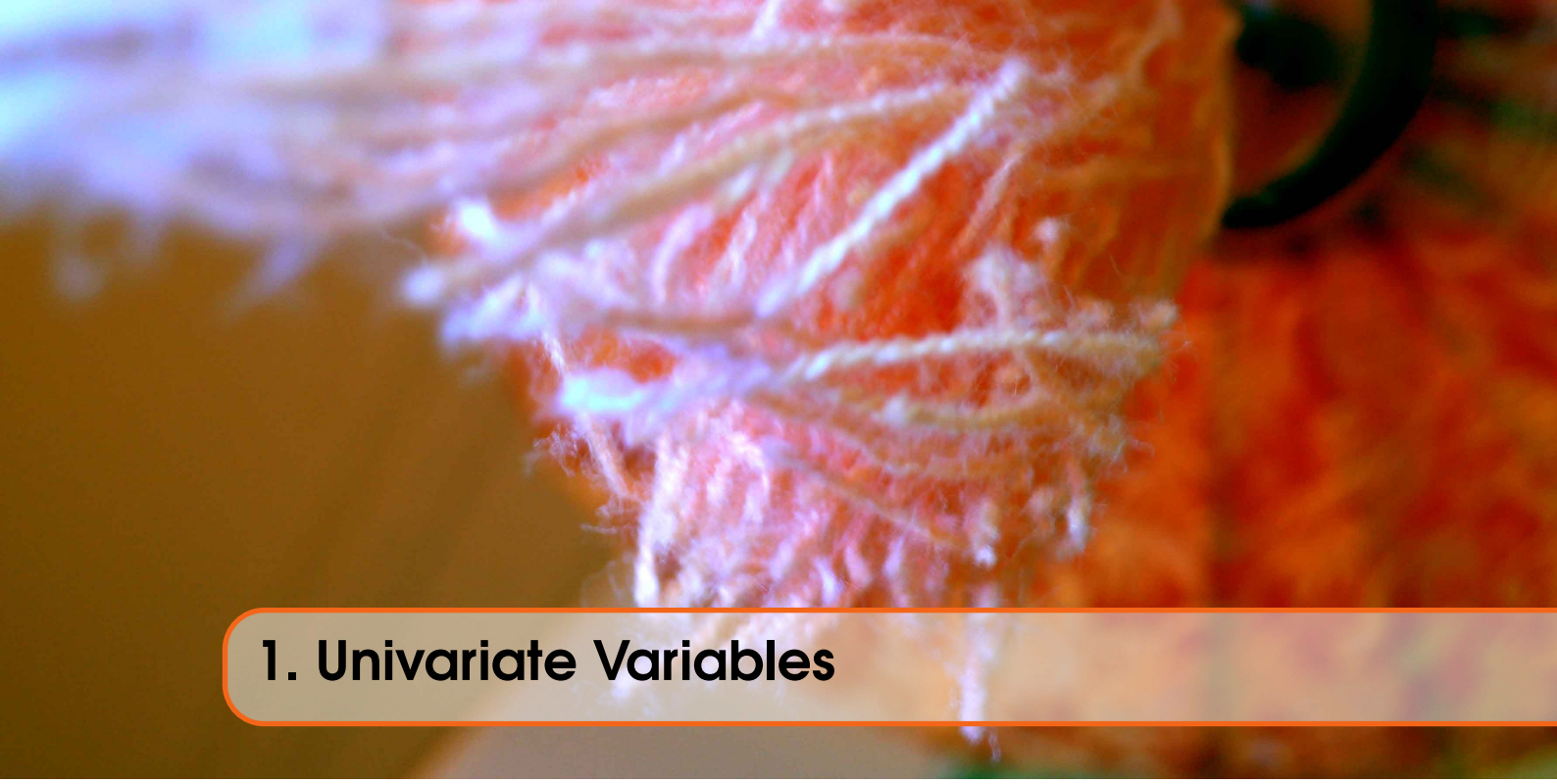
S

STAT 330 - Mathematical Statistics

Prof. Reza Ramezan

Contents



1. Univariate Variables

1.1 Probability

Definition 1.1.1 A **sample space** S is the set of all distinct outcomes for a random experiment with the property that in a trial, only one of the outcomes occurs.

Definition 1.1.2 $B \subseteq \mathcal{P}(S)$, where S is a sample space, is a **sigma algebra** if

1. $\emptyset \in B$
2. if $A \in B$, then $S \setminus A \in B$
3. if $A_1, A_2, \ldots \in B$, then $\bigcup_{i=1}^{\infty} A_i \in B$

Definition 1.1.3 Let B be a sigma algebra associated with sample space S. A **probability set function** is $P : B \to \mathbb{R}$ such that

1. $P(A) \geq 0$ for all $A \in B$.
2. $P(S) = 1$.
3. If $A_1, A_2, \ldots \in B$ are pairwise mutually exclusive, then $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Proposition 1.1.4 Let B be a sigma algebra with sample space S and $P : B \to \mathbb{R}$ be a probability set function. Let $C, D \in B$. Then

1. $P(\emptyset) = 0$.
2. $P(C) \leq 1$.
3. $P(C) = 1 - P(S \setminus C)$.
4. if $C \subseteq D$, then $P(C) \leq P(D)$.
5. $P(C \cup D) = P(C) + P(D) - P(C \cap D)$.

Proof. (1) Clearly $\emptyset \cap \emptyset = \emptyset$, so $P(\emptyset \cup \emptyset) = P(\emptyset) = P(\emptyset) + P(\emptyset)$. Since $P(\emptyset) \geq 0$ by Def. 1.1.3, $P(\emptyset) = 0$.
(2) $(S \setminus C) \cap C = \emptyset$, so by Def. 1.1.3(3), $P((S \setminus C) \cup C) = P(S \setminus C) + P(C)$, but $P((S \setminus C) \cup C) = P(S) = 1$ by Def. 1.1.3(2), so $P(C) = 1 - P(S \setminus C)$. Now $P(S \setminus C) \geq 0$ by Def. 1.1.3(1), so it follows that $P(C) \leq 1$.
(3) See the proof of (2).

(4) Write $T = D \setminus C$. It follows that $T \cap C = \emptyset$. Moreover, $T \cup C = D$. By Def. 1.1.3(3), $P(D) = P(T) + P(C)$. Re-arranging yields $P(C) = P(D) - P(T)$ where $P(T) \geq 0$ by Def. 1.1.3(1). Thus $P(C) \leq P(D)$.
(5) Denote $D^c = S \setminus D$. Observe that $C = (C \cap D) \cup (C \cap D^c)$ where $C \cap D$ and $C \cap D^c$ are mutually exclusive. Thus

$$P(C \cap D^c) = P(C) - P(C \cap D)$$

and similarly

$$P(D \cap C^c) = P(D) - P(C \cap D)$$

Now $C \cup D = (C^c \cap D) \cup (C \cap D) \cup (C \cap D^c)$ and these three events are pairwise mutually exclusive, so $P(C \cup D) = P(D \cap C^c) + P(C \cap D) + P(C \cap D^c) = P(C) - P(C \cap D) + P(C \cap D) + P(D) - P(C \cap D) = P(C) + P(D) - P(C \cap D)$ as required. ■

Proposition 1.1.5 Let $\mathcal{B}$ be a sigma algebra over sample space S and $P : \mathcal{B} \to \mathbb{R}$ a probability function, then

1. (Boole's Inequality) if $A_1, A_2, \ldots \in \mathcal{B}$, then

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

.

2. (Bonferroni's Inequality) if $A_1, A_2, \ldots A_k \in \mathcal{B}$, then

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

3. (Continuity Property) if $A_1 \subseteq A_2 \subseteq \ldots$ is a sequence of nested sets in $\mathcal{B}$ and $A = \bigcup_{i=1}^{\infty} A_i$, then

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

On the other hand, if $B_1 \supseteq B_2 \supseteq \ldots$ are sets in $\mathcal{B}$ and $B = \bigcap_{i=1}^{\infty} B_i$, then

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

Proof. (1) Clearly the case for $n = 1$ holds. Suppose inductively that for $n \in \mathbb{N}$, $P\left(\bigcup_{i=1}^{n} A_i\right) \leq$

$\sum_{i=1}^{n} P(A_i)$, then by Proposition 1.1.4(5),

$$\begin{aligned} P\left(\bigcup_{i=1}^{n+1} A_i\right) &= P\left(\bigcup_{i=1}^{n} A_i \cup A_{n+1}\right) \\ &= P\left(\bigcup_{i=1}^{n} A_i\right) + P(A_{n+1}) - P\left(\bigcup_{i=1}^{n} A_i \cap A_{n+1}\right) \\ &\leq \sum_{i=1}^{n} P(A_i) + P(A_{n+1}) - P\left(\bigcup_{i=1}^{n} A_i \cap A_{n+1}\right) \\ &= \sum_{i=1}^{n+1} P(A_i) - P\left(\bigcup_{i=1}^{n} A_i \cap A_{n+1}\right) \\ &\leq \sum_{i=1}^{n+1} P(A_i) \end{aligned}$$

as required.

(2) By De Morgan's Law, $\left(\bigcup_{i=1}^{k} A_i^c\right)^c = \bigcap_{i=1}^{k} (A_i^c)^c = \bigcap_{i=1}^{k} A_i$, and so $P\left[\left(\bigcup_{i=1}^{k} A_i^c\right)^c\right] = P\left(\bigcap_{i=1}^{k} (A_i^c)^c\right) = P\left(\bigcap_{i=1}^{k} A_i\right)$.

But $P\left[\left(\bigcup_{i=1}^{k} A_i^c\right)^c\right] = 1 - P\left(\bigcup_{i=1}^{k} A_i^c\right)$ where $P\left(\bigcup_{i=1}^{k} A_i^c\right) \leq \sum_{i=1}^{k} P(A_i^c)$ by part (1), so it follows that

$$P\left(\bigcap_{i=1}^{k} A_i\right) = P\left[\left(\bigcup_{i=1}^{k} A_i^c\right)^c\right] \geq 1 - \sum_{i=1}^{k} P(A_i^c).$$

(3) Define $C_1 = A_1, C_2 = A_2 \setminus A_1, C_3 = A_3 \setminus A_2, \ldots$ and recursively $C_n = A_n \setminus A_{n-1}$. It follows that $C_i \cap C_{i+1} = \emptyset$ for all $i \in \mathbb{N}$ and in general $\bigcup_{i=1}^{n} C_i = A_n$ for $n \geq 1$.

Consequently $\bigcup_{i=1}^{n} C_i = \bigcup_{i=1}^{n} A_i$, except the C_i's are by construction mutually exclusive. By Def. 1.1.3(3),

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

and thus

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

as required.

For the second part, note that by definition $B_1^c \subseteq B_2^c \subseteq \ldots$. By what we have just proved,

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

Apply De Morgan's Law to both sides yields

$$\lim_{n\to\infty} P\left[\left(\bigcap_{i=1}^{n} B_i\right)^c\right] = P\left[\left(\bigcap_{i=1}^{\infty} B_i\right)^c\right]$$

and it follows that

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

as required. ■

Definition 1.1.6 Let S be a sample space, B be a sigma algebra (with respect to S), and $P : B \to \mathbb{R}$ a probability function, then (S,B) is called a **measurable space**, while (S,B,P) is called a **probability space**.

Definition 1.1.7 Given a probability space (S,B,P), suppose $C,D \in B$ with $P(D) > 0$, then the **conditional probability** of event C given event D is

$$P(C|D) = \frac{P(C \cap D)}{P(D)}$$

Definition 1.1.8 Given probability space (S,B,P), $C,D \in B$, C and D are **independent**, written $C \perp D$, if

$$P(C \cap D) = P(C) \cdot P(D)$$

1.2 Random Variables

Definition 1.2.1 Let (S,B,P) be a probability space. The function

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

is a **random variable**, or **rv**, if for arbitrary $x \in \mathbb{R}$,

$$\{w \in S : X(w) \leq x\} \in B.$$

We write the set $\{w \in S : X(w) \leq x\}$ to be $P(x \leq X)$. Note that $P(x \leq X)$ is well-defined for all $x \in \mathbb{R}$.

Definition 1.2.2 The **cumulative distribution function**, or **distribution function**, or **cdf**, of a random variable X on probability space (S,B,P) is

$$\begin{aligned} F : \mathbb{R} &\to \mathbb{R} \\ x &\mapsto P(x \leq X) \end{aligned}$$

such that F is well-defined for all $x \in \mathbb{R}$.

Remark 1.2.3 It follows from Def. 1.2.1 and 1.2.2 that if F is a cdf on (S,B,P) with rv X, then

1. $F(x_1) \leq F(x_2)$ for all $x_1 < x_2, x_1, x_2 \in \mathbb{R}$.
2. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.
3. If $a \in \mathbb{R}$, then
$$\lim_{x \to a^+} F(x) = F(a).$$
4. If $a < b$, then
$$P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$
5. For all $b \in \mathbb{R}$,
$$P(X = b) = F(b) - \lim_{a \to b^-} F(a)$$

1.3 Discrete Random Variables

Definition 1.3.1 If S is a discrete sample space and X is a rv on S, then X is a **discrete random variable**.

Definition 1.3.2 Let X be a discrete rv on (S,B,P) and F be the cdf of X, then the **probability mass function**, or **pmf**, of X is given by

$$\begin{aligned} f : \mathbb{R} &\to [0,1] \\ x &\mapsto P(X=x) \end{aligned}$$

where $P(X=x) = F(x) - \lim_{\varepsilon\to 0^+} F(x-\varepsilon) = F(x) - \lim_{y\to x^-} F(y)$.
The set $A = \{x \in \mathbb{R} : f(x) > 0\}$ is the **support** of X.

Remark 1.3.3 Let f be the pf of discrete rv X on (S,B,P), then
1. $f(x) \leq 0$ for all $x \in \mathbb{R}$.
2. $\sum_{a\in A} f(a) = 1$ where A is the support of X.

■ **Example 1.3.4** Let X with pmf

$$\begin{aligned} f : \mathbb{N} &\to [0,1] \\ x &\mapsto \frac{-(1-p)^x}{x \log p} \end{aligned}$$

for some fixed $p \in (0,1)$. Let A be the support of X. We can verify that $\sum_{a\in A} f(a) = 1$.

Clearly $A = \mathbb{N}$. We have

$$\begin{aligned} \sum_{a\in A} f(a) &= \sum_{i=1}^{\infty} f(i) \\ &= \sum_{i=1}^{\infty} \frac{-(1-p)^i}{i \log p} \\ &= \sum_{i=1}^{\infty} \frac{(-1)(-1)^i (p-1)^i}{i \log p} \\ &= \frac{1}{\log p} \sum_{i=1}^{\infty} \frac{(-1)^{i+1}(p-1)^i}{i} \\ &= \frac{1}{\log p} \cdot \log(p-1+1) \\ &= 1 \end{aligned}$$

■

1.4 Continuous Random Variables

Definition 1.4.1 Let X be a rv on (S,B,P) with cdf F. If F is continuous for all $x \in \mathbb{R}$ and differentiable except at most countably many points on $\mathbb{R}$, then X is a **continuous random variable** on (S,B,P).

