0)\left(I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right)\right)^2 + o_p(1).
\end{aligned}$$

Therefore, we obtain

$$-2\log\Lambda = \left(I(\theta_0)^{-1/2}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right)\right)^2 + o_p(1)$$
$$\xrightarrow{d} (N(0,1))^2 \sim \chi^2(1).$$

– We made an implicit assumption that

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta^2} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta_0)}{\partial\theta^2} + o_p(1),$$

which would imply

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta^2} = I(\theta_0) + o_p(1),$$

because

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f(X_i;\theta_0)}{\partial\theta^2} = I(\theta_0) + o_p(1)$$

by law of large numbers.

- Wald test rejects when

$$\left(\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\right)'\widehat{I}\left(\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\right) \geq \chi^2_\alpha(1)$$

where $\widehat{I}$ denotes a consistent estimator of $I(\theta)$

- The score test rejects the null when

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta_0)}{\partial\theta}\right)' I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f(X_i;\theta_0)}{\partial\theta}\right) \geq \chi^2_\alpha(1)$$

6.9 Maximum Likelihood Tests- Vector Case (Time Permitting)

- Suppose that we want to test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. There are three tests associated with the MLE $\widehat{\theta}$.

- We now assume that θ is a vector.

- Wald test rejects when

$$\left(\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\right)'\widehat{I}\left(\sqrt{n}\left(\widehat{\theta}-\theta_0\right)\right)\geq\chi^2_\alpha\left(\dim\left(\theta\right)\right)$$

  where $\widehat{I}$ denotes a consistent estimator of $I\left(\theta\right)$.

  - This is an asymptotic version of the test discussed in Example 10. There, we had $X\sim N\left(\mu,\Sigma\right)$, and that $H_0:\mu=0$ vs $H_1:\mu\neq 0$. We imposed a particular weight $(V=\Sigma)$, and concluded that the Neyman-Pearson test boils down to rejecting $H_0:\mu=0$ for large values of $X'\Sigma^{-\mathbf{1}}X$. Here, $\sqrt{n}\left(\widehat{\theta}-\theta_0\right)$ plays the role of X, and $I\left(\theta_0\right)^{-1}$ plays the role of Σ.

- Likelihood ratio test rejects H_0 when

$$\Lambda=\frac{L\left(\theta_0\right)}{L\left(\widehat{\theta}\right)}$$

  is small, where $L\left(\theta\right)=\prod_{i=1}^n f\left(X_i;\theta\right)$. More precisely, the null is rejected if $-2\log\Lambda\geq\chi^2_\alpha\left(\dim\left(\theta\right)\right)$.

  - The argument below requires familiarity with vector calculus, and presented here for completeness sake. Under the null, we have

$$\begin{aligned}-2\log\Lambda&=2\log L\left(\widehat{\theta}\right)-2\log L\left(\theta_0\right)\\&=2\sum_{i=1}^n\log f\left(X_i;\widehat{\theta}\right)-2\sum_{i=1}^n\log f\left(X_i;\theta_0\right)\\&=2\left(\sum_{i=1}^n\frac{\partial\log f\left(X_i;\theta_0\right)}{\partial\theta'}\right)\left(\widehat{\theta}-\theta_0\right)+\left(\widehat{\theta}-\theta_0\right)'\left(\sum_{i=1}^n\frac{\partial^2\log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta\partial\theta'}\right)\left(\widehat{\theta}-\theta_0\right)\end{aligned}$$

    for some $\widetilde{\theta}$ in between $\widehat{\theta}$ and θ_0. Using (6.10), we write

$$\begin{aligned}2\left(\sum_{i=1}^n\frac{\partial\log f\left(X_i;\theta_0\right)}{\partial\theta'}\right)\left(\widehat{\theta}-\theta_0\right)&=2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i\right)'\sqrt{n}\left(\widehat{\theta}-\theta_0\right)+o_p\left(1\right)\\&=2\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i\right)'I\left(\theta_0\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i\right)+o_p\left(1\right)\end{aligned}$$

    and

$$\begin{aligned}&\left(\widehat{\theta}-\theta_0\right)'\left(\sum_{i=1}^n\frac{\partial^2\log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta\partial\theta'}\right)\left(\widehat{\theta}-\theta_0\right)\\&=-\left(I\left(\theta_0\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i\right)\right)'I\left(\theta_0\right)\left(I\left(\theta_0\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i\right)\right)+o_p\left(1\right)\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}-2\log\Lambda &= \left(I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_i\right)\right)' I(\theta_0)\left(I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}s_i\right)\right)+o_p(1)\\ &\xrightarrow{d}\left(N\left(0,I^{-1}(\theta_0)\right)\right)'\cdot I(\theta_0)^{-1}\cdot N\left(0,I^{-1}(\theta_0)\right)\\ &\sim\chi^2(\dim(\theta))\end{aligned}$$