Definition 1.4.2 Let X be a continuous rv on (S,B,P) with cdf F. Suppose $x \in \mathbb{R}$ such that F is differentiable at x, then the **probability density function**, or **density function**, or **pdf**, of X at x is

$$f(x) = F'(x).$$

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

Remark 1.4.3 If F is not differentiable at a certain $y \in \mathbb{R}$, we may conveniently assign a value to $f(y)$, as long as $f(y) \geq 0$.

We use the operator Pr interchangeably with P in the remainder of the text.

Remark 1.4.4 let X be a continuous rv with pdf f and cdf F, then

1. $f(x) \geq 0$ for all $x \in \mathbb{R}$.
2. $\int_{-\infty}^{\infty} f(x)dx = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1 - 0 = 1$.
3. $F(x) = \int_{-\infty}^{x} f(t)dt$ for all $x \in \mathbb{R}$.
4. $\Pr(a \leq X \leq b) = \int_a^b f(x)dx$.
5. $\Pr(X = b) = F(b) - \lim_{a\to b^-} F(a) = F(b) - F(b) = 0$ for all $b \in \mathbb{R}$.
6. By (5), we have $\Pr(a \leq X \leq b) = \Pr(a < X \leq b) = \Pr(a < X < b) = \Pr(a \leq X < b)$ for all $a < b \in \mathbb{R}$.

Definition 1.4.5 The **gamma function** is

$$\begin{aligned}\Gamma : (0,\infty) &\to \mathbb{R} \\ \alpha &\mapsto \int_0^\infty y^{\alpha-1}e^{-y}dy\end{aligned}$$

Proposition 1.4.6 Let $\alpha > 1$, $n \in \mathbb{N}$, then

1. $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$.
2. $\Gamma(n) = (n-1)!$.
3. $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

Proof. (1) Using integration by parts, we have

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy = uv|_{v=0}^{\infty} - \int_0^\infty vdu$$

where $u = y^{\alpha-1}$, $dv = e^{-y}dy$, $v = -e^{-y}$, $du = (\alpha-1)y^{\alpha-2}$. It follows that

$$\begin{aligned}\Gamma(\alpha) &= y^{\alpha-1}(-e^{-y})\big|_{y=0}^{\infty} + \int_0^\infty e^{-y}(\alpha-1)y^{\alpha-2}dy \\ &= 0 + (\alpha-1)\Gamma(\alpha-1)\end{aligned}$$

as required.
(2) Let $n = 1$, then $\Gamma(1) = \int_0^\infty y^0 e^{-y}dy = -e^{-y}|_0^\infty = 1$. Suppose inductively that for $n \in \mathbb{N}$, $\Gamma(n) = (n-1)!$, then by part (1), $\Gamma(n+1) = n\Gamma(n) = n\cdot(n-1)! = n!$. This completes the inductive step.
(3) We have $\Gamma\left(\frac{1}{2}\right) = \int_0^\infty y^{-1/2}e^{-y}dy$. Substitute $y = u^2$, then $dy = 2udu$ and

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty u^{-1}2ue^{-u^2}du = 2\int_0^\infty e^{-u^2}du.$$

Now $\Gamma\left(\frac{1}{2}\right)^2 = 4\int_0^\infty\int_0^\infty e^{-u^2-v^2}dudv$. Define $G : \mathbb{R} \to \mathbb{R}$ by $G(u,v) = e^{-u^2-v^2}$. Define a one-to-one mapping $f : [0,\infty)^2 \to [0,\infty)^2$ by $f(u,v) = (r\cos\theta, r\sin\theta)$ where $r = \sqrt{u^2+v^2}$ and $\theta =$

$\arctan(v/u)$. Using the change of variables technique from calculus to get

$$\begin{aligned}
&\int_0^\infty \int_0^\infty e^{-(u^2+v^2)} du dv \\
&= \int_0^{\frac{\pi}{2}} \int_0^\infty e^{-(r^2\cos^2\theta + r^2\sin^2\theta)} \left| \begin{matrix} \frac{\partial u}{\partial r} & \frac{\partial u}{\partial \theta} \\ \frac{\partial v}{\partial r} & \frac{\partial v}{\partial \theta} \end{matrix} \right| dr d\theta \\
&= \int_0^{\frac{\pi}{2}} \int_0^\infty e^{-r^2} \left| \begin{matrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{matrix} \right| dr d\theta \\
&= \int_0^{\frac{\pi}{2}} \int_0^\infty e^{-r^2} r dr d\theta \\
&= \left(\int_0^{\frac{\pi}{2}} d\theta \right) \left(\int_0^\infty e^{-r^2} r dr \right) \\
&= \frac{\pi}{2} \cdot \frac{-1}{2} \int_0^\infty e^u du \\
&= \frac{\pi}{4}
\end{aligned}$$

Now

$$\Gamma\left(\frac{\pi}{2}\right)^2 = 4 \int_0^\infty \int_0^\infty e^{-(u^2+v^2)} du dv = 4 \cdot \frac{\pi}{4} = \pi,$$

and since $e^{-(u^2+v^2)} \geq 0$, $\Gamma\left(\frac{\pi}{2}\right) \geq 0$, so

$$\Gamma\left(\frac{\pi}{2}\right) = \sqrt{\pi}$$

as required. ■

Definition 1.4.7 Fix $\alpha, \beta > 0$ and define continuous rv X with pdf

$$\begin{aligned} f : \mathbb{R} &\to [0,1] \\ x &\mapsto \frac{x^{\alpha-1} e^{-x/\beta}}{\Gamma(\alpha)\beta^\alpha} \mathbf{1}_{\{x>0\}} \end{aligned}$$

then X is said to follow a **Gamma distribution**, or $X \sim \gamma(\alpha, \beta)$.

If $\alpha = 1$, then $f(x) = \frac{1}{\beta} e^{-\frac{x}{\beta}} \mathbf{1}_{\{x>0\}}$ and X is said to follow an **exponential distribution**, or $X \sim \text{Exp}(\beta)$.

If X follows the pdf

$$f(x) = \frac{\beta}{\alpha^\beta} x^{\beta-1} e^{-\left(\frac{x}{\theta}\right)^\beta} \mathbf{1}_{\{x>0\}}$$

then X is said to follow a **Weibull distribution**, or $X \sim \text{Weibull}(\alpha, \beta)$.

■ **Example 1.4.8** Fix $\alpha, \beta > 0$. Suppose $X \sim \gamma(\alpha, \beta)$. We can verify $\int_{-\infty}^\infty f(x)dx = 1$:

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

Take $u = \frac{x}{\beta}$, then $du = \frac{1}{\beta}$ and $x = \beta u$. Continuing:

$$\int_{-\infty}^\infty f(x)dx = \frac{1}{\Gamma(\alpha)\beta^{\alpha-1}} \int_0^\infty \beta^{\alpha-1} u^{\alpha-1} e^{-u} du = \frac{1}{\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} e^{-u} du = \frac{1}{\Gamma(\alpha)} \Gamma(\alpha) = 1$$

So f is a well-defined pdf.
Now suppose $X \sim \text{Weibull}(\alpha, \beta)$. We can perform a similar verification:

$$\int_{-\infty}^{\infty} f(x)dx = \frac{\beta}{\alpha^\beta}\int_0^\infty \alpha^\beta \frac{1}{\beta} e^{-\frac{x^\beta}{\alpha^\beta}} \frac{1}{\alpha^\beta}\beta x^{\beta-1}dx$$

Let $u = \frac{x^\beta}{\alpha^\beta}$, then $du = \frac{\beta}{\alpha^\beta}x^{\beta-1}dx$ and

$$\int_{-\infty}^{\infty} f(x)dx = \int_0^\infty e^{-u}du = 1$$

which is as expected. ∎

Remark 1.4.9 If $\beta = 1$ in a Weibull distribution, then $f(x) = \frac{1}{\alpha}e^{-x/\alpha}$, $x > 0$, $\beta > 0$. This is again the exponential distribution $X \sim \text{Exp}(\alpha)$.

1.5 Functions of a Random Variable

Definition 1.5.1 Let X be a discrete rv with pmf $\Pr := P$ and $Y = h(X)$ be a discrete function of X, then for $y \in \mathbb{R}$

$$\Pr(Y = y) = \sum_{x:h(x)=y} \Pr(X = x)$$

If X is a continuous rv with pdf f, $Y = h(X)$ is a discrete function of X, and $A := \{x \in \mathbb{R} : h(x) = y\}$, then the pmf of Y is

$$\Pr(Y = y) = \int_A f(x)dx$$

■ **Example 1.5.2** Let X be a rv with pmf

$$f : \mathbb{N} \cup \{0\} \to [0,1]$$
$$x \mapsto \frac{e^{-1}}{x!}$$

Suppose Y is another rv with $Y = (X-1)^2$. We can find the pmf of Y. Note that first of all Y only takes values $0, 1, 4, 9, \ldots$, and

$$\Pr(Y = 0) = \Pr(X = 1) = e^{-1}$$
$$\Pr(Y = 1) = \Pr(X = 0) + \Pr(X = 2) = e^{-1} + \frac{e^{-1}}{2}$$
$$\Pr(Y = 4) = \Pr(X = 3) = \frac{e^{-1}}{3!}$$
$$\vdots$$
$$\Pr(Y = k) = \Pr(X = \sqrt{k}+1) = \frac{e^{-1}}{(\sqrt{k}+1)!}$$

Thus the pmf of Y is

$$f(y) = \begin{cases} \frac{e^{-1}}{(1+\sqrt{y})!} & \text{if} \quad y = 0, 4, 9, \ldots \\ \frac{3}{2}e^{-1} & \text{if} \quad y = 1 \\ 0 & \text{if} \quad \text{otherwise} \end{cases}$$

∎

Definition 1.5.3 If X is a continuous rv and $Y = h(X)$ is a continuous function of X, then the cdf of Y is

$$\begin{aligned} F_Y : \mathbb{R} &\to [0,1] \\ y &\mapsto \Pr(Y \leq y) = \Pr(h(x) \leq y) \end{aligned}$$

Remark 1.5.4 We may solve $\Pr(h(x) \leq y)$ with respect to x and calculate the probability $F_Y(y)$ by the cdf of X, then differentiate $F_Y(y)$ with respect to y to get the pdf of Y.

Theorem 1.5.5 Let X be a continuous rv with cdf F, then the rv Y defined by

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

has a Uniform$(0,1)$ distribution.

Proof. Denote $A := \{x \in \mathbb{R} : f(x) > 0\}$ to be the support of X. For $a \in A$, $F(a)$ is an increasing function, so for all $a \in A$, F has an inverse function F^{-1}. Moreover $Y \in [0,1]$ for all X, so we construct the cdf of Y:

$$G(Y) = \Pr(Y \leq y) = \Pr(F(X) \leq y) = \Pr(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y, y \in [0,1]$$

This is the cdf of a uniformly distributed rv, so $Y \sim$ Uniform$(0,1)$ as required. ■

Remark 1.5.6 The formula $Y = F(X) = \int_{-\infty}^{X} f(t)dt$ is known as the **probability integral transformation**. The following theorem establishes the converse of the above theorem.

Theorem 1.5.7 Let Y be a continuous rv and F be the cdf of Y. Then the rv $X = F^{-1}(U)$ has cdf F if $U \sim$ Uniform$(0,1)$.

Proof. Since $U \sim$ Uniform$(0,1)$, we have $\Pr(U \leq u) = u$ for all $u \in (0,1)$. Let A be the support of $X = F^{-1}(U)$, then for $x \in A$,

$$\Pr(X \leq x) = \Pr(F^{-1}(U) \leq x) = \Pr(U \leq F(x)) = F(x)$$

where the last step comes from the cdf of a uniform distribution. This shows that F is the cdf of X ■

■ **Example 1.5.8** From a random sample that follows the uniform distribution, we may extract from it a randome sample of any continuous distribution that has an invertible cdf.

Suppose $U \sim$ Uniform$(0,1)$ and fix $\theta > 0$. We can find a mapping $T : [0,1] \to \mathbb{R}$ such that the rv $T(U) \sim$ Exp(θ). First of all, if $X \sim$ Exp(θ), then X has cdf

$$F_X(x) = 1 - e^{-\frac{x}{\theta}}, x \in \mathbb{R}$$

Write $U = 1 - e^{-\frac{x}{\theta}}$. Re-arranging yields $x = -\theta \log(1-U)$. Define $T(U) = -\theta \log(1-U)$. By Thm. 1.5.7, $T(U)$ has the same cdf as Exp(θ) and $T(U) \sim$ Exp(θ). ■

1.6 Location and Scale Parameters

Definition 1.6.1 Let X be a continuous rv with parameter θ such that $f(x;\theta)$ is the pdf of X. Let $f_0(x) := f(x;\theta = 0)$. θ is called a **location parameter** if

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

Let $f_1(x) := f(x;\theta=1)$, then θ is called a **scale parameter** if

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

■ **Example 1.6.2** Suppose $\beta > 0$ and $\alpha \in \mathbb{R}$. Let X be a continuous rv with pdf

$$\begin{aligned} f : \mathbb{R} &\to [0,1] \\ x &\mapsto \frac{1}{\beta} e^{-(x-\alpha)/\beta} \mathbf{1}_{x \geq \alpha} \end{aligned}$$

Fix $\alpha = \theta$ and $\beta = 1$, then $f(x) = e^{-(x-\theta)}$. Using the above definition we have $f_0(x) = e^{-x}$ and $f_0(x-\theta) = e^{-(x-\theta)} = f(x)$. Hence here θ is a location parameter by definition.

Fix $\alpha = 0$ and $\beta = \theta$, then $f(x) = \frac{1}{\theta} e^{-x/\theta}$. Using the above definition we have $f_1(x) = e^{-x}$ and $\frac{1}{\theta} f_1\left(\frac{x}{\theta}\right) = \frac{1}{\theta} e^{-x/\theta} = f(x)$, so θ here is a scale parameter for X. ■

Remark 1.6.3 The continuous rv in the above example is said to follow a **Double Exponential Distribution** and is written $X \sim \text{DoubleExp.}(\alpha, \beta)$.

Remark 1.6.4 Constructing confidence intervals for location and scale parameters are easier. See Thm. 5.7.9 and its following examples.

1.7 Expectation

Definition 1.7.1 Let X be a discrete rv with pmf $f(x)$ and support A, then the **expectation** or the **mean** of X is

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

provided that $\sum_{a \in A} |a| f(a) < \infty$. If $\sum_{a \in A} |a| f(a)$ diverges, then $E(X)$ does not exist.

If X is a continuous rv with pdf $f(x)$ and support A, then the **expectation** or the **mean** of X is

$$\mu_X = E(X) = \int_A x f(x) dx$$

provided that $\int_A |x| f(x) dx < \infty$. Otherwise $E(X)$ does not exist.

Theorem 1.7.2 Let X be a non-negative continuous rv with cdf F such that $E(X) < \infty$, then

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

Proof. By the definition of cdf, $1-F(x) = \Pr(X \geq x) = \int_x^\infty f(t)dt$ where f is the pdf of X. Hence

$$\begin{aligned}
&\int_0^\infty 1-F(x)dx \\
=&\int_0^\infty \left(\int_x^\infty f(t)dt\right) dx \\
=&\int_0^\infty \left(\int_0^t f(t)dx\right) dt \\
=&\int_0^\infty f(t)dt \cdot \int_0^t dx \\
=&\int_0^\infty f(t)tdt \\
=&E(X)
\end{aligned}$$

as required. The logic from the second line to the third line is that $t \in [x,\infty)$ and $x \in [0,\infty)$ implies $0 \leq x \leq t < \infty$ which in turn implies $x \in [0,t)$ and $t \in [0,\infty)$. ■

■ **Example 1.7.3** Suppose $X \sim \text{Exp}(\theta)$, then $F(x) = 1-e^{-x/\theta}$. X is non-negative by definition of the exponential distribution, so by the above theorem

$$E(X) = \int_0^\infty 1-F(x)dx = \int_0^\infty e^{-x/\theta}dx = \theta.$$

■

Theorem 1.7.4 Suppose X is an rv with probability function f, then for $a,b \in \mathbb{R}$ and g,h real-valued functions,

$$E(aX+b) = aE(X)+b$$

and

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

Proof. This is a direct consequence of the linearity of integrals if X is continuous, and straightforward from the definition of expectation if X is discrete. ■

Definition 1.7.5 Let X be a rv.

The **k-th moment** of X about a real number a is

$$E((X-a)^k).$$

The **k-th factorial moment** of X is

$$E(X^{(k)}) = E(X(X-1)\dots(X-k+1)).$$

The **variance** of X is its second moment about its mean μ_X:

$$\text{Var}(X) = \sigma_X^2 = E((X-\mu)^2).$$

Theorem 1.7.6 Let X be a rv, then

$$\text{Var}(X) = E(X^2)-\mu_X^2 = E(X(X-1))+\mu_X-\mu_X^2 = E(X^{(2)}+\mu_X-\mu_X^2$$

and

$$\text{Var}(aX+b) = a^2 \text{Var}(X), a \in \mathbb{R}.$$

Proof. For convenience, denote $\mu := \mu_X$. By definition of variance

$$\text{Var}(X) = E(X^2 - 2\mu X + \mu^2)$$

By linearity of expectation this equals to

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

as required.

Meanwhile

$$\text{Var}(X) = E(X^2) - \mu^2 = E(X(X-1)+X) - \mu^2 = E(X(X-1)) + \mu - \mu^2$$

as required. Finally

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

as required. ■

■ **Example 1.7.7** Let X follow a **negative binomial distribution** $\text{NB}(k,p)$, $k \in \mathbb{N}$ being the number of successes before the experiment is stopped and $p \in [0,1]$ is the chance of success each trial. Recall that X is a discrete rv following pmf

$$f(x) = \binom{-k}{x} p^k (p-1)^x = \binom{x+k-1}{x} (1-p)^x p^k, x \in \{0,1,\ldots,k\}$$

where

$$\binom{-k}{x} = \frac{(-k)(-k-1)(-k-2)\ldots(-k-x+1)}{x!}$$

and

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

We may compute $E(X^{(j)})$ for $j \in \mathbb{N}$ where $x^{(j)} = x(x-1)\ldots(x-j+1)$. To do this we first recall the identity

$$x^{(j)}\binom{n}{x} = n^{(j)}\binom{n-j}{x-j}.$$

Then

$$\begin{aligned}
E(X^{(j)}) &= \sum_{x=0}^{\infty} x^{(j)} \binom{-k}{x} p^k (p-1)^x \\
&= \sum_{x=j}^{\infty} (-k)^{(j)} \binom{-k-j}{x-j} p^k (p-1)^j (p-1)^{x-j} \\
&= p^k (p-1)^j (-k)^{(j)} \sum_{x=j}^{\infty} \binom{-k-j}{x-j} (p-1)^{x-j} \\
&= p^k (p-1)^j (-k)^{(j)} \sum_{y=0}^{-k-j} \binom{-k-j}{y} (p-1)^y (1)^{-k-j-y} \\
&= p^k (p-1)^j (-k)^{(j)} (p-1+1)^{-k-j} \\
&= \left(\frac{p-1}{p}\right)^j (-k)^{(j)}
\end{aligned}$$

where $(-k)^{(j)} = (-k)(-k-1)\dots(-k-j+1)$. ∎

■ **Example 1.7.8** Let $X \sim \text{Weibull}(\theta, \beta)$, then recall that X has pdf

$$f(x) = \frac{\theta}{\beta^\theta} x^{\theta-1} e^{-\left(\frac{x}{\beta}\right)^\beta} \mathbf{1}_{x>0}$$

The k-th moment for X is therefore

$$E(X^k) = \int_0^\infty x^k \frac{\theta}{\beta^\theta} x^{\theta-1} e^{-\left(\frac{x}{\beta}\right)^\beta} dx.$$

Let $y = \left(\frac{x}{\beta}\right)^\theta$, then $dy = \theta x^{\theta-1} \left(\frac{1}{\beta}\right)^\theta dx$ and $x = \beta y^{\frac{1}{\theta}}$. Consequently

$$E(X^k) = \int_0^\infty \left(\beta y^{\frac{1}{\theta}}\right)^k e^{-y} dy = \beta^k \int_0^\infty y^{\frac{k}{\theta}} e^{-y} dy = \beta^k \Gamma\left(\frac{k}{\theta}+1\right)$$

■

1.8 Inequalities

Remark 1.8.1 Probability inequalities help prove limit theorems of sequences of random variables.

Theorem 1.8.2 — Markov's Inequality. Let X be a rv. Suppose $u(X)$ is a non-negative function such that $E(u(X))$ exists, then for all $c \geq 0$,

$$\Pr(u(X) \geq c) \leq \frac{E(u(X))}{c}$$

Proof. Let $A := \{x \in \mathbb{R} : u(x) \geq c\}$, then

$$E(u(X)) = \int_{-\infty}^{\infty} u(x) f(x) dx = \int_A u(x) f(x) dx + \int_{A^c} u(x) f(x) dx$$

where f is the pdf of X.

Clearly $\int_{A^c} u(x) f(x) dx > 0$ since both $u(x)$ and $f(x)$ are nonnegative. We have

$$\begin{aligned}
E(u(X)) &\geq \int_A u(x) f(x) dx \geq \int_A c f(x) dx = c \int_A f(x) dx \\
&= c \Pr(x \in A) = c \Pr(u(X) \geq c)
\end{aligned}$$

Re-arranging yields

$$\Pr(u(X) \geq c) \leq \frac{E(u(X))}{c}.$$

■

Corollary 1.8.3 Let X be a rv, then for all $k \in \mathbb{N}$,

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

Proof. $x \mapsto |x|^k$ is a nonnegative function, so this follows directly from Thm. 1.8.2. ■

Corollary 1.8.4 — Chebyshev's Inequality. Suppose X is a rv with finite mean μ and finite variance σ^2, then for all $k \geq 0$, we have

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

Proof. The mapping $u : x \mapsto |x - \mu|^2$ is nonnegative and $E(u(X)) = \text{Var}(X) = \sigma^2$, so by Thm. 1.8.2, we have

$$\Pr(u(X) \geq k^2\sigma^2) \leq \frac{E(u(X))}{k^2\sigma^2} = \frac{\sigma^2}{k^2\sigma^2} = \frac{1}{k^2},$$

but $u(X) = |X - \mu|^2 \geq k^2\sigma^2$ if and only if $|X - \mu| \geq k\sigma$, so this yields

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

■

Remark 1.8.5 Chebyshev's Inequality immediately yields

$$\Pr(|X - \mu_X| \geq k) \leq \frac{\sigma_X^2}{k^2}$$

for rv X. This formulation is perhaps the most well-known.

■ **Example 1.8.6** A post office handles 10000 letters on average per day. What is the probability that it handles 15000 letters tomorrow?

We employ Thm. 1.8.2 directly by taking $u(X) = X$ and $c = 15000$ to get

$$\Pr(X \geq 15000) \leq \frac{E(X)}{15000} = \frac{10000}{15000} = \frac{2}{3},$$

So the probability is at most $\frac{2}{3}$. ■

■ **Example 1.8.7** A post office handles 10000 letters per day on average with a variance of 2000. What can be said about $\Pr(8000 \leq X \leq 12000)$? What about $\Pr(X \geq 15000)$?

Let $k = \sqrt{2000}$. Use Chebyshev's Inequality to get

$$\Pr(|X - E(X)| \geq k\sigma_X) = \Pr(|X - 10000| \geq \sqrt{2000}\sqrt{2000}) = \Pr(|X - 10000| \geq 2000) \leq \frac{1}{k^2} = \frac{1}{2000}.$$

Hence $\Pr(8000 \leq X \leq 12000) = \Pr(|X - 10000| \leq 2000) \geq 1 - \frac{1}{2000} = \frac{1999}{2000}$.
For $\Pr(X \geq 15000)$, note that

$$\Pr(X \geq 15000) = \Pr(X - 10000 \geq 5000) \leq \Pr(|X - 10000| \geq 5000).$$

Let $k = \frac{5000}{\sqrt{2000}}$, then $k\sigma = 5000$ and by Chebyshev's Inequality we have

$$\Pr(|X - 10000| \geq 5000) \leq \frac{1}{k^2} = \frac{2000}{5000^2}.$$

Hence $\Pr(X \geq 15000) \leq \frac{2}{25000}$. ∎

1.9 Moment Generating Functions

Remark 1.9.1 Moment generating functions provide a way other than pdf or cdf to uniquely identify the distribution of a random variable. They are closely related to Laplace and Fourier transforms.

Definition 1.9.2 Let X be a rv. The **moment generating function**, or **MGF**, of X is

$$M_X : t \mapsto E(e^{tX})$$

if this expectation exists for $t \in (-h, h)$ for some fixed $h > 0$. h is called the **radius of convegence** for $M_X(t)$.

Remark 1.9.3 Immediately we note that $M_X(0) = E(e^0) = E(1) = 1$.

Due to the fact that expectation are not guaranteed to exist, not all random variables have a moment generating function.

Proposition 1.9.4 Let X be a rv and suppose its MGF $M_X(t)$ exists on some interval $(-h, h)$, $h > 0$, then the k-th derivative of M_X with respect to t evaluated at $t = 0$ is the k-th moment of X, i.e.

$$M_X^{(k)}(t)|_{t=0} = \left.\frac{d^k M_X(t)}{dt^k}\right|_{t=0} = E(X^k).$$

Proof. We have $M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)dx$ where f is the cdf of X. Consequently

$$\frac{d^k}{dt^k} \int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{-\infty}^{\infty} \frac{d^k}{dt^k} e^{tx} f(x)dx$$

since the integrand is absolutely convergent by assumption. Now $\frac{d^k}{dt^k} e^{tx} = x^k e^{tx}$, so

$$\frac{d^k}{dt^k} M_X(t) = \int_{-\infty}^{\infty} x^k e^{tx} f(x)dx = E(X^k e^{tx}).$$

Evaluated at $t = 0$, $\frac{d^k}{dt^k} M_X(0) = E(X^k)$ as required. ∎

Corollary 1.9.5 Suppose X is a rv with well-defined MGF $M_X(t)$ on some interval $(-h, h)$, then the MacLaurin series of $M_X(t)$ is

$$M_X(t) = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!} t^k.$$

Proof. By definition of MacLaurin series

$$M_X(t) = \sum_{k=0}^{\infty} \frac{M_X^{(k)}(t)|_{t=0}}{k!} t^k$$

where $M_X^{(k)}(t)|_{t=0} = E(X^k)$ by the above proposition, so

$$M_X(t) = \sum_{k=0}^{\infty} \frac{E(X^k)}{k!} t^k$$

as required. ■

■ **Example 1.9.6** The above proofs are done as if X is continuous, but suppose X is a discrete rv with pmf

$$f(x) = \left(\frac{1}{2}\right)^{x+1} \mathbf{1}_{x \in \mathbb{N} \cup \{0\}},$$

the operations with its MGF are analogous to the continuous case. To see this, observe that

$$M_X(t) = E(e^{tX}) = \sum_{x=0}^{\infty} e^{tx} \left(\frac{1}{2}\right)^{x+1} = \frac{1}{2} \sum_{x=0}^{\infty} \left(\frac{1}{2} e^t\right)^x$$

which converges to $\frac{1}{2}\left(\frac{1}{1-\frac{1}{2}e^t}\right)$ if and only if $\left|\frac{1}{2}e^t\right| = \frac{1}{2}e^t < 1$ if and only if $t < \log 2$. Thus the MGF of X is

$$\begin{aligned} M_X : (-\log 2, \log 2) &\to \mathbb{R} \\ t &\mapsto \frac{1}{2 - e^t}. \end{aligned}$$

With the explicit formula for X's MGF we can easily find its expectation and variance via Proposition 1.9.4:

$$E(X) = M_X'(t)|_{t=0} = 1, E(X^2) = M_X^{(2)}(t)|_{t=0} = 3,$$

and $\text{Var}(X) = E(X^2) - E(X)^2 = 3 - 1 = 2$. ■

Theorem 1.9.7 Let X be a rv with MGF $M_X(t)$, $t \in (-h, h)$. Let $Y = aX + b$ where $a, b \in \mathbb{R}$, $a \neq 0$, then the MGF for Y is

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

Proof. By definition of MGF,

$$M_Y(t) = E(e^{taX+tb}) = e^{tb} E(e^{taX}) = e^{tb} M_X(at), |at| < h$$

where $|at| < h$ if and only if $|t| < \frac{h}{|a|}$. This completes the proof. ■

Theorem 1.9.8 Suppose X, Y are rv's, and that they have MGF's M_X and M_Y respectively. If $M_X(t) = M_Y(t)$ for all values of t, then X and Y follow the same distribution.

Proof. We omit the proof. ■

■ **Example 1.9.9** The MGF of well-known distributions are documented in formula sheets. Suppose $X \sim N(\mu, \sigma^2)$ and $Y \sim \chi_1^2$, i.e. X follows a normal distribution and Y follows a chi-squared distribution with 1 degree of freedom, then it is known (calculated) that

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}, t \in \mathbb{R}$$

and

$$M_Y(t) = \frac{1}{\sqrt{1-2t}}, |t| < \frac{1}{2}.$$

We can show that $Y_0 := \left(\frac{X-\mu}{\sigma}\right)^2 \sim \chi_1^2$. To do this first define $Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$. This is a linear combination of X. Use Thm. 1.9.7 to get

$$M_Z(t) = e^{-\frac{\mu}{\sigma}t} M_X\left(\frac{1}{\sigma}t\right) = e^{\frac{t^2}{2}}, t \in \mathbb{R}.$$

This shows that $Z \sim N(0,1)$. Now $Y_0 = Z^2$, so

$$\begin{aligned}
M_Y(t) = E(e^{tY}) = E(e^{tZ^2}) &= \int_{-\infty}^{\infty} e^{tz^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{\left(t-\frac{1}{2}\right)z^2} dz \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{z^2}{\left(\frac{1}{\sqrt{1-2t}}\right)^2}\right)\right) dz \\
&= \frac{1}{\sqrt{1-2t}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\frac{1}{\sqrt{1-2t}}} \exp\left(-\frac{1}{2}\left(\frac{z^2}{\left(\frac{1}{\sqrt{1-2t}}\right)^2}\right)\right) dz \\
&= \frac{1}{\sqrt{1-2t}}, t < \frac{1}{2}
\end{aligned}$$

where $\frac{1}{\sqrt{1-2t}}$ is the MGF of χ_1^2. Hence $Y_0 \sim \chi_1^2$ by Thm. 1.9.8.