- The score test rejects the null when

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\log f(X_i;\theta_0)}{\partial\theta}\right)' I(\theta_0)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial\log f(X_i;\theta_0)}{\partial\theta}\right)\geq\chi^2_\alpha(\dim(\theta))$$

6.10 LR Test - An Extension

- We consider the extension where $\dim(\theta)=2$, and we only care about one component of θ.
- Recall that our generic notation is $H_0:\theta\in\omega_0$ vs. $H_1:\theta\in\omega_1$, where $\omega_0\cup\omega_1=\Omega$
- In this situation, our LR statistic takes the form

$$-2\log\frac{\sup_{\theta\in\omega_0}L(\theta)}{\sup_{\theta\in\Omega}L(\theta)}=2\sup_{\theta\in\Omega}\log L(\theta)-2\sup_{\theta\in\omega_0}\log L(\theta)$$

- As an example, we suppose that $X_1,\ldots,X_n$ is a random sample from $N(\mu,\sigma^2)$, and we would like to test $H_0:\mu=\mu_0$ against $H_1:\mu\neq\mu_0$.
- Note that

$$\begin{aligned}\log L(\theta)&=\log\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(X_i-\mu)^2}{2\sigma^2}\right)\\ &=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log\sigma^2-\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\sigma^2}\end{aligned}$$

- We first consider $\sup_{\theta\in\Omega}\log L(\theta)$.
  - Note that μ,σ^2 are unrestricted.
  - Therefore, the FOC is

$$\begin{aligned}0&=\frac{\partial\log L(\theta)}{\partial\mu}=\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma^2},\\ 0&=\frac{\partial\log L(\theta)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2(\sigma^2)^2}.\end{aligned}$$

– Therefore, the maximum is obtained at

$$\widehat{\mu} = \overline{X},$$
$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n}.$$

– It follows that

$$\begin{aligned}\sup_{\theta\in\Omega} \log L(\theta) = \log L\left(\widehat{\mu}, \widehat{\sigma}^2\right) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widehat{\sigma}^2 - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\widehat{\sigma}^2} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widehat{\sigma}^2 - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{n}} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\widehat{\sigma}^2.\end{aligned}$$

- We now consider $\sup_{\theta\in\omega_0} \log L(\theta)$.

  – Note that μ is restricted to be equal to μ_0, but σ^2 is unrestricted.

  – We therefore want to maximize

$$\log L\left(\mu_0, \sigma^2\right) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{2\sigma^2}$$

  with respect to σ^2 alone.

  – The FOC is

$$0 = \frac{\partial \log L(\mu_0, \sigma^2)}{\partial\sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{2\left(\sigma^2\right)^2}$$

  and therefore, the maximum is obtained at

$$\widetilde{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{n}.$$

  – It follows that

$$\begin{aligned}\sup_{\theta\in\omega_0} \log L(\theta) = \log L\left(\mu_0, \widetilde{\sigma}^2\right) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widetilde{\sigma}^2 - \frac{\sum_{i=1}^{n}(X_i-\mu)^2}{2\widetilde{\sigma}^2} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\widetilde{\sigma}^2 - \frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{2\frac{\sum_{i=1}^{n}(X_i-\mu_0)^2}{n}} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\widetilde{\sigma}^2.\end{aligned}$$

- To summarize, we have

$$\begin{aligned}
-2\log\frac{\sup_{\theta\in\omega_0}L(\theta)}{\sup_{\theta\in\Omega}L(\theta)} &= 2\sup_{\theta\in\Omega}\log L(\theta) - 2\sup_{\theta\in\omega_0}\log L(\theta) \\
&= 2\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\widehat{\sigma}^2\right) - 2\left(-\frac{n}{2}\log(2\pi) - \frac{n}{2} - \frac{n}{2}\log\widetilde{\sigma}^2\right) \\
&= n\log\widetilde{\sigma}^2 - n\log\widehat{\sigma}^2 \\
&= n\log\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2}
\end{aligned}$$

- Because

$$\begin{aligned}
\frac{\widetilde{\sigma}^2}{\widehat{\sigma}^2} &= \frac{\frac{\sum_{i=1}^n (X_i-\mu_0)^2}{n}}{\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n}} = \frac{\sum_{i=1}^n (X_i-\mu_0)^2}{\sum_{i=1}^n (X_i-\overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i-\overline{X})^2 + 2\sum_{i=1}^n (X_i-\overline{X})(\overline{X}-\mu_0) + n(\overline{X}-\mu_0)^2}{\sum_{i=1}^n (X_i-\overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i-\overline{X})^2 + n(\overline{X}-\mu_0)^2}{\sum_{i=1}^n (X_i-\overline{X})^2} = 1 + \frac{1}{n-1}\frac{(\overline{X}-\mu_0)^2}{\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n-1}\Big/ n} \\
&= 1 + \frac{1}{n-1}\left(\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right)^2
\end{aligned}$$

  where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i-\overline{X})^2$$

  we can write

$$-2\log\frac{\sup_{\theta\in\omega_0}L(\theta)}{\sup_{\theta\in\Omega}L(\theta)} = n\log\left(1 + \frac{1}{n-1}\left(\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right)^2\right).$$