Note that the integral in the second-last line above is simply $\int_{-\infty}^{\infty} g(z)dz$ where g is the pdf of the $N\left(0, \frac{1}{1-2t}\right)$ distribution. ■

2. Multi-Variate Random Variables

2.1 Joint and Marginal Cumulative Distribution Functions

Definition 2.1.1 Suppose X and Y are random variables defined on sample space S, then (X,Y) forms a (bivariate) **random vector** and it has **joint cumulative distribution function**

$$\begin{aligned} F &: \mathbb{R}^2 \to [0,1] \\ &(x,y) \mapsto \Pr(X \leq x, Y \leq y). \end{aligned}$$

Remark 2.1.2 We may generalise the above definition to random vectors of dimention $k > 2$. The cdf F has the following properties.

1. F is non-decreasing if x stays fixed and $y \to \infty$, or if y stays fixed and $x \to \infty$.
2. $\lim_{x\to-\infty} F(x,y) = 0 = \lim_{y\to-\infty} F(x,y)$.
3. $\lim_{(x,y)\to(-\infty,-\infty)} F(x,y) = 0$ and $\lim_{(x,y)\to(\infty,\infty)} F(x,y) = 1$.

Definition 2.1.3 Let X,Y be rv's on sample space S with joint cdf F. The **marginal cumulative distribution function** of X is given by

$$F_X(x) = \lim_{y\to\infty} F(x,y) = \Pr(X \leq x), x \in \mathbb{R}$$

and similarly the marginal cdf for Y is

$$F_Y(y) = \lim_{x\to\infty} F(x,y) = \Pr(Y \leq y), y \in \mathbb{R}.$$

Remark 2.1.4 The above definitions work for both discrete and continuous rv's.

2.2 Bivariate Discrete Distributions

Definition 2.2.1 Suppose X,Y are discrete rv's on sample space S. The **joint probability mass**

function of (X,Y) is

$$f : \mathbb{R}^2 \to [0,1]$$
$$(x,y) \mapsto \Pr(X = x, Y = y).$$

and $A := \{(x,y) \in \mathbb{R}^2 : f(x,y) > 0\}$ is the **support** of (X,Y).

Remark 2.2.2 A few immediate properties of joint pmf's:

1. $f(x,y) \geq 0$ for all $(x,y) \in \mathbb{R}^2$.
2. $\sum_{(x,y)\in A} f(x,y) = 1$.
3. For all $R \subseteq \mathbb{R}^2$,

$$\Pr((X,Y) \in R) = \sum_{(x,y)\in R} f(x,y).$$

Definition 2.2.3 Let X,Y be discrete rv's with joint pmf $f : \mathbb{R}^2 \to [0,1]$, then the **marginal probability function** of X is given by

$$f_X(x) = \Pr(X = x) = \sum_y f(x,y), x \in \mathbb{R}$$

while the **marginal probability function** of Y is given by

$$f_Y(y) = \Pr(Y = y) = \sum_x f(x,y), y \in \mathbb{R}.$$

■ **Example 2.2.4 — Hardy-Weinberg Law.** Under certain conditions, the relative frequencies with which three genotypes AA, Aa, and aa occur in the population will be θ^2, $2\theta(1-\theta)$ and $(1-\theta)^2$ respectively for some fixed $\theta \in (0,1)$. Suppose n members of the population are selected at random.

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

1. The joint pmf of (X,Y) is

$$f(x,y) = \frac{n!}{x!y!(n-x-y)!}(\theta^2)^x(2\theta(1-\theta))^y((1-\theta)^2)^{n-x-y}$$

   where $x+y \leq n$, $x,y \geq 0$.
2. The marginal probability function of X is

$$\begin{aligned}\sum_y f(x,y) &= \sum_{y=0}^{n-x} f(x,y) \\ &= \frac{n!}{x!(n-x)!}\theta^{2x}\sum_{y=0}^{n-x}\frac{(n-x)!}{y!(n-x-y)!}(2\theta(1-\theta))^y((1-\theta)^2)^{n-x-y} \\ &= \binom{n}{x}\theta^{2x}(2\theta(1-\theta)+(1-\theta)^2)^{n-x} \\ &= \binom{n}{x}(\theta^2)^x(1-\theta^2)^{n-x}, x \leq n\end{aligned}$$

   which is the pmf of the binomial distribution with n trials and success probability θ^2, so $X \sim \text{Bin}(n,\theta^2)$.

3. The marginal pmf of Y is

$$\begin{aligned}\sum_x f(x,y) &= \frac{n!}{y!(n-y)!}\sum_{x=0}^{n-y}\frac{(n-y)!}{x!(n-y-x)!}(\theta^2)^x((1-\theta)^2)^{n-y-x}\\ &= \binom{n}{y}(\theta^2+(1-\theta)^2)^{n-y}(2\theta(1-\theta))^y\\ &= \binom{n}{y}(1-(2\theta-2\theta^2))^{n-y}(2\theta-2\theta^2)^y, y\le n\end{aligned}$$

which is the pmf of the binomial distribution with n trials and success probability $2\theta-2\theta^2$, so $Y\sim \text{Bin}(n,2\theta-2\theta^2)$.

4. Fix $0\le t\le n$, then

$$\begin{aligned}\Pr(X+Y=t) &= \sum_{(x,y):x+y=t} f(x,t-x)\\ &= \sum_{x=0}^{t}\frac{n!}{x!(t-x)!(n-t)!}(\theta^2)^x(2\theta(1-\theta))^{t-x}((1-\theta)^2)^{n-t}\\ &= \binom{n}{t}((1-\theta)^2)^{n-t}\sum_{x=0}^{t}\binom{t}{x}(\theta^2)^x(2\theta(1-\theta))^{t-x}\\ &= \binom{n}{t}((1-\theta)^2)^{n-t}(\theta^2+2\theta(1-\theta))^t\\ &= \binom{n}{t}(1-(2\theta-\theta^2))^{n-t}(2\theta-\theta^2)^t.\end{aligned}$$

Thus $T=X+Y\sim\text{Bin}(n,2\theta-\theta^2)$.

■

2.3 Bivariate Continuous Distributions

Definition 2.3.1 Let X,Y be continuous rv's, with joint cumulative distribution function $F:\mathbb{R}^2\to[0,1]$, then the **joint probability density function** for (X,Y) is

$$\begin{aligned}f:\mathbb{R}^2&\to[0,1]\\ (x,y)&\mapsto\frac{\partial^2}{\partial x\partial y}F(x,y)\end{aligned}$$

if the partial derivative exists and is a continuous function on $\mathbb{R}^2$ except possibly along a finite number of curves.

The set $A:=\{(x,y)\in\mathbb{R}^2: f(x,y)>0\}$ is the **support** of (X,Y).

Remark 2.3.2 Some properties of the joint pdf:

1. $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}f(x,y)dxdy=1$.
2. $f(x,y)\ge 0$ for all $(x,y)\in\mathbb{R}^2$.
3. For any $R\subseteq\mathbb{R}^2$, $\Pr((X,Y)\in R)=\iint_R f(x,y)dxdy$.

Definition 2.3.3 Let (X,Y) be a random vector with joint pdf f. The **marginal probability density function** of X is given by

$$f_X: x\mapsto\int_{-\infty}^{\infty}f(x,y)dy$$

while the **marginal pdf** of Y is given by

$$f_Y : y \mapsto \int_{-\infty}^{\infty} f(x,y)dx.$$

Both f_X and f_Y above map to $\mathbb{R}$ to $[0,1]$.

2.4 Independent Random Variables

Definition 2.4.1 Two random variables X and Y are **independent** if and only if

$$\Pr(X \in A, Y \in B) = \Pr(X \in A)\Pr(Y \in B)$$

for all sets $A, B \subseteq \mathbb{R}$. We write $X \perp Y$.

Proposition 2.4.2 Two rv's X and Y are independent if and only if

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

where F is the joint cdf of (X,Y), F_X is the marginal cdf of X, and F_Y is the marginal cdf of Y.

Proof. Let $(x,y) \in \mathbb{R}^2$. Define $A_X := \{s \in \mathbb{R} : s \leq x\}$ and $A_Y := \{s \in \mathbb{R} : s \leq y\}$, then $X \perp Y$ if and only if

$$\Pr(X \in A_X, Y \in A_Y) = \Pr(X \in A_X)\Pr(Y \in A_Y)$$

Rewrite both sides of the equation as

$$\Pr(X \leq x, Y \leq y) = \Pr(X \leq x)\Pr(Y \leq y)$$

which is equivalent to $F(x,y) = F_X(x)F_Y(y)$, as required. ■

Corollary 2.4.3 Two rv's X and Y are independent if and only if

$$f(x,y) = f_X(x)f_Y(y) \quad \forall (x,y) \in A_1 \times A_2,$$

where f is the joint pdf of (X,Y), f_X is the marginal pdf of X, f_Y is the marginal pdf of Y, $A_1 = \{r \in \mathbb{R} : f_X(r) > 0\}$, and $A_2 = \{r \in \mathbb{R} : f_Y(r) > 0\}$.

Proof. From the proposition above, $X \perp Y$ if and only if $F(x,y) = F_X(x)F_Y(y)$. Differentiate F_X and F_Y we get f_X and f_Y defined on A_1, A_2 respectively. Taking partial derivative for F yields $f(x,y)$. The result follows. ■

Theorem 2.4.4 Let X, Y be continuous rv's with joint pdf $f(x,y)$, A be the support of (X,Y), A_X be the support of X, and A_Y be the support of Y, then $X \perp Y$ if and only if there exist non-negative functions $g, h : \mathbb{R} \to \mathbb{R}^+$ such that

$$f(x,y) = g(x)h(y) \quad \forall (x,y) \in A_X \times A_Y.$$

Proof. ($\Rightarrow$) Suppose $X \perp Y$, then by Corollary 2.4.3 $f(x,y) = f_X(x)f_Y(y)$ for all $(x,y) \in A_X \times A_Y$. Let $g = f_X$ and $h = f_Y$, and we are done.
($\Leftarrow$) Suppose the converse, then for $x \in \mathbb{R}$

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{-\infty}^{\infty} h(x)g(y)dy = h(x)\int_{y \in A_Y} g(y)dy = c_1 h(x)$$

for some $c_1 \in \mathbb{R}^+$.

Similarly, for $y \in \mathbb{R}$, $f_Y(y) = c_2 g(y)$ for some $c_2 \in \mathbb{R}^+$.

Now observe that

$$\begin{aligned} 1 &= \int_{y\in A_Y}\int_{x\in A_X} f_X(x)f_Y(y)dxdy \\ &= \int_{y\in A_Y}\int_{x\in A_X} c_1 h(x) c_2 g(y)dxdy \\ &= c_1c_2\int_{y\in A_Y}\int_{x\in A_X} h(x)g(y)dxdy \\ &= c_1c_2 \end{aligned}$$

Since $c_1c_2 = 1$, so

$$f(x,y) = h(x)g(y) = c_1h(x)c_2g(y) = f_X(x)f_Y(y)$$

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

By Corollary 2.4.3, $X \perp Y$, as required. ■

2.5 Conditional Distributions

Definition 2.5.1 Let X,Y be rv's with joint pdf f, marginal pdf f_X and f_Y, and A be the support of (X,Y). The **conditional probability density function** of X **given** $Y = y$ is

$$\begin{aligned} f_X : \mathbb{R} &\to [0,1] \\ x &\mapsto \frac{f(x,y)}{f_Y(y)} \end{aligned}$$

and is denoted $f_X(x|y)$.

Analagously, the **conditional probability density function** of Y **given** $X = x$ is

$$\begin{aligned} f_Y : \mathbb{R} &\to [0,1] \\ y &\mapsto \frac{f(x,y)}{f_X(x)}. \end{aligned}$$

In both definitions we require $f_Y(y) \neq 0$ and $f_X(x) \neq 0$.

Remark 2.5.2 It follows from Def. 2.5.1 that

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

Remark 2.5.3 It follows from Def. 2.5.1 and Corollary 2.4.3 that $X \perp Y$ if and only if one of the following holds:

1. $f_X(x|y) = f_X(x) \quad \forall x \in A_X$.
2. $f_Y(y|x) = f_Y(y) \quad \forall y \in A_Y$.

2.6 Joint Expectations

Definition 2.6.1 Suppose X,Y are discrete rv's with joint pmf f and support set $A \subseteq \mathbb{Z}^2$. Let $h : \mathbb{Z}^2 \to \mathbb{R}$ be a function, then

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

provided that this sum converges absolutely.

If X,Y are continuous rv's, then

$$E(h(X,Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y)f(x,y)dxdy$$

provided that the integral converges absolutely.

Remark 2.6.2 By the linearity of summations and integrals, if X,Y are rv's with joint pdf $f(x,y)$, $a,b \in \mathbb{R}$, and $h,g : \mathbb{R}^2 \to \mathbb{R}$ are functions, then

$$E(ah(X,Y)+bh(X,Y)) = aE(g(X,Y))+bE(h(X,Y)).$$

Remark 2.6.3 It follows immediately from Remark 2.6.2 that

$$E(aX+bY) = aE(X)+bE(Y),$$

and if $X_1,\dots,X_n$ are random variables, $a_1,\dots,a_n \in \mathbb{R}$ are constants, then

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

Finally, if each $E(X_i) = \mu$, then

$$E(\bar{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

Theorem 2.6.4 If X and Y are independent rv's, and $g,h : \mathbb{R} \to \mathbb{R}$ are functions, then

$$E(g(X)h(Y)) = E(g(X))E(h(Y)).$$

Proof. If $X \perp Y$, then $f(x,y) = f_X(x)f_Y(y)$ by Corollary 2.4.3. Thus

$$\begin{aligned} E(g(X)h(Y)) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x)h(y)f_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} g(x)f_X(x)dx \int_{-\infty}^{\infty} h(y)f_Y(y)dy \\ &= E(g(X))E(h(Y)) \end{aligned}$$

as required. ■

Remark 2.6.5 It follows by an induction argument that, in general, if $X_1,\dots,X_n$ are pairwise independent, and $h_1,\dots,h_n \in \mathbb{R}$, then

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

Definition 2.6.6 Let X,Y be rv's. The **covariance** of X and Y is

$$\mathrm{Cov}(X,Y) = E((X-E(X))(Y-E(Y))).$$

Definition 2.6.7 If X, Y are rv's such that $\text{Cov}(X,Y) = 0$, then X and Y are said to be **uncorrelated**.

Theorem 2.6.8 If X, Y are rv's, then

$$\text{Cov}(X,Y) = E(XY) - E(X)E(Y).$$

Proof. For convenience we denote $\mu_X := E(X)$, $\mu_Y := E(Y)$. We have

$$\begin{aligned}
\text{Cov}(X,Y) &= E((X-\mu_X)(Y-\mu_Y)) \\
&= E(XY - Y\mu_X - X\mu_Y + \mu_X\mu_Y) \\
&= E(XY) - \mu_X E(Y) - \mu_Y E(X) + E(\mu_X\mu_Y) \\
&= E(XY) - \mu_X\mu_Y - \mu_Y\mu_X + \mu_X\mu_Y \\
&= E(XY) - \mu_X\mu_Y
\end{aligned}$$

as required. ■

Corollary 2.6.9 If $X \perp Y$, then $\text{Cov}(X,Y) = 0$.