- The LR test is usually understood to be an asymptotic test in the sense that the critical value is obtained by examining the asymptotic distribution of the test statistic. We can still attempt to obtain a critical value such that the size of the test is exact. For this purpose, note that the LR test rejects for a large value of $\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right|$. Also note that under the null, the "test statistic" $\frac{\overline{X}-\mu_0}{S/\sqrt{n}}$ has $t(n-1)$ distribution. Therefore, the exact finite sample version of the LR test rejects when $\left|\frac{\overline{X}-\mu_0}{S/\sqrt{n}}\right| \geq t_{\alpha/2}(n-1)$. Does it look familiar?

6.11 Homework

1. (Textbook p. 428) Let $X_1, \ldots, X_{10}$ denote a random sample from $N(0, \sigma^2)$. Find a best critical region of size $\alpha = 5\%$ for testing $H_0 : \sigma^2 = 1$ against $H_1 : \sigma^2 = 2$. Does your answer change if the alternative is now $H_1 : \sigma^2 = 4$?

2. (Textbook p. 436) Let $X_1, \ldots, X_n$ denote a random sample from $N(0, \theta)$. Show that the set $\{(x_1, \ldots, x_n) : \sum_{i=1}^n x_i^2 \leq c\}$ is a uniformly most powerful test for $H_0 : \theta = 1$ against $H_1 : \theta < 1$.

3. (Textbook p. 436) Let $X_1, \ldots, X_n$ denote a random sample from $N(0, \theta)$. Show that there is no uniformly most powerful test for $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$.

4. (Textbook p. 428) Let $X_1, \ldots, X_n$ denote a random sample from a distribution with pdf of the form $f(x; \theta) = \theta x^{\theta - 1}$, $0 < x < 1$. Derive the best critical region for testing $H_0 : \theta = 1$ against $H_1 : \theta = 2$. Does it look like $\{(x_1, \ldots, x_n) : \prod_{i=1}^n x_i \geq c\}$?

5. Textbook (p. 428) Let $X_1, \ldots, X_n$ denote a random sample from $N(\theta, 10^2)$. Show that $\{(x_1, \ldots, x_n) : \overline{x} = n^{-1} \sum_{i=1}^n x_i \geq c\}$ is the best critical region for testing $H_0 : \theta = 75$ against $H_1 : \theta = 78$. Find n and c such that the size of the test is 5% and the power of the test is 90%, approximately.

6. Let $X_1, \ldots, X_n$ denote a random sample from $N(\theta, 1)$. We would like to test $H_0 : \theta = 5$ against $H_1 : \theta \neq 5$.

   (a) Prove that the LR statistic is

   $$2\left(\log L\left(\widehat{\theta}\right) - \log L(5)\right) = \sum_{i=1}^n \left((X_i - 5)^2 - \left(X_i - \overline{X}\right)^2\right)$$

   Hint: We have

   $$f(x; \theta) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x - \theta)^2}{2}\right)$$

   so

   $$\log L(\theta) = \sum_{i=1}^n \left(-\frac{1}{2}\log(2\pi) - \frac{(X_i - \theta)^2}{2}\right)$$

   The FOC for the problem $\max_\theta \log L(\theta)$ is

   $$\sum_{i=1}^n \left(X_i - \widehat{\theta}\right) = 0$$

   from which we derive the MLE $\widehat{\theta} = \overline{X}$.

   (b) Prove that the Wald statistic, using $I\left(\widehat{\theta}\right)^{-1}$ for $\widehat{I}$, is equal to

   $$\left(\sqrt{n}\left(\overline{X} - 5\right)\right)'\left(\sqrt{n}\left(\overline{X} - 5\right)\right) = n\left(\overline{X} - 5\right)^2.$$

   Hint: We have

   $$\frac{\partial^2 \log f(x; \theta)}{\partial \theta^2} = \frac{\partial^2\left(-\frac{1}{2}\log(2\pi) - \frac{(x-\theta)^2}{2}\right)}{\partial \theta^2} = -1$$

so

$$I(\theta) = -E[-1] = 1,$$

i.e., it does not depend on θ.

(c) Prove that the score test statistic is equal to

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 5)\right)'\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i - 5)\right) = n\left(\overline{X} - 5\right)^2.$$

Hint: We have

$$\frac{\partial \log f(X_i;\theta_0)}{\partial\theta} = \left.\frac{\partial\left(-\frac{1}{2}\log(2\pi) - \frac{(X_i-\theta)^2}{2}\right)}{\partial\theta}\right|_{\theta=5} = X_i - 5$$

and

$$I(\theta_0) = I(5) = 1.$$

7. (Textbook p. 340) Let $X_1, \ldots, X_n$ denote a random sample from a Poisson distribution with mean θ. We would like to test $H_0 : \theta = 5$ against $H_1 : \theta \neq 5$.