Proof. If $X \perp Y$, then $E(XY) = E(X)E(Y)$ by Thm. 2.6.4. By Thm. 2.6.8,

$$\text{Cov}(XY) = E(XY) - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

■

Theorem 2.6.10 If X, Y are rv's, and $a, b \in \mathbb{R}$ are constants, then

$$\text{Var}(aX+bY) = a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X,Y).$$

Proof. $aX+bY$ has mean $a\mu_X + b\mu_Y$. Thus by the definition of variance,

$$\begin{aligned}
&\text{Var}(aX+bY) \\
=&E((aX+bY-a\mu_X-b\mu_Y)^2) \\
=&E((a(X-\mu_X)+b(Y-\mu_Y))^2) \\
=&E(a^2(X-\mu_X)^2 + b^2(Y-\mu_Y)^2 + 2ab(X-\mu_X)(Y-\mu_Y)) \\
=&a^2E((X-\mu_X)^2) + b^2E((Y-\mu_Y)^2) + 2abE((X-\mu_X)(Y-\mu_Y)) \\
=&a^2\,\text{Var}(X) + b^2\,\text{Var}(Y) + 2ab\,\text{Cov}(X,Y)
\end{aligned}$$

as required. ■

Remark 2.6.11 It follows by an induction argument that, in general, if $X_1, \ldots, X_n$ are rv's, $a_1, \ldots, a_n \in \mathbb{R}$ are constants, then

$$\text{Var}\left(\sum_{i=1}^n a_i X_i\right) = \sum_{i=1}^n a_i\,\text{Var}(X_i) + \sum_{i\neq j} 2a_i a_j\,\text{Cov}(X_i,X_j)$$

where $\sum_{i\neq j} 2a_i a_j\,\text{Cov}(X_i,X_j)$ are all possible pairwise combinations of the covariances.

Remark 2.6.12 As with the case of Remark 2.6.3 and 2.6.5, if $X_1, \ldots, X_n$ are pairwise independent and $a_1, \ldots, a_n \in \mathbb{R}$ are constants, then

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i \,\mathrm{Var}(X_i).$$

If moreover each $\mathrm{Var}(X_i) = \sigma^2$ for some $\sigma \in \mathbb{R}$, then

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

Definition 2.6.13 The **correlation coefficient** between rv's X and Y is

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

where $\sigma_X = \sqrt{\mathrm{Var}(X)}$, $\sigma_Y = \sqrt{\mathrm{Var}(Y)}$.

Theorem 2.6.14 Let X, Y be rv's, then $|\rho(X,Y)| \leq 1$. If $\rho(X,Y) = 1$, then $Y = aX + b$ for some $a > 0, b \in \mathbb{R}$; if $\rho(X,Y) = -1$, then $Y = aX + b$ for some $a < 0, b \in \mathbb{R}$.

Proof. We omit the proof. ■

2.7 Conditional Expectation

Definition 2.7.1 Let X, Y be rv's and $g : \mathbb{R} \to \mathbb{R}$ be a function. The **conditional expectation** of $g(Y)|X = x$ is

$$E(g(Y)|X = x) = \begin{cases} \sum_y g(y) f_Y(y|X = x) & \text{if} \quad Y|X = x \text{ is discrete} \\ \int_{-\infty}^{\infty} g(y) f_Y(y|X = x) dy & \text{if} \quad Y|X = x \text{ is continuous} \end{cases}$$

Taking g to be the identity function, we get the **conditional mean** of $Y|X = x$ to be

$$E(Y|X = x) = \begin{cases} \sum_y y f_Y(y|X = x) & \text{if} \quad Y|X = x \text{ is discrete} \\ \int_{-\infty}^{\infty} y f_Y(y|X = x) dy & \text{if} \quad Y|X = x \text{ is continuous} \end{cases}$$

$E(Y|X = x)$ is often denoted $\mu_{Y|x}$.

Definition 2.7.2 The **conditional variance** of rv Y given $X = x$ is

$$\mathrm{Var}(Y|X = x) = E((Y - \mu_{Y|x}|X = x)^2).$$

Theorem 2.7.3 If X and Y are independent rv's, and $g, h : \mathbb{R} \to \mathbb{R}$ are functions, then $E(g(X)|Y = y) = E(g(X))$ and $E(h(Y)|X = x) = E(h(Y))$.

Proof. We have

$$\begin{aligned} E(g(X)|Y = y) &= \int_{-\infty}^{\infty} g(x) f_X(x|y) dx \\ &= \int_{-\infty}^{\infty} g(x) f_X(x) dx \text{ by Remark 2.5.3} \\ &= E(g(X)) \end{aligned}$$

as required. The proof for the other conditional expectation of analogous. ■

Definition 2.7.4 Let X, Y be rv's and $g : \mathbb{R} \to \mathbb{R}$ a function, then

$$E(g(X)|Y)$$

is a function of the rv Y, and realises $E(g(X)|Y = y)$ when $Y = y$, while

$$\mathrm{Var}(g(X)|Y)$$

is a function of the rv Y, and realises $\mathrm{Var}(g(X)|Y = y)$ when $Y = y$.

Remark 2.7.5 This means that $E(g(X)|Y)$ and $\mathrm{Var}(g(X)|Y)$ are random variables in themselves.

Theorem 2.7.6 — Law of Total Expectation. Suppose X, Y are rv's, $g : \mathbb{R} \to \mathbb{R}$ a function, then

$$E(E(g(X)|Y)) = E(g(X)).$$

Proof. Since $E(g(X)|Y)$ is a function of Y, we have

$$\begin{aligned}
E(E(g(X)|Y)) &= \int_{-\infty}^{\infty} E(g(X)|y) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} g(x) f_X(x|y) dx \right) f_Y(y) dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) f_X(x|y) f_Y(y) dx dy \\
&= \int_{-\infty}^{\infty} g(x) \left(\int_{-\infty}^{\infty} f_X(x|y) f_Y(y) dy \right) dx \\
&= \int_{-\infty}^{\infty} g(x) f_X(x) dx \\
&= E(g(X))
\end{aligned}$$

as required. ∎

Remark 2.7.7 It follows immediately from Thm. 2.7.6 that $E(E(X|Y)) = E(X)$, which is a more popularly stated version of the Law of Total Expectation.

Theorem 2.7.8 — Law of Total Variance. Let X, Y be rv's, then

$$\mathrm{Var}(X) = E(\mathrm{Var}(X|Y)) + \mathrm{Var}(E(X|Y)).$$

Proof. We have $\mathrm{Var}(X) = E(X^2) - E(X)^2$. Apply Law of Total Expectation to each term to get

$$\mathrm{Var}(X) = E(E(X^2|Y)) - E(E(X|Y))^2.$$

Re-write this as

$$E(E(X^2|Y)) - E(E(X|Y)^2) + E(E(X|Y)^2) - E(E(X|Y))^2$$

and we realise that the latter two terms combine to be $\mathrm{Var}(E(X|Y))$ while the first two terms, due to the linearity of expectation, combine to be $E(E(X^2|Y) - E(X|Y)^2) = E(\mathrm{Var}(X|Y))$. This yields the final result

$$\mathrm{Var}(X) = E(\mathrm{Var}(X|Y)) + \mathrm{Var}(E(X|Y)).$$

∎

Remark 2.7.9 We have assumed that X,Y are continuous in the various theorems above. The proofs are analogous to the continuous case if any one of X,Y are discrete.

We follow a similar procedure in the theorems that follow.

2.8 Joint Moment Generating Functions

Definition 2.8.1 Let X,Y be rv's. The **joint moment generating function** for X and Y is

$$\begin{aligned} M &: \mathbb{R}^2 \to \mathbb{R} \\ (t_1,t_2) &\mapsto E(e^{t_1X+t_2Y}) \end{aligned}$$

provided that the expectation exists for $t_i \in (-h_i,h_i)$, $h_i > 0$, $i = 1,2$.

More generally, if $X_1,\ldots,X_k$ are rv's, then the joint MGF of all X_i's is

$$\begin{aligned} M &: \mathbb{R}^k \to \mathbb{R} \\ (t_1,\ldots,t_k) &\mapsto E\left(\exp\left(\sum_{i=1}^{k} t_iX_i\right)\right) \end{aligned}$$

provided that the expectation exists for $t_i \in (-h_i,h_i)$, $1 \le i \le k$, $h_i > 0$.

The set $\prod_{i=1}^{k}(-h_i,h_i) \subseteq \mathbb{R}^k$ is the **region of convergence**.

Remark 2.8.2 We mostly deal with the case where $k = 2$. Note, however, that for a joint MGF $M : \mathbb{R}^k \to \mathbb{R}$,

$$M(0,\ldots,0,t_j,0,\ldots,0) = E(\exp(t_jX_j)),$$

which is just the MGF for the j-th rv X_j. This is how to get marginal MGF's from a joint MGF.

In this course we mostly deal with the case where $k = 2$.

Theorem 2.8.3 If X,Y are rv's with joint MGF $M : \mathbb{R}^2 \to \mathbb{R}$, then

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

for $j,k \ge 1$.

Proof. We have $M(t_1,t_2) = E(\exp(t_1X+t_2Y)) < \infty$ for $(t_1,t_2) \in (-h_1,h_1)\times(-h_2,h_2) \subseteq \mathbb{R}^2$. This allows us to switch integral signs and differential operators with ease in algebraic manipulations.

We have

$$E(\exp(t_1X+t_2Y)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t_1x+t_2y} f(x,y)\,dx\,dy$$

and

$$\frac{\partial^j}{\partial t_1^j} M(t_1,t_2) = \int_{-\infty}^{\infty} \frac{\partial^j}{\partial t_1^j} e^{t_1x}\left(\int_{-\infty}^{\infty} e^{t_2y} f(x,y)\,dy\right)dx = \int_{-\infty}^{\infty} x^j e^{t_1x}\left(\int_{-\infty}^{\infty} e^{t_2y} f(x,y)\,dy\right)dx$$

Take the k-th derivative with respect to t_2 yields

$$\begin{aligned}
&\frac{\partial^k}{\partial t_2^k}\left(\frac{\partial^j}{\partial t_1^j}M(t_1,t_2)\right)\\
&=\int_{-\infty}^{\infty}\frac{\partial^k}{\partial t_2^k}e^{t_2y}\left(\int_{-\infty}^{\infty}x^j e^{t_1x}f(x,y)dx\right)dy\\
&=\int_{-\infty}^{\infty}y^k x^j\int_{-\infty}^{\infty}e^{t_1x}e^{t_2y}f(x,y)dxdy
\end{aligned}$$

Evaluate this at $t_1=t_2=0$ yields

$$\left.\frac{\partial^{j+k}}{\partial t_1^j\partial t_2^k}M(t_1,t_2)\right|_{t_1=t_2=0}=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^j y^k f(x,y)dxdy=E(X^jY^k).$$

■

Theorem 2.8.4 Suppose X,Y have joint MGF $M:\mathbb{R}^2\to\mathbb{R}$ that is well-defined in on $(-h_1,h_1)\times(-h_2,h_2)$, then $X\perp Y$ if and only if

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

for $(t_1,t_2)\in(-h_1,h_1)\times(-h_2,h_2)$.

Proof. ($\Rightarrow$) Suppose $X\perp Y$, then $f(x,y)=f(x)f(y)$ and

$$\begin{aligned}
M(t_1,t_2)&=E(e^{t_1X}e^{t_2Y})=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{t_1x}e^{t_2y}f(x,y)dxdy\\
&=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}e^{t_1x}e^{t_2y}f(x)f(y)dxdy\\
&=\int_{-\infty}^{\infty}e^{t_1x}f(x)dx\int_{-\infty}^{\infty}e^{t_2y}f(y)dy\\
&=E(t_1X)E(t_2Y)\\
&=M_X(t_1)M_Y(t_2)
\end{aligned}$$

as required.

($\Leftarrow$) Straightforward from the derivation above. ■

■ **Example 2.8.5** Let X,Y be discrete rv's with the following joint distribution function:

Pr	$X=-1$	$X=1$
$Y=1$	0.5	0.3
$Y=2$	0.1	0.1

The joint MGF is

$$\begin{aligned}
M(t_1,t_2)&=\sum_x\sum_y E(e^{t_1x+t_2y})=\sum_{x=-1,1}\sum_{y=1,2}E(e^{t_1x+t_2y})\\
&=\sum_{x=-1,1}e^{t_1x+t_2}0.5+e^{t_1x+2t_2}0.1\\
&=0.5e^{-t_1+t_2}+0.1e^{-t_1+2t_2}+0.3e^{t_1+t_2}+0.1e^{t_1+2t_2}
\end{aligned}$$

which converges for all $t_1, t_2 \in \mathbb{R}$.

By Thm. 2.8.3 we have

$$E(XY) = \frac{\partial^2}{\partial t_1 \partial t_2} M(t_1, t_2) = -0.5e^{t_1+t_2} - 0.2e^{-t_1+2t_2} + 0.3e^{t_1+t_2} + 0.2e^{t_1+2t_2}$$

evaluated at $t_1 = t_2 = 0$. Hence $E(XY) = -0.7 + 0.5 = -0.2$.

Finally, the marginal MGF for X is

$$M_X(t) = M(t, 0) = 0.6e^{-t} + 0.4e^{t}, t \in \mathbb{R}$$

while the marginal MGF for Y is

$$M_Y(t) = M(0, t) = 0.2e^{2t} + 0.8e^{t}, t \in \mathbb{R}.$$

■

2.9 Multinomial Distribution

Definition 2.9.1 Fix $n \in \mathbb{N}$. Suppose $X_1, \ldots, X_k$ are discrete rv's, $k \geq 1$, and they have joint probability mass function

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

where $\sum_{i=1}^k x_i = n$, each $0 \leq x_i \leq n$ for $i = 1, \ldots, k$, $\sum_{i=1}^k p_i = 1$, each $p_i \in [0, 1]$ for $i = 1, \ldots, k$, then $(X_1, \ldots, X_k)$ are said to follow a **multinomial distribution** and we write

$$X_1, \ldots, X_k \sim \text{MUL}(n; p_1, \ldots, p_k).$$

Remark 2.9.2 Note that in the definition above, the value of x_k depends completely on $x_1, \ldots, x_{k-1}$, and similarly p_k completely depends on $p_1, \ldots, p_{k-1}$. This means that we can write the joint pmf for $X_1, \ldots, X_k$ as

$$f(x_1, \ldots, x_{k-1}) = \frac{n!}{\prod_{i=1}^{k-1}(x_i!)\left(1 - \sum_{i=1}^{k-1} x_i\right)!} \prod_{i=1}^{k-1} p_i^{x_i} \left(1 - \sum_{i=1}^{k-1} p_i\right)^{1-\sum_{i=1}^{k-1} x_i}.$$

Theorem 2.9.3 Suppose $X_1, \ldots, X_k \sim \text{MUL}(n; p_1, \ldots, p_k)$, then $X_1, \ldots, X_{k-1}$ has joint MGF

$$M(t_1, \ldots, t_{k-1}) = (p_1 e^{t_1} + \cdots + p_{k-1} e^{t_{k-1}} + p_k)^n, t_1, \ldots, t_{k-1} \in \mathbb{R}$$

where $p_k = 1 - \sum_{i=1}^{i-1} p_i$.

Proof. By the definition of MGF we have

$$\begin{aligned} M(t_1, \ldots, t_{k-1}) &= \sum_{x_1} e^{t_1 x_1} f(x_1, \ldots x_k) \sum_{x_2} f(x_1, \ldots, x_k) \ldots \sum_{x_{k-1}} e^{t_{k-1} x_{k-1}} f(x_1, \ldots, x_k) \\ &= \sum_{x_1} \sum_{x_2} \ldots \sum_{x_{k-1}} e^{t_1 x_1} \ldots e^{t_{k-1} x_{k-1}} f(x_1, \ldots, x_k) \\ &= \sum_{x_1} \ldots \sum_{x_{k-1}} e^{t_1 x_1} \ldots e^{t_{k-1} x_{k-1}} \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_{k-1}^{x_{k-1}} p_k^{x_k} \end{aligned}$$

where $\sum_{i=1}^{k} x_k = n$ and $\sum_{i=1}^{k} p_k = 1$. This evaluates to

$$\sum_{\sum_{i=1}^{k} x_i = n} (e^{t_1} p_1)^{x_1} \dots (e^{t_{k-1}} p_{k-1})^{x_{k-1}} p_k^{x_k} \frac{n!}{x_1! \dots x_k!}.$$

By the multinomial expansion theorem (a generalisation of the binomial expansion theorem), the above evaluates to

$$(e^{t_1} p_1 + \dots + e^{t_{k-1}} p_{k-1} + p_k)^n$$

as required. ■

Corollary 2.9.4 Suppose $X_1, \dots, X_k \sim \text{MUL}(n; p_1, \dots, p_k)$, then any subset of $\{X_1, \dots, X_k\}$ also have a multinomial distribution, i.e. if $\{X_{k_1}, \dots, X_{k_m}\} \subseteq \{X_1, \dots, X_k\}$, then

$$X_{k_1}, \dots, X_{k_m} \sim \text{MUL}(n; p_{k_1}, \dots, p_{k_m})$$

where $\sum_{i=1}^{m} X_{k_i} = n$ and $\sum_{i=1}^{m} p_{k_i} = 1$.
In particular, for each i, $1 \leq i \leq k$, $X_i \sim \text{Bin}(n; p_i)$.

Proof. The MGF of $X_{k_1}, \dots, X_{k_{m-1}}$ is

$$\begin{aligned} M(t_{k_1}, \dots, t_{k_{m-1}}) &= \sum_{\sum_{i=1}^{m} x_{k_i} = n} (e^{t_{k_1}} p_{k_1})^{x_{k_1}} \dots (e^{t_{k_{m-1}}} p_{k_{m-1}})^{x_{k_{m-1}}} \frac{n!}{x_{k_1}! \dots x_{k_m}!} p_{k_m}^{x_{k_m}} \\ &= (e^{t_{k_1}} p_{k_1} + \dots + e^{t_{k_{m-1}}} p_{k_{m-1}} + p_{k_m})^n \end{aligned}$$

by analogous reasoning to Thm. 2.9.3. By the uniqueness of MGFs and Thm. 2.9.3 we have

$$(X_{k_1}, \dots, X_{k_m}) \sim \text{MUL}(n; p_{k_1}, \dots, p_{k_m}).$$

It follows immediately that for each i, $1 \leq i \leq k$,

$$X_i \sim \text{Bin}(n; p_1).$$

■

Corollary 2.9.5 Suppose $X_1, \dots, X_k \sim \text{MUL}(n; p_1, \dots, p_k)$. Set $1 \leq i < j \leq k$ and define $T = X_i + X_j$, then $T \sim \text{Bin}(n; p_i + p_j)$.

Proof. By Thm. 2.9.3, The MGF of (X_i, X_j) is

$$M(0, \dots, 0, t, 0, \dots, 0, t, 0, \dots, 0) = (e^t p_i + e^t p_j + (1 - p_i - p_j))^n$$

where $M : \mathbb{R}^k \to \mathbb{R}$ has the i-th and j-th arguments to be t and 0 otherwise. Hence the MGF of T is

$$E(e^{tT}) = (e^t (p_i + p_j) + (1 - p_i - p_j))^n, t \in \mathbb{R}$$

This is the MGF of the $\text{Bin}(n; p_i + p_j)$ distribution. By the uniqueness of MGFs, $T \sim \text{Bin}(n; p_i + p_j)$. ■

Corollary 2.9.6 Let $X_1,\dots,X_k \sim \text{MUL}(n;p_1,\dots,p_k)$. Set $1 \le i < j \le k$, then $\text{Cov}(X_i,X_j) = -np_ip_j$.

Proof. We use Thm. 2.8.3 to get

$$E(X_iX_j) = \frac{\partial^2}{\partial t_1 \partial t_2} M(t_1,t_2)\Big|_{t_1=t_2=0}$$

where $M(t_1,t_2) = E(e^{t_1x_i} + e^{t_2x_j})$. By Corollary. 2.9.4 we have

$$M(t_1,t_2) = (e^{t_1}p_i + e^{t_2}p_j + (1-p_i-p_j))^n$$

and therefore

$$\frac{\partial^2}{\partial t_1 \partial t_2} M(t_1,t_2) = n(n-1)(e^{t_1}p_i + e^{t_2}p_j + (1-p_i-p_j))^{n-2} p_i e^{t_1} p_j e^{t_2}$$

which, at $t_1 = t_2 = 0$, equals to $n(n-1)p_ip_j$. Now

$$\text{Cov}(X_i,X_j) = E(X_iX_j) - E(X_i)E(X_j) = n(n-1)p_ip_j - np_inp_j = -np_ip_j$$

since $X_i \sim \text{Bin}(n;p_i)$, $X_j \sim \text{Bin}(n;p_j)$, and so $E(X_i) = np_i$, $E(X_j) = np_j$. ■

Corollary 2.9.7 Suppose $X_1,\dots,X_k \sim \text{MUL}(n;p_1,\dots,p_k)$, then for $0 < m < k$,

$$X_1,\dots,X_m | (X_{m+1},\dots,X_k) = (x_{m+1},\dots,x_k) \sim \text{MUL}\left(n - \sum_{i=m+1}^{k} x_i; \frac{p_1}{\sum_{i=1}^m p_i},\dots,\frac{p_m}{\sum_{i=1}^m p_i}\right).$$

In particular, if $1 \le i < j \le k$, then

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

Proof. We omit the proof. ■

Corollary 2.9.8 Using the same rv's as Corollary 2.9.7, the conditional distribution of X_i given $T := X_i + X_j = t$ is

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

Proof. Using the definition of conditional distribution,

$$f_{X_i|T}(x_i|x_i+x_j) = \frac{f_{X_i,T}(x_i,x_i+x_j)}{f_T(x_i+x_j)}$$

where $f_{X_i,T}(x_i,x_i+x_j) = \Pr(X_i = x_i, T = x_i+x_j) = \Pr(X_i = x_i, X_j = x_j)$. By Corollary 2.9.4,

$$\Pr(X_i = x_i, X_j = x_j) = \frac{n!}{x_i!x_j!} p_i^{x_i} p_j^{x_j}$$

since $X_i, X_j \sim \text{MUL}(n;p_i,p_j)$. Moreover, by Corollary 2.9.5,

$$f_T(x_i+x_j) = f_T(t) = \frac{n!}{(x_i+x_j)!(n-x_i-x_j)!}(p_i+p_j)^{x_i+x_j}(1-p_i-p_j)^{n-x_i-x_j}.$$

Note however that since $x_i + x_j = n$, we have

$$f_T(x_i + x_j) = \frac{n!}{(x_i + x_j)!}(p_i + p_j)^{x_i + x_j} = (p_i + p_j)^{x_i + x_j}.$$

Hence

$$f_{X_i|T}(x_i|x_i + x_j) = \frac{(x_i + x_j)!}{x_i! x_j!}\left(\frac{p_i}{p_i + p_j}\right)^{x_i}\left(\frac{p_j}{p_i + p_j}\right)^{x_j}$$

which shows that

$$X_i|X_i = x_i, X_j = x_j \sim \text{Bin}\left(x_i + x_j; \frac{p_i}{p_i + p_j}\right).$$

Consequently,

$$X_i|T = t \sim \text{Bin}\left(t; \frac{p_i}{p_i + p_j}\right).$$

■

2.10 Bivariate Normal Distribution

Definition 2.10.1 Suppose (X_1, X_2) is a random vector with joint probability density function

$$\begin{aligned} f &: \mathbb{R}^2 \to [0,1] \\ x = (x_1, x_2) &\mapsto \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right) \end{aligned}$$

where

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \sigma_1, \sigma_2 \in \mathbb{R}, \rho \in [0,1]$$

such that Σ is non-singular (invertible), then $X := (X_1, X_2)$ is said to have a **bivariate normal distribution**, written as $\text{BVN}(\mu, \Sigma)$.

μ is known as the **mean vector** while the 2×2 matrix Σ is known as the **variance-covariance matrix**.

Remark 2.10.2 A more general definition is possible with $n \times n$ a covariance-variance matrix and a mean vector in $\mathbb{R}^n$. Such a definition is for **multivariate normal distribution**. However, in this course we restrict the discussion to the bivariate case.

Recall that for $\Sigma \in \mathbb{R}_{2\times 2}$, as defined above, we have

$$|\Sigma| = \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2$$

and so

$$\Sigma^{-1} = \frac{1}{|\Sigma|}\begin{bmatrix} \sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2 \end{bmatrix} = \frac{1}{1-\rho^2}\begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix}.$$

Thus

$$\begin{aligned}
&(x-\mu)^T\Sigma^{-1}(x-\mu)\\
=&\frac{1}{1-\rho^2}\begin{bmatrix}x_1-\mu_1 & x_2-\mu_2\end{bmatrix}\begin{bmatrix}\frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma^2}\\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}\\
=&\frac{1}{1-\rho^2}\begin{bmatrix}\frac{x_1-\mu_1}{\sigma_1^2}-\frac{(x_2-\mu_2)\rho}{\sigma_1\sigma_2} & \frac{-(x_1-\mu_1)\rho}{\sigma_1\sigma_2}+\frac{x_2-\mu_2}{\sigma_2^2}\end{bmatrix}\begin{bmatrix}x_1-\mu_1\\ x_2-\mu_2\end{bmatrix}\\
=&\frac{1}{1-\rho^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right).
\end{aligned}$$

With this we can write a non-matrix version of the joint pdf of the BVN distribution:

$$\begin{aligned}
&f:\mathbb{R}^2\to[0,1]\\
&(x_1,x_2)\mapsto\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left(\frac{-1}{2(1-\rho)^2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right)\right).
\end{aligned}$$

Lemma 2.10.3 Suppose $t=\begin{bmatrix}t_1 & t_2\end{bmatrix}^T\in\mathbb{R}^2$, $x=\begin{bmatrix}x_1 & x_2\end{bmatrix}^T\in\mathbb{R}^2$, $\mu=\begin{bmatrix}\mu_1 & \mu_2\end{bmatrix}^T\in\mathbb{R}^2$, $\Sigma\in\mathbb{R}_{2\times2}$ is symmetric, then

$$(x-\mu)^T\Sigma^{-1}(x-\mu)-2t^Tx=(x-(\mu+\Sigma t))^T\Sigma^{-1}(x-(\mu+\Sigma t))-2\mu^Tt-t^T\Sigma t.$$

Proof. Re-write the RHS as

$$\begin{aligned}
&((x-\mu)-\Sigma t)^T\Sigma^{-1}((x-\mu)-\Sigma t)-2\mu^Tt-t^T\Sigma t\\
=&(x-\mu)^T\Sigma^{-1}(x-\mu)-(x-\mu)^T\Sigma^{-1}\Sigma t-t^T\Sigma^T\Sigma^{-1}(x-\mu)+t^T\Sigma^T\Sigma^{-1}\Sigma t-2\mu^Tt+t^T\Sigma t\\
=&(x-\mu)^T\Sigma^{-1}(x-\mu)-(x-\mu)^Tt-t^T(x-\mu)+t^T\Sigma t-2\mu^Tt-t^T\Sigma t\\
=&(x-\mu)^T\Sigma^{-1}(x-\mu)-x^Tt+\mu^Tt-t^Tx+t^T\mu-2\mu^Tt\\
=&(x-\mu)^T\Sigma^{-1}(x-\mu)-x^Tt-\mu^Tt-t^Tx+t^T\mu.
\end{aligned}$$

However, x^Tt, μ^Tt, t^Tx, and $t^T\mu$ are all real numbers, and so $\mu^Tt=t^T\mu$, and $x^Tt=t^Tx$. Substituting this into the above yields

$$RHS=(x-\mu)^T\Sigma^{-1}(x-\mu)-2t^Tx=LHS$$

as required. ■

Proposition 2.10.4 Suppose $X\sim\text{BVN}(\mu,\Sigma)$, then X has MGF

$$M(t_1,t_2)=\exp\left(\mu^Tt+\frac{1}{2}t^T\Sigma t\right),t=[t_1,t_2]^T\in\mathbb{R}^2.$$

Proof. By definition

$$\begin{aligned}
M(t_1,t_2) &= E(\exp(t_1X_1+t_2X_2)) \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t^Tx} f(x_1,x_2)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{t^Tx} \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}((x-\mu)^T\Sigma^{-1}(x-\mu)-2t^Tx)\right)dx_1dx_2 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}((x-(\mu+\Sigma t))^T\Sigma^{-1}(x-(\mu+\Sigma t))-2\mu^Tt-t^T\Sigma t)\right)dx_1dx_2 \\
&= \exp\left(\mu^Tt+\frac{1}{2}t^T\Sigma t\right)\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-(\mu+\Sigma t))^T\Sigma^{-1}(x-(\mu+\Sigma t))\right)dx_1dx_2 \\
&= \exp\left(\mu^Tt+\frac{1}{2}t^T\Sigma t\right), t\in\mathbb{R}^2
\end{aligned}$$

as required. Note that the third to the fourth line uses Lemma 2.10.3. ■

Corollary 2.10.5 If $X = \binom{X_1}{X_2} \sim \text{BVN}(\mu,\Sigma)$, then $X_1 \sim N(\mu_1,\sigma_1^2)$ and $X_2 \sim N(\mu_2,\sigma_2^2)$.

Proof. Use Remark 2.8.2 and Proposition 2.10.4 to get the marginal MGF of X_1:

$$\begin{aligned}
M(t_1,0) &= \exp\left(\mu_1t_1+\frac{1}{2}\begin{bmatrix}t_1 & 0\end{bmatrix}\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}\begin{bmatrix}t_1\\0\end{bmatrix}\right) \\
&= \exp\left(\mu_1t_1+\frac{1}{2}t_1^2\sigma_1^2\right), t_1\in\mathbb{R}
\end{aligned}$$

which is the MGF of the $N(\mu_1,\sigma_1^2)$ distribution. By the uniqueness of the MGF of distributions, $X_1 \sim N(\mu_1,\sigma_1^2)$. By symmetry, we also get $X_2 \sim N(\mu_2,\sigma_2^2)$. ■

Corollary 2.10.6 If $X = \binom{X_1}{X_2} \sim \text{BVN}(\mu,\Sigma)$, then $\text{Cov}(X_1,X_2) = \rho\sigma_1\sigma_2$.