(a) Prove that the LR statistic is equal to

$$2\left(\log L\left(\widehat{\theta}\right) - \log L(5)\right) = 2n\left(\overline{X}\log\overline{X} - 5\log 5 - \overline{X} + 5\right).$$

Hint: We have

$$f(x;\theta) = \frac{\theta^x e^{-\theta}}{x!}$$

so

$$\begin{aligned}\log L(\theta) &= \sum_{i=1}^{n}\log\frac{\theta^{X_i}e^{-\theta}}{X_i!} = \sum_{i=1}^{n}X_i\log\theta - \sum_{i=1}^{n}\theta - \sum_{i=1}^{n}\log(X_i!)\\ &= n\overline{X}\log\theta - n\theta - \sum_{i=1}^{n}\log(X_i!)\end{aligned}$$

The FOC for the problem $\max_\theta \log L(\theta)$ is

$$0 = \frac{n\overline{X}}{\widehat{\theta}} - n$$

from which we derive the MLE $\widehat{\theta} = \overline{X}$.

(b) Prove that the Wald statistic, using $I\left(\widehat{\theta}\right)^{-1}$ for $\widehat{I}$, is equal to

$$\left(\sqrt{n}\left(\overline{X} - 5\right)\right)'\frac{1}{\overline{X}}\left(\sqrt{n}\left(\overline{X} - 5\right)\right) = \frac{n\left(\overline{X} - 5\right)^2}{\overline{X}}.$$

Hint: We have

$$\frac{\partial^2 \log f(x;\theta)}{\partial \theta^2} = \frac{\partial^2 (x \log \theta - \theta - \log(x!))}{\partial \theta^2} = -\frac{x}{\theta^2}$$

so

$$I(\theta) = -E\left[-\frac{X}{\theta^2}\right] = \frac{E[X]}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}$$

and

$$I\left(\widehat{\theta}\right) = \frac{1}{\widehat{\theta}}$$

(c) Prove that the score test statistic is

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i}{5}-1\right)\right)'\left(\frac{1}{5}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{X_i}{5}-1\right)\right) = \frac{n\left(\overline{X}-5\right)^2}{5}.$$

Hint: We have

$$\frac{\partial \log f(X_i;\theta_0)}{\partial \theta} = \left.\frac{\partial (X_i \log \theta - \theta - \log(X_i!))}{\partial \theta}\right|_{\theta=5} = \frac{X_i}{5} - 1$$

and

$$I(\theta_0) = I(5) = \frac{1}{5}.$$

6.12 Technical Details - Some Algebra

Lemma 2

$$\frac{\int \frac{1}{\left(\sigma\sqrt{2\pi}\right)^k} \exp\left(-\frac{(x-\mu)'(x-\mu)}{2\sigma^2}\right)\Gamma(d\mu)}{\frac{1}{\left(\sigma\sqrt{2\pi}\right)^k}\exp\left(-\frac{x'x}{2\sigma^2}\right)} = \exp\left(-\frac{\beta^2}{2\sigma^2}\right)\int_0^{\beta}\cosh\left(\frac{|x|}{\sigma^2}u\right)2f(u)\,du$$

Proof. We first write the numerator as

$$\int \frac{1}{\left(\sigma\sqrt{2\pi}\right)^k}\exp\left(-\frac{(x-\mu)'(x-\mu)}{2\sigma^2}\right)\Gamma(d\mu) = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^k}E\left[\exp\left(-\frac{(x-U)'(x-U)}{2\sigma^2}\right)\right].$$

Because

$$(x-U)'(x-U) = -x'x + 2x'U - U'U = -x'x + 2x'U - \beta^2,$$

and because $x'U$ has the same distribution as $|x|U_1$, we get

$$\frac{1}{\left(\sigma\sqrt{2\pi}\right)^k}E\left[\exp\left(-\frac{(x-U)'(x-U)}{2\sigma^2}\right)\right] = \frac{1}{\left(\sigma\sqrt{2\pi}\right)^k}\exp\left(-\frac{x'x+\beta^2}{2\sigma^2}\right)E\left[\exp\left(\frac{|x|}{\sigma^2}U_1\right)\right].$$

The counterpart of the ratio on the far LHS then becomes

$$\frac{\int \frac{1}{\left(\sigma\sqrt{2\pi}\right)^k} \exp\left(-\frac{(x-\mu)'(x-\mu)}{2\sigma^2}\right) \Gamma\left(d\mu\right)}{\frac{1}{\left(\sigma\sqrt{2\pi}\right)^k} \exp\left(-\frac{x'x}{2\sigma^2}\right)} = \exp\left(-\frac{\beta^2}{2\sigma^2}\right) E\left[\exp\left(\frac{|x|}{\sigma^2}U_1\right)\right].$$