Proof. Differentiate $M(t_1,t_2)$ in Proposition 2.10.4 to get

$$E(X_1X_2) = \frac{\partial^2}{\partial t_1\partial t_2}M(t_1,t_2)\bigg|_{t_1=t_2=0} = \rho\sigma_1\sigma_2+\mu_1\mu_2.$$

From Corollary 2.10.5, $E(X_1)=\mu_1$, $E(X_2)=\mu_2$, so

$$\text{Cov}(X_1,X_2) = \rho\sigma_1\sigma_2+\mu_1\mu_2-\mu_1\mu_2 = \rho\sigma_1\sigma_2.$$

■

Remark 2.10.7 It is clear now why the notations ρ,σ_1,σ_2 are used in the definition of the variance-covariance matrix. ρ is the correlation between X_1 and X_2, and σ_1,σ_2 are the standard deviations of X_1, X_2 respectively.

Corollary 2.10.8 If $X = \binom{X_1}{X_2} \sim \text{BVN}(\mu, \Sigma)$, then $X_1 \perp X_2$ if and only if the correlation $\rho_{X_1X_2} = 0$.

Proof. ($\Rightarrow$) If $X_1 \perp X_2$, then by Corollary 2.6.9 $\text{Cov}(X_1, X_2) = 0$, and so

$$\rho_{X_1X_2} = \frac{\text{Cov}(X_1, X_2)}{\sigma_1 \sigma_2} = 0.$$

($\Leftarrow$) Suppose $\rho_{X_1X_2} = 0$, then X_1 has MGF

$$M(t_1) = \exp\left(\mu_1 t_1 + \frac{1}{2} t_1^2 \sigma_1^2\right), t_1 \in \mathbb{R}$$

and X_1 has MGF

$$M(t_2) = \exp\left(\mu_2 t_2 + \frac{1}{2} t_2^2 \sigma_2^2\right), t_2 \in \mathbb{R}.$$

Note that since $\rho = 0$, the covariance matrix is $\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$ and the joint MGF for $X = \binom{X_1}{X_2}$ is

$$M(t_1, t_2) = \exp\left(\mu_1 t_1 + \mu_2 t_2 + \frac{1}{2} t_1^2 \sigma_1^2 + \frac{1}{2} t_2^2 \sigma_2^2\right) = M(t_1) M(t_2).$$

Hence $X_1 \perp X_2$ by Thm. 2.8.4. ■

Remark 2.10.9 For multi-dimensional random vectors in general, independence implies zero correlation, but zero correlation does not, in general, imply independence.

Corollary 2.10.10 Let $X = \binom{X_1}{X_2} \sim \text{BVN}(\mu, \Sigma)$. If $A \in \mathbb{R}_{2\times 2}$ with $|A| \neq 0$, i.e. A is invertible, and $b \in \mathbb{R}_{2\times 1}$, then

$$Y := AX + b \sim \text{BVN}(A\mu + b, A\Sigma A^T).$$

Proof. We omit the proof. ■

Corollary 2.10.11 Let $X = \binom{X_1}{X_2} \sim \text{BVN}(\mu, \Sigma)$. If $c = \binom{c_1}{c_2} \neq 0$, then

$$\begin{bmatrix} c_1 & c_2 \end{bmatrix} X \sim N(c^T \mu, c^T \Sigma c).$$

Proof. We omit the proof. ■

Proposition 2.10.12 Let $X = \binom{X_1}{X_2} \sim \text{BVN}(\mu, \Sigma)$, then

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

and

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

Proof. We omit the proof. ■

Proposition 2.10.13 Let $X = \binom{X_1}{X_2} \sim \mathrm{BVN}(\mu, \Sigma)$, then

$$(X-\mu)^T \Sigma^{-1} (X-\mu) \sim \chi^2_2$$

where χ^2_2 is the chi-squared distribution with 2 degrees of freedom.

Proof. We omit the proof. ■

3. Functions of Two or More Random Variables

3.1 Using the Cumulative Distribution Function Technique

Remark 3.1.1 Suppose $X_1, \ldots, X_n$ are continuous rv's with joint pdf $f : (x_1, \ldots, x_n) \to [0,1]$, then the pdf of $Y := h(X_1, \ldots, X_n)$, where h is a real-valued function, can be determined by first finding $\Pr(Y \leq y) = \Pr(h(X_1, \ldots, X_n) \leq y)$. Some care needed to be taken with the support of the X_i's and Y. We demonstrate this using an example.

■ **Example 3.1.2** Suppose X, Y are continuous rv's with pdf

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

Define $S = \frac{Y}{X}$. We first find the cdf of S: $\Pr(S \leq s) = \Pr(Y/X \leq s)$. But what is the support of S? If $s \leq 1$, then $\Pr(Y/X \leq s) = \Pr(Y \leq sX) = 0$, because $\Pr(X > Y) = 0$ by the definition of the joint pdf of (X,Y).

If $s \geq 1$, then we have the shaded area in the following picture as the set $Y \leq sX$:

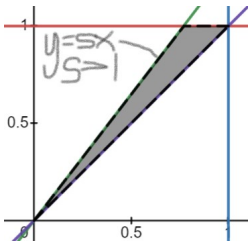


Figure 3.1: $Y \leq sX$

It follows that

$$\Pr\left(\frac{Y}{X} \leq s\right) = \Pr(Y \leq sX) = 1 - \Pr(y > sX) = 1 - \int_0^{\frac{1}{s}} \int_{sx}^1 f(x,y)dydx = 1 - \frac{1}{s}.$$

Hence the cdf of S is

$$F(s) = \left(1 - \frac{1}{s}\right) \mathbf{1}_{s \geq 1}, s \in \mathbb{R}.$$

Differentiate with respect to s yields the pdf of S:

$$f(s) = \frac{1}{s^2}\mathbf{1}_{s\geq 1}.$$

■

3.2 One-to-One Transformations

Definition 3.2.1 Suppose X, Y are continuous rv's. A **one-to-one transformation** S of X and Y are two functions

$$U = h_1(X,Y)$$
$$V = h_2(X,Y)$$

such that for all (x,y) in the support of (X,Y) (denoted R_{XY}, h_1, h_2 maps (x,y) to $R_{UV} \subseteq \mathbb{R}^2$, the support for (U,V).

Definition 3.2.2 Suppose $S : \mathbb{R}^2 \to \mathbb{R}^2$ is defined by $S(x,y) = (h_1(x,y), h_2(x,y)) = (u,v)$ for some $h_1, h_2 : \mathbb{R}^2 \to \mathbb{R}$, then the **Jacobian matrix** of S is

$$JS(x,y) = \begin{bmatrix} \frac{\partial u}{\partial x} & \frac{\partial u}{\partial y} \\ \frac{\partial v}{\partial x} & \frac{\partial v}{\partial y} \end{bmatrix}$$

and the **Jacobian** of S is the determinant of the Jacobian matrix of S:

$$\frac{\partial(u,v)}{\partial(x,y)} = |JS(x,y)|.$$

Theorem 3.2.3 — Inverse Mapping Theorem. Let $R \subseteq \mathbb{R}^2$. Suppose $S : R \to \mathbb{R}^2$ defined by

$$S(x,y) = (h_1(x,y), h_2(x,y)) = (u,v)$$

is a transformation such that $\frac{\partial u}{\partial x}$, $\frac{\partial u}{\partial y}$, $\frac{\partial v}{\partial x}$, $\frac{\partial v}{\partial y}$ are continuous functions and $\frac{\partial(u,v)}{\partial(x,y)} \neq 0$ for all $(x,y) \in \mathbb{R}$, then S is one-to-one on R and $S^{-1} : \mathbb{R}^2 \to R$ exists.

Proof. We omit the proof. ■

Remark 3.2.4 Thm. 3.2.3 provides a sufficient but not necessary condition for the existence of S^{-1}. We now use this to give a formula for the pdf of a transformation of a 2-dimensional random vector using the Jacobian matrix of the transformation.

Theorem 3.2.5 Let X, Y be continuous rv's with joint pdf $f : \mathbb{R}^2 \to [0,1]$ and support $R_{XY} \subseteq \mathbb{R}^2$. Suppose S is a one-to-one transformation

$$S : R_{XY} \to R_{UV}$$
$$(x,y) \mapsto (h_1(x,y), h_2(x,y))$$

where $h_1, h_2 : \mathbb{R}^2 \to \mathbb{R}$ are functions, $R_{UV} \subseteq \mathbb{R}^2$, with its inverse transformation

$$S^{-1} : R_{UV} \to R_{XY}$$
$$(u,v) \mapsto (w_1(u,v), w_2(u,v)),$$

where $w_1, w_2 : \mathbb{R}^2 \to \mathbb{R}$ are functions, then the pdf of (U,V), $g : R_{UV} \to [0,1]$, is given by

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

Proof. Suppose S^{-1} maps from $A \subseteq R_{UV}$ onto $B \subseteq R_{XY}$, then

$$\begin{aligned} \Pr((U,V) \leq (u,v)) &= \iint_A g(u,v) du dv \\ &= \iint_B f(x,y) dx dy \\ &= \iint_A f(w_1(u,v), w_2(u,v)) \left\| \begin{matrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{matrix} \right\| du dv \end{aligned}$$

where the last line follows from a theorem in calculus. Differentiate the above gives

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

■

3.3 Moment Generating Function Technique

Theorem 3.3.1 Suppose $X_1, \ldots, X_n$ are independent rv's and each X_i has MGF $M_i : (-h_i, h_i) \to \mathbb{R}$, then the MGF of $Y := \sum_{i=1}^n X_i$ is given by

$$\begin{aligned} M_Y : (-h,h) &\to \mathbb{R} \\ t &\mapsto \prod_{i=1}^n M_i(t). \end{aligned}$$

In particular, if the X_i's have identical distributions, then Y has the MGF

$$\begin{aligned} M_Y : (-h,h) &\to \mathbb{R} \\ t &\mapsto (M(t))^n. \end{aligned}$$

Proof. By the definition of MGFs, we have

$$\begin{aligned} M_Y(t) &= E(e^{Yt}) \\ &= E\left(\exp\left(t \sum_{i=1}^n X_i \right) \right) \\ &= E\left(\exp\left(\sum_{i=1}^n t X_i \right) \right) \\ &= E\left(\prod_{i=1}^n \exp(t X_i) \right). \end{aligned}$$

Since the X_i's are independent, $E\left(\prod_{i=1}^n \exp(tX_i)\right) = \prod_{i=1}^n E(tX_i)$, and hence

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

as required.

If the X_i's follow the same distribution, then all the $M_i(t)$'s are the same, and $M_Y(t) = (M(t))^n$ immediately. ■

Theorem 3.3.2 If $X_i \sim N(\mu_i, \sigma_i^2)$ independently, $1 \leq i \leq n$, then

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

Proof. Using Thm. 3.3.1 and Thm. 1.9.7 we have

$$M_Y(t) = \prod_{i=1}^{n} M_{\alpha_i X_i}(t) = \prod_{i=1}^{n} M_{X_i}(\alpha_i t)$$

where each

$$M_{X_i}(\alpha_i t) = \exp\left(\mu_i t + \sigma_i^2 + \frac{\alpha_i^2 t^2}{2}\right).$$

Hence

$$M_Y(t) = \exp\left(\sum_{i=1}^{n} \alpha_i \mu_i t + \frac{\sigma_i^2 \alpha_i^2 t^2}{2}\right) = \exp\left(t \sum_{i=1}^{n} \alpha_i \mu_i + \frac{t^2}{2} \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2\right), t \in \mathbb{R}$$

which is the MGF of

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

By the uniqueness of MGF's, this completes the proof. ■

Corollary 3.3.3 If each $X_i \sim N(\sigma, \sigma^2)$ independently, $1 \leq i \leq n$, then

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2)$$

and

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

Proof. Take Thm. 3.3.2 and put each $\alpha_i = 1$, $\sigma_i = \sigma$, $\mu_i = \mu$, and the result immediately follows. For the $\bar{X}$ result, take $\mu_i = \mu$, $\sigma_i = \sigma$, and $\alpha_i = \frac{1}{n}$. ■

Proposition 3.3.4 If $X_i \sim \chi^2_{k_i}$ independently, $1 \leq i \leq n$, then

$$Y := \sum_{i=1}^{n} X_i \sim \chi^2_{\sum_{i=1}^{n} k_i}.$$

Proof. Each X_i has MGF

$$M_i(t) = (1-2t)^{-k_i/2}, t < \frac{1}{2}.$$

By Thm. 3.3.1 we have

$$M_Y(t) = \prod_{i=1}^{n} (1-2t)^{-k_i/2} = (1-2t)^{\sum_{i=1}^{n} \frac{-k_i}{2}} = (1-2t)^{-\frac{\sum_{i=1}^{n} k+i}{2}}, t < \frac{1}{2}$$

which is the MGF of

$$\chi^2_{\sum_{i=1}^n k_i}.$$

By the uniqueness of MGF's, $Y \sim \chi^2_{\sum_{i=1}^n k_i}$ as required. ■

Proposition 3.3.5 If each $X_i \sim N(\mu, \sigma^2)$ independently, $1 \leq i \leq n$, then

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

Proof. Denote, for each i, $Z_i = \frac{X_i - \mu}{\sigma}$, then each $Z_i \sim N(0,1)$. Define each $U_i := Z_i^2$, then, if $u_i \geq 0$,

$$\begin{aligned}
\Pr(U_i \leq u_i) &= \Pr(Z_i^2 \leq u_i) \\
&= \Pr(-\sqrt{u_i} \leq Z_i \leq \sqrt{u_i}) \\
&= \int_{-\sqrt{u_i}}^{\sqrt{u_i}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z_i^2}{2} \right) dz_i \\
&= -\int_0^{-\sqrt{u_i}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z_i^2}{2} \right) dz_i + \int_0^{\sqrt{u_i}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-z_i}{2} \right) dz_i.
\end{aligned}$$

Differentiate $\Pr(U_i \leq u_i)$ with respect to u_i using the Fundamental Thm of Calculus to get the pdf of each U_i to be

$$f(u_i) = -\frac{-1}{2\sqrt{u_i}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u_i}{2} \right) + \frac{1}{2\sqrt{u_i}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u_i}{2} \right) = \frac{1}{\sqrt{u_i}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-u_i}{2} \right), u_i \geq 0.$$

By Proposition 1.4.6, $\sqrt{\pi} = \Gamma\left(\frac{1}{2}\right)$, so re-write

$$f(u_i) = \frac{u_i^{\frac{1}{2}-1} e^{\frac{-u_i}{2}}}{2^{\frac{1}{2}} \Gamma\left(\frac{1}{2}\right)}, u_i \geq 0.$$

This shows that each

$$\left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2_1.$$

Using Proposition 3.3.4, we have

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

as required. ■

Remark 3.3.6 The next results illustrate techniques used in hypothesis testing. It starts out with a technique lemma and ends with a theorem concerning the student t-distribution.

Lemma 3.3.7 Suppose $X_i \sim N(\mu, \sigma^2)$ independently, $1 \leq i \leq n$. Define random variables

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

then $\bar{X} \perp S^2$.