We will now use an intuitive property of U_1, i.e., that it has a symmetric distribution on $(-\beta, \beta)$. Therefore, we can write

$$E\left[\exp\left(\frac{|x|}{\sigma^2}U_1\right)\right] = \int_{-\beta}^{\beta} \exp\left(\frac{|x|}{\sigma^2}u\right) g\left(u\right) du,$$

where $g\left(u\right)$ is an even function. Therefore,

$$\begin{aligned}
E\left[\exp\left(\frac{|x|}{\sigma^2}U_1\right)\right] &= \int_0^{\beta} \exp\left(\frac{|x|}{\sigma^2}u\right) g\left(u\right) du + \int_{-\beta}^{0} \exp\left(\frac{|x|}{\sigma^2}u\right) g\left(u\right) du \\
&= \int_0^{\beta} \exp\left(\frac{|x|}{\sigma^2}u\right) g\left(u\right) du + \int_{\beta}^{0} \exp\left(-\frac{|x|}{\sigma^2}v\right) g\left(-v\right) (-dv) \\
&= \int_0^{\beta} \exp\left(\frac{|x|}{\sigma^2}u\right) g\left(u\right) du + \int_0^{\beta} \exp\left(-\frac{|x|}{\sigma^2}u\right) g\left(u\right) du \\
&= \int_0^{\beta} \cosh\left(\frac{|x|}{\sigma^2}u\right) 2g\left(u\right) du,
\end{aligned}$$

where we used the change-of-variable $u = -v$ for the second equality. ■

Lemma 3 *If $V = \Sigma$, (6.5) is proportional to $\exp\left(\frac{1}{4}x'\Sigma^{-1}x\right)$.*

Proof. When $V = \Sigma$, we have (6.5) is equal to

$$\begin{aligned}
&\frac{\int \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(x-\mu\right)'\Sigma^{-1}\left(x-\mu\right)\right) \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\mu'\Sigma^{-1}\mu\right) d\mu}{\frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}x'\Sigma^{-1}x\right)} \\
&= \frac{1}{(2\pi)^{k/2}\left|\Sigma\right|^{1/2}} \int \exp\left(-\frac{1}{2}\left(x-\mu\right)'\Sigma^{-1}\left(x-\mu\right)\right) \exp\left(-\frac{1}{2}\mu'\Sigma^{-1}\mu\right) \exp\left(\frac{1}{2}x'\Sigma^{-1}x\right) d\mu \\
&= \frac{1}{(2\pi)^{k/2}\left|\Sigma\right|^{1/2}} \int \exp\left(-\frac{1}{2}\left(x-\mu\right)'\Sigma^{-1}\left(x-\mu\right) - \frac{1}{2}\mu'\Sigma^{-1}\mu + \frac{1}{2}x'\Sigma^{-1}x\right) d\mu
\end{aligned}$$

Because

$$\begin{aligned}
&-\frac{1}{2}\left(x-\mu\right)'\Sigma^{-1}\left(x-\mu\right) - \frac{1}{2}\mu'\Sigma^{-1}\mu + \frac{1}{2}x'\Sigma^{-1}x \\
&= -\mu'\Sigma^{-1}\mu + \mu'\Sigma^{-1}x \\
&= -\left(\mu - \frac{1}{2}x\right)'\Sigma^{-1}\left(\mu - \frac{1}{2}x\right) + \frac{1}{4}x'\Sigma^{-1}x \\
&= -\frac{1}{2}\left(\mu - \frac{1}{2}x\right)'\left(\frac{1}{2}\Sigma\right)^{-1}\left(\mu - \frac{1}{2}x\right) + \frac{1}{4}x'\Sigma^{-1}x
\end{aligned}$$

Therefore, (6.5) is equal to

$$\begin{aligned}
&\frac{1}{(2\pi)^{k/2}\left|\Sigma\right|^{1/2}}\exp\left(\frac{1}{4}x'\Sigma^{-\mathbf{1}}x\right)\int\exp\left(-\frac{1}{2}\left(\mu-\frac{1}{2}x\right)'\left(\frac{1}{2}\Sigma\right)^{-\mathbf{1}}\left(\mu-\frac{1}{2}x\right)\right)d\mu \\
&=\frac{(2\pi)^{k/2}\left|\frac{1}{2}\Sigma\right|^{1/2}}{(2\pi)^{k/2}\left|\Sigma\right|^{1/2}}\exp\left(\frac{1}{4}x'\Sigma^{-\mathbf{1}}x\right)\int\frac{1}{(2\pi)^{k/2}\left|\frac{1}{2}\Sigma\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\mu-\frac{1}{2}x\right)'\left(\frac{1}{2}\Sigma\right)^{-\mathbf{1}}\left(\mu-\frac{1}{2}x\right)\right)d\mu \\
&=\frac{\left|\frac{1}{2}\Sigma\right|^{1/2}}{\left|\Sigma\right|^{1/2}}\exp\left(\frac{1}{4}x'\Sigma^{-\mathbf{1}}x\right),
\end{aligned}$$

where we used the fact that

$$\frac{1}{(2\pi)^{k/2}\left|\frac{1}{2}\Sigma\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\mu-\frac{1}{2}x\right)'\left(\frac{1}{2}\Sigma\right)^{-\mathbf{1}}\left(\mu-\frac{1}{2}x\right)\right)$$

integrates up to 1 because it is the PDF of $N\left(\frac{1}{2}x,\frac{1}{2}\Sigma\right)$. ■

6.13 Why (6.9)

- We can see that the power at $\theta_n=\frac{h}{\sqrt{n}}$ is equal to

$$\begin{aligned}
&P_{\theta_n}\left(\sqrt{n}\left(T_n-\mu\left(0\right)\right)>\sigma\left(0\right)z_\alpha\right) \\
&=P_{\theta_n}\left(\sqrt{n}\left(T_n-\mu\left(\theta_n\right)\right)+\sqrt{n}\left(\mu\left(\theta_n\right)-\mu\left(0\right)\right)>\sigma\left(0\right)z_\alpha\right) \\
&=P_{\theta_n}\left(\frac{\sqrt{n}\left(T_n-\mu\left(\theta_n\right)\right)}{\sigma\left(\theta_n\right)}>\frac{\sigma\left(0\right)}{\sigma\left(\theta_n\right)}z_\alpha-\sqrt{n}\frac{\mu\left(\theta_n\right)-\mu\left(0\right)}{\sigma\left(\theta_n\right)}\right)
\end{aligned}$$

  Assuming that (i) $\sigma\left(\cdot\right)$ is continuous; and (ii) $\mu\left(\cdot\right)$ is differentiable, we can see that the power converges to

$$P\left(Z>z_\alpha-\frac{\mu'\left(0\right)h}{\sigma\left(0\right)}\right)=1-\Phi\left(z_\alpha-\frac{\mu'\left(0\right)}{\sigma\left(0\right)}h\right)=\Phi\left(\frac{\mu'\left(0\right)}{\sigma\left(0\right)}h-z_\alpha\right)$$

  where $Z\sim N\left(0,1\right)$. So an efficient test would maximize $\frac{\mu'(0)h}{\sigma(0)}$.

- In order to get some insight about this sort of analysis, assume (as is often the case) that under $\theta=0$,

$$\sqrt{n}\left(T_n-\mu\left(0\right)\right)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi_i+o_p\left(1\right)$$

  where $\psi_i=\psi\left(X_i\right)$ has a zero expectation. By the CLT, we should have $\sqrt{n}\left(T_n-\mu\left(0\right)\right)\rightarrow N\left(0,E\left[\psi_i^2\right]\right)$ under $\theta=0$. There is a result called Le Cam's Third Lemma, which is

useful in this situation. According to the result, we should have $\sqrt{n}\left(T_n - \mu\left(0\right)\right) \rightarrow N\left(E\left[s_i\psi_i\right]h, E\left[\psi_i^2\right]\right)$ under $\theta_n = \frac{h}{\sqrt{n}}$. Comparing with

$$\begin{aligned}\sqrt{n}\left(T_n - \mu\left(\theta_n\right)\right) &= \sqrt{n}\left(T_n - \mu\left(0\right)\right) - \sqrt{n}\left(\mu\left(\theta_n\right) - \mu\left(0\right)\right) \\ &\rightarrow \sqrt{n}\left(T_n - \mu\left(0\right)\right) - \mu'\left(0\right)h \\ &= N\left(\left(E\left[s_i\psi_i\right] - \mu'\left(0\right)\right)h, E\left[\psi_i^2\right]\right)\end{aligned}$$

we can see that

$$\begin{aligned}\mu'\left(0\right) &= E\left[s_i\psi_i\right] \\ \sigma^2\left(0\right) &= E\left[\psi_i^2\right]\end{aligned}$$

so we may want to maximize

$$\frac{E\left[s_i\psi_i\right]}{\sqrt{E\left[\psi_i^2\right]}}$$

Because

$$\frac{\left(E\left[s_i\psi_i\right]\right)^2}{E\left[\psi_i^2\right]} \leq \frac{E\left[s_i^2\right]E\left[\psi_i^2\right]}{E\left[\psi_i^2\right]} = E\left[s_i^2\right]$$

we can see that the maximum can be obtained by choosing $\psi_i \propto s_i$, and the optimal test would have the property

$$P_{\theta_n}\left(\sqrt{n}\left(T_n - \mu\left(0\right)\right) > \sigma\left(0\right)z_\alpha\right) = \Phi\left(\sqrt{E\left[s_i^2\right]}h - z_\alpha\right)$$

Chapter 7

Introduction to MLE (Time Permitting)

- Suppose that $X_1, \ldots, X_n$ are iid with PDF $f(x;\theta)$.
- We estimate θ by

$$\arg\max \prod_{i=1}^{n} f(X_i;\theta) = \arg\max \sum_{i=1}^{n} \log f(X_i;\theta)$$

Example 12 *If X_is are Bernoulli(θ) with the PDF $\theta^z (1-\theta)^{1-z}$, the MLE maximizes*

$$\begin{aligned}\sum_{i=1}^{n} \log\left(\theta^{X_i}(1-\theta)^{1-X_i}\right) &= \sum_{i=1}^{n} \left(X_i \log\theta + (1-X_i)\log(1-\theta)\right) \\ &= \left(\sum_{i=1}^{n} X_i\right)\log\theta + \left(n - \sum_{i=1}^{n} X_i\right)\log(1-\theta) \\ &= n\left(\overline{X}\log\theta + \left(1-\overline{X}\right)\log(1-\theta)\right)\end{aligned}$$

A straightforward calculation reveals that the MLE should be equal to $\overline{X}$ in this case.

- The MLE can be motivated by the following. Let θ^* denote the true value of θ. Suppose that for all $\theta \in \Theta$ such that $\theta \neq \theta^*$, there exists a set A of values for z such that

$$\int_A f(x;\theta)\,dx \neq \int_A f(x;\theta^*)\,dx$$

  for all x. Then, $E[\log f(Z;\theta)]$ is uniquely maximized at θ^*.

  - Because log is a concave function, we can use Jensen's Inequality and conclude that

$$E[\log f(X;\theta)] - E[\log f(X;\theta^*)] = E\left[\log\frac{f(X;\theta)}{f(X;\theta^*)}\right] \leq \log E\left[\frac{f(X;\theta)}{f(X;\theta^*)}\right]$$

But

$$E\left[\frac{f(X;\theta)}{f(X;\theta^*)}\right] = \int \frac{f(x;\theta)}{f(x;\theta^*)} f(x;\theta^*)\,dx = \int f(x;\theta)\,dz = 1$$

and hence

$$E[\log f(X;\theta)] - E[\log f(X;\theta^*)] \leq \log(1) = 0$$

for all θ. In other words,

$$E[\log f(X;\theta)] \leq E[\log f(X;\theta^*)]$$

for all θ. Because the log is a strictly concave function, the inequality is strict as long as $f(X;\theta)/f(X;\theta^*)$ is not degenerate, which is ruled out by our assumption. Therefore, θ^* uniquely maximizes $Q(\theta) = E[\log f(X;\theta)]$.

Lemma 4 $E[s(X;\theta)] = 0$

Proof. Because

$$1 = \int f(x;\theta)\,dx$$

we have

$$0 = \int \frac{\partial f(x;\theta)}{\partial \theta} dx = \int \frac{\frac{\partial f(x;\theta)}{\partial \theta}}{f(x;\theta)} f(x;\theta)\,dx = \int s(x;\theta) f(x;\theta)\,dx$$

■

Definition 38 (Fisher Information)

$$I(\theta) = \int s(x;\theta) s(x;\theta)' f(x;\theta) dx = E\left[s(X;\theta) s(X;\theta)'\right].$$

Theorem 68

$$I(\theta) = -\int \frac{\partial^2 \log f(x;\theta)}{\partial\theta\partial\theta'} f(x;\theta)\,dz = -E\left[\frac{\partial^2 \log f(X;\theta)}{\partial\theta\partial\theta'}\right].$$

Proof. Because

$$0 = \int s(x;\theta) f(x;\theta)\,dx$$

we have

$$\begin{aligned}
0 &= \int \frac{\partial s(x;\theta)}{\partial\theta'} f(x;\theta)\,dx + \int s(x;\theta) \frac{\partial f(x;\theta)}{\partial\theta'} dx \\
&= \int \frac{\partial}{\partial\theta'}\left(\frac{\partial^2 \log f(x;\theta)}{\partial\theta}\right) f(x;\theta)\,dx + \int s(x;\theta) \frac{\frac{\partial f(x;\theta)}{\partial\theta'}}{f(x;\theta)} f(x;\theta)\,dx \\
&= \int \frac{\partial^2 \log f(x;\theta)}{\partial\theta\partial\theta'} f(x;\theta)\,dx + \int s(x;\theta) s(x;\theta)' f(x;\theta)\,dx
\end{aligned}$$

■

Example 13 *Suppose that* $X \sim N(\mu, \sigma^2)$*. Assume that* σ^2 *is known. The Fisher information* $I(\mu)$ *can be calculated in the following way. Notice that*

$$f(x;\mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]$$

so that

$$\log f(x;\mu) = C - \frac{(x-\mu)^2}{2\sigma^2}$$

where C *denotes the part of the* $\log f$ *wich does not depend on* μ*. Because*

$$s(x;\mu) = \frac{x-\mu}{\sigma^2}$$

we have

$$I(\mu) = E\left[s(X;\mu)^2\right] = \frac{1}{\sigma^4} E\left[(X-\mu)^2\right] = \frac{1}{\sigma^2}$$

Remark 17 *In the multivariate case where* $\theta = (\theta_1, \ldots, \theta_K)$*, we let*

$$s(x;\theta) = \frac{\partial \log f(x;\theta)}{\partial \theta} \equiv \begin{pmatrix} \frac{\partial \log f(x;\theta)}{\partial \theta_1} \\ \vdots \\ \frac{\partial \log f(x;\theta)}{\partial \theta_K} \end{pmatrix}.$$

and

$$I(\theta) = E\left[s(X;\theta)\, s(X;\theta)'\right] = -E\left[\frac{\partial^2 \log f(x;\theta)}{\partial \theta \partial \theta'}\right] = -E\begin{bmatrix} \frac{\partial^2 \log f(X;\theta)}{\partial \theta_1 \partial \theta_1} & \frac{\partial^2 \log f(X;\theta)}{\partial \theta_1 \partial \theta_K} \\ \\ \frac{\partial^2 \log f(X;\theta)}{\partial \theta_K \partial \theta_1} & \frac{\partial^2 \log f(X;\theta)}{\partial \theta_K \partial \theta_K} \end{bmatrix}$$

Example 14 *Suppose that* X *is from* $N(\theta_1, \theta_2)$*. Then,*

$$\log f(x;\theta_1,\theta_2) = -\frac{1}{2}\log(2\pi\theta_2) - \frac{(x-\theta_1)^2}{2\theta_2}$$

so that

$$s(x;\theta_1,\theta_2) = \begin{pmatrix} \partial \log f/\partial\theta_1 \\ \partial \log f/\partial\theta_2 \end{pmatrix} = \begin{pmatrix} \frac{x-\theta_1}{\theta_2} \\ -\frac{1}{2\theta_2} + \frac{(x-\theta_1)^2}{2\theta_2^2} \end{pmatrix}$$

from which we obtain

$$E[ss'] = \begin{pmatrix} \frac{1}{\theta_2} & 0 \\ 0 & \frac{1}{2\theta_2^2} \end{pmatrix}$$

In this calculation, I used the fact that $E[Z^{2m}] = (2m)!/(2^m m!)$ *and* $E[Z^{2m-1}] = 0$ *if* $Z \sim N(0,1)$*.*

Theorem 69 (Asymptotic Normality of MLE)

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}-\theta^*\right) &= I\left(\theta^*\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\right)+o_p\left(1\right)\\ &\xrightarrow{d}\mathcal{N}\left(0,I^{-1}\left(\theta^*\right)\right)\end{aligned}$$

Sketch of Proof. We will assume that the MLE is consistent. By the FOC, we have

$$\begin{aligned}0&=\sum_{i=1}^{n}\frac{\partial \log f\left(X_i;\widehat{\theta}\right)}{\partial\theta}\\ &=\sum_{i=1}^{n}\frac{\partial \log f\left(X_i;\theta^*\right)}{\partial\theta}+\left(\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta\partial\theta'}\right)\left(\widehat{\theta}-\theta^*\right)\end{aligned}$$

where the second equality is justified by the mean value theorem. Here, the $\widetilde{\theta}$ is on the line segment adjoining $\widehat{\theta}$ and θ^*. It follows that

$$\sqrt{n}\left(\widehat{\theta}-\theta^*\right)=-\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f\left(X_i;\theta^*\right)}{\partial\theta}\right)$$

It can be shown that, under some regularity conditions,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta\partial\theta'}-\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\theta^*\right)}{\partial\theta\partial\theta'}\to 0$$

in probability. Because

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\theta^*\right)}{\partial\theta\partial\theta'}\to E\left[\frac{\partial^2 \log f\left(X_i;\theta^*\right)}{\partial\theta\partial\theta'}\right]$$

in probability, we conclude that

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \log f\left(X_i;\widetilde{\theta}\right)}{\partial\theta\partial\theta'}=E\left[\frac{\partial^2 \log f\left(X_i;\theta^*\right)}{\partial\theta\partial\theta'}\right]=-I\left(\theta^*\right) \tag{7.1}$$

We also note that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial \log f\left(X_i;\theta^*\right)}{\partial\theta}\xrightarrow{d}\mathcal{N}\left(0,I\left(\theta^*\right)\right) \tag{7.2}$$

by the central limit theorem. Combining (7.1) and (7.2), we obtain the desired conclusion. ■