Proof. For notational convenience, all $\sum$ and $\prod$ in this proof are over i from $i = 1$ to $i = n$.
For each i, define $U_i := X_i - \bar{X}$, and U to be the random vector $U := (U_1, \ldots, U_n)$.
Write the joint MGF of U and $\bar{X}$ to be

$$M(s_1, \ldots, s_n, s) = E(\exp(\sum s_i U_i + s\bar{X})).$$

Define $t_i = s_i - \bar{s} + \frac{s}{n}$, where $\bar{s} = \frac{1}{n}\sum s_i$. It follows that

$$\begin{aligned}
&E(\exp(\sum t_i X_i)) \\
=&E\left(\exp\left(\sum s_i X_i - \sum \bar{s} X_i + \sum \frac{s}{n} X_i\right)\right) \\
=&E\left(\exp\left(\sum s_i X_i - \frac{1}{n}\sum s_i X_i + \frac{s}{n}\sum X_i\right)\right) \\
=&E\left(\exp\left(\sum s_i \left(X_i - \frac{X_i}{n}\right) + s\bar{X}\right)\right) \\
=&E\left(\exp\left(\sum s_i (X_i - \bar{X}) + s\bar{X}\right)\right) \\
=&E(\exp(\sum s_i U_i + s\bar{X})) \\
=&M(s_1, \ldots, s_n, s).
\end{aligned}$$

Hence, since each $X_i \sim N(\mu, \sigma^2)$,

$$\begin{aligned}
M(s_1, \ldots, s_n, s) &= E(\exp(\sum t_i X_i)) \\
&= \prod E(\exp(t_i X_i)) \\
&= \prod \exp\left(\mu t_i + \frac{\sigma^2 t_i^2}{2}\right) \\
&= \exp\left(\sum \left(\mu t_i + \frac{\sigma^2 t_i^2}{2}\right)\right).
\end{aligned}$$

Now $\sum t_i = \sum \left(s_i - \bar{s} + \frac{s}{n}\right) = n\bar{s} - n\bar{s} + n\frac{s}{n} = s$, and

$$\begin{aligned}
\sum t_i^2 &= \sum (s_i - \bar{s})^2 + 2\sum (s_i - \bar{s})\left(\frac{s}{n}\right) + \sum \left(\frac{s}{n}\right)^2 \\
&= \sum (s_i - \bar{s})^2 + \frac{1}{n^2} n s^2 \\
&= \sum (s_i - \bar{s})^2 + \frac{s^2}{n}
\end{aligned}$$

which yields

$$\begin{aligned}
M(s_1, \ldots, s_n, s) &= \exp\left(\mu \sum t_i + \frac{\sigma^2}{2}\sum t_i^2\right) \\
&= \exp\left(\mu s + \frac{\sigma^2}{2}\sum (s_i - \bar{s})^2 + \frac{\sigma^2}{2}\frac{s^2}{n}\right).
\end{aligned}$$

This means that the joint MGF of $U_1, \ldots, U_n$ is

$$M_U(s_1, \ldots, s_n) = M(s_1, \ldots, s_n, s = 0) = \exp\left(\frac{\sigma^2}{2}\sum (s_i - \bar{s})^2\right)$$

and the MGF of $\bar{X}$ is

$$M_{\bar{X}}(s) = M(s_1 = 0, \ldots, s_n = 0, s) = \exp\left(\mu s + \frac{\sigma^2 s^2}{2n}\right).$$

Hence

$$M(s_1, \ldots, s_n, s) = M_U(s_1, \ldots, s_n) M_{\bar{X}}(s)$$

and so $U \perp \bar{X}$ by Thm. 2.8.4. By the definition of U, each $U_i = X_i - \bar{X}$ is independent of $\bar{X}$, and consequently

$$S^2 = \frac{(X_i - \bar{X})^2}{n-1} \perp \bar{X}$$

as required. ■

Proposition 3.3.8 If $X_i \sim N(\mu, \sigma^2)$ independently, $1 \leq i \leq n$, and we use the same rv S^2 as in Lemma 3.3.7, then

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

Proof. As before, all summations are over i from $i = 1$ to $i = n$. Recall that $\bar{X} = \frac{1}{n}\sum X_i$. We have

$$\begin{aligned}
\sum (X_i - \mu)^2 &= \sum (X_i - \bar{X} + \bar{X} - \mu)^2 \\
&= \sum (X_i - \bar{X})^2 + 2\sum (X_i - \bar{X})(\bar{X} - \mu) + \sum (\bar{X} - \mu)^2 \\
&= \sum (X_i - \bar{X})^2 + 0 - n(\bar{X} - \mu)^2 \\
&= \sum (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.
\end{aligned}$$

Hence

$$\sum \left(\frac{X_i - \mu}{\sigma}\right)^2 = \frac{(n-1)S^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

Denote $Y := \sum \left(\frac{X_i - \mu}{\sigma}\right)^2$, $U := \frac{(n-1)S^2}{\sigma^2}$, $V := \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2$.
Since $S^2 \perp \bar{X}$ by Lemma 3.3.7, we have $U \perp V$, and moreover by Proposition 3.3.5,

$$Y = \sum \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n$$

with MGF

$$M_Y(t) = (1 - 2t)^{\frac{-n}{2}}, t < \frac{1}{2}.$$

By Corollary 3.3.3,

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

and so

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

By Proposition 3.3.5 again,

$$V = \left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi_1^2$$

with MGF

$$M_V(t) = (1-2t)^{-\frac{1}{2}}, t < \frac{1}{2}.$$

Now $Y = U + V$ and $U \perp V$, so

$$M_Y(t) = M_U(t) M_V(t)$$

which gives

$$M_U(t) = E(e^{tU}) = \frac{M_Y(t)}{M_V(t)} = \frac{(1-2t)^{-\frac{n}{2}}}{(1-2t)^{-\frac{1}{2}}} = (1-2t)^{-\frac{n-1}{2}}, t < \frac{1}{2}.$$

This is the MGF of χ_{n-1}^2. By the uniqueness of MGF's, $\frac{(n-1)S^2}{\sigma^2} = U \sim \chi_{n-1}^2$ as required. ■

Lemma 3.3.9 If $X \sim \chi_n^2$, $Z \sim N(0,1)$, and $X \perp Z$, then

$$T := \frac{Z}{\sqrt{X/n}} \sim t_n$$

where t_n is the student t-distribution with n degrees of freedom.

Proof. Define a one-to-one transformation $S : \mathbb{R}^2 \to \mathbb{R}^2$:

$$S(Z,X) = (T,U) = \left(\frac{Z}{\sqrt{X/n}}, X\right).$$

S has the reverse transformation

$$S^{-1}(T,U) = (Z,X) = \left(T\left(\frac{U}{n}\right)^{\frac{1}{2}}, U\right).$$

Since $X \sim \chi_n^2$, $Z \sim N(0,1)$ and $X \perp Z$, the joint pdf of (Z,X) is the product of the marginal pdf's, namely

$$f_{ZX}(z,x) = f_Z(z) f_X(x) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \cdot \frac{x^{\frac{n}{2}-1} e^{-\frac{x}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)}, x > 0, z \in \mathbb{R}.$$

Using Thm. 3.2.5, the joint pdf of (T,U) is

$$\begin{aligned}
g(t,u) &= f\left(t\left(\frac{u}{n}\right)^{\frac{1}{2}}, u\right)\left|\frac{\partial(z,x)}{\partial(t,u)}\right| \\
&= \frac{e^{-\frac{t^2 u}{2n}}}{\sqrt{2\pi}} \cdot \frac{u^{\frac{n}{2}-1} e^{-\frac{u}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \left|\left| \begin{matrix} \frac{\partial z}{\partial t} & \frac{\partial z}{\partial u} \\ \frac{\partial x}{\partial t} & \frac{\partial x}{\partial u} \end{matrix} \right|\right| \\
&= \frac{e^{-\frac{t^2 u}{2n}}}{\sqrt{2\pi}} \cdot \frac{u^{\frac{n}{2}-1} e^{-\frac{u}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \left|\left| \begin{matrix} \left(\frac{u}{n}\right)^{\frac{1}{2}} & 0 \\ 0 & 1 \end{matrix} \right|\right| \\
&= \frac{e^{-\frac{t^2 u}{2n}}}{\sqrt{2\pi}} \cdot \frac{u^{\frac{n}{2}-1} e^{-\frac{u}{2}}}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} \left(\frac{u}{n}\right)^{\frac{1}{2}}, t \in \mathbb{R}, u > 0.
\end{aligned}$$

From this we get the marginal pdf of T:

$$f_T(t) = \int_{-\infty}^{\infty} g(t,u)du.$$

This eventually yields

$$f_T(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{\pi n}}\left(1+\frac{t^2}{2}\right)^{-\frac{n+1}{2}}, t \in bR,$$

which is exactly the pdf of the Student t-distribution of n degrees of freedom, so $T \sim t_n$ as required.

The trick with the integration above is to let

$$y = u\left(\frac{1}{2}+\frac{t^2}{2n}\right)$$

and substitute u and du in terms of y and dy. ■

Theorem 3.3.10 If $X_i \sim N(\mu,\sigma^2)$ independently, $1 \leq i \leq n$, then

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

where

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

Proof. Re-write

$$\begin{aligned}\frac{\bar{X}-\mu}{S/\sqrt{n}} &= \frac{\bar{X}-\mu}{S/\sqrt{n}} \cdot \frac{\sigma/\sqrt{n}}{\sigma/\sqrt{n}} \\ &= \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{S}{\sqrt{n}}\cdot\frac{\sqrt{n}}{\sigma}} \\ &= \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\frac{S(n-1)/\sigma}{n-1}} \\ &= \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{S^2(n-1)/\sigma^2}{n-1}}}\end{aligned}$$

where $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \perp \frac{S^2(n-1)}{\sigma^2}$ by Proposition 3.3.8. Furthermore

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

by Corollary 3.3.3 and

$$\frac{S^2(n-1)}{\sigma^2} \sim \chi^2_{n-1}$$

by Proposition 3.3.8. Apply Lemma 3.3.9 to get

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

as required. ■

Remark 3.3.11 S^2 in Proposition 3.3.8 and Thm. 3.3.10 is the **sample variance** of a random sample $X_1, \ldots, X_n$, while S is the **sample standard deviation**. Proposition 3.3.8 and Thm. 3.3.10 are useful when constructing estimates and confidence intervals for the σ^2 parameter given a random sample with assumed underlying distribution of $N(\mu, \sigma^2)$.

4. Limiting Asymptotic Distributions

4.1 Convergence in Distribution

Definition 4.1.1 Let $(X_n)_{n=1}^{\infty}$ be a sequence of rv's and $(F_n(x))_{n=1}^{\infty}$ be the corresponding cdf's. Let X be a rv with cdf F. We say that $(X_n)_n$ **converges in distribution** to X and write

$$X_n \xrightarrow{D} X$$

if $\lim_{n\to\infty} F_n(x) = F(x)$ pointwise for all $x \in \mathbb{R}$ at which F is continuous.

Remark 4.1.2 We state some facts from real analysis that helps with finding limiting distributions.

1. Suppose $a < b$ and $f : [a,b] \to \mathbb{R}$ is in C^{∞}, i.e. infinitely differentiable and all derivatives are continuous on $[a,b]$. Let $c \in [a,b]$, then for all $x \in [a,b]$ and $k \in \mathbb{N}$, we have

$$f(x) = \sum_{i=0}^{k} \frac{f^{(i)}(c)(x-c)^i}{i!} + \frac{f^{(k+1)}(\xi_x)(x-c)^{k+1}}{(k+1)!}$$

   for some $\xi_x \in [x,c]$ or $[c,x]$.

2. Let $b, c \in \mathbb{R}$, Φ a function from $\mathbb{R}$ to $\mathbb{R}$ with $\lim_{n\to\infty} \Phi(n) = 0$, then

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

   Proof. Consider the Taylor expansion of $\log(1+x)$ about $c = 0$:

$$\begin{aligned}
\log(1+x) &= \sum_{i=0}^{1} \frac{f^{(i)}(0)(x-0)^i}{i!} + \frac{f^{(2)}(\xi_x)(x-0)^2}{2!} \\
&= \frac{\log(1)x^0}{0!} + \frac{\frac{1}{1+0}(x-0)^i}{1!} + \frac{\frac{-1}{(1-\xi_x)^2}x^2}{2!} \\
&= x - \frac{x^2}{2(1+\xi_x)^2}, \xi_x \in [0,x].
\end{aligned}$$

Therefore

$$\log\left(1+\frac{b}{n}+\frac{\Phi(n)}{n}\right)=\frac{b}{n}+\frac{\Phi(n)}{n}-\frac{\left(\frac{b}{n}+\frac{\Phi(n)}{n}\right)^2}{2(1+\xi_x)^2}$$

and

$$cn\cdot\log\left(1+\frac{b}{n}+\frac{\Phi(n)}{n}\right)=bc+\Phi(n)c-\frac{(b+\Phi(n))^2}{2n(1+\xi_x)^2}$$

which yields

$$\lim_{n\to\infty} cn\cdot\log\left(1+\frac{b}{n}+\frac{\Phi(n)}{n}\right)=bc$$

as required. ■

3. Let $b,c\in\mathbb{R}$, then

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

■ **Example 4.1.3** Consider a sequence $(X_i)_i$ where each

$$X_i\sim\text{Uniform}\left(0,\frac{1}{i}\right),$$

and denote the cdf of X_i to be F_i.

Define X such that $\Pr(X=0)=1$. Consider the cdf $F_X:\mathbb{R}\to[0,1]$ of X. Clearly $F_X(x)=\mathbf{1}_{x\geq 0}$. In particular F_X has a jump discontinuity at $x=0$.

Fix $x\neq 0$.

If $x<0$, then each $0\leq F_i(x)=\Pr(X_i\leq x)\leq\Pr(X_i<0)=0$, so $F_i(x)=0$.

For each $i\in\mathbb{N}$, if $x\in\left[0,\frac{1}{i}\right]$, then $F_i(x)=\int_0^x 1/(1/i)dx=ix$. If $x>\frac{1}{i}$, $F_i(x)=1$.

In summary

$$F_i(x)=\Pr(X_i\leq x)=\begin{cases}0 & \text{if} \quad x<0\\ ix & \text{if} \quad x\in\left[0,\frac{1}{i}\right]\\ 1 & \text{if} \quad x>\frac{1}{i}\end{cases}$$

As $i\to\infty$, $\frac{1}{i}\to 0$. Now note that the cdf of X is

$$F_X(x)=\begin{cases}0 & \text{if} \quad x<0\\ 1 & \text{if} \quad x\geq 0\end{cases}$$

Thus for every $x\neq 0$, $F_i(x)\to F(x)$ pointwise as $i\to\infty$. Thus $X_i\xrightarrow{D}X$ by definition.

This naturally gives rise to the following definition. ■

Definition 4.1.4 Fix $c\in\mathbb{R}$. A rv Y has a **degenerate distribution** at $y=c$ if Y has the cdf

$$F_Y(y)=\mathbf{1}_{y\geq c}, y\in\mathbb{R},$$

i.e. Y has pdf $f_Y(y)=\mathbf{1}_{y=c}, y\in\mathbb{R}$.

Definition 4.1.5 $(X_i)_i$ is said to **converge stochastically to a constant** c if $X_i \xrightarrow{D} X$, where X is a degenerate distribution at c.

4.2 Convergence in Probability

Definition 4.2.1 A sequence of rv's $(X_n)_{n=1}^{\infty}$ **converges in probability** to a random variable X if for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - X| \geq \varepsilon) = 0,$$

or equivalently

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

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

Theorem 4.2.2 Let $(X_n)_n$ be a sequence of rv's. If $X_n \xrightarrow{P} X$ for some rv X, then $X_n \xrightarrow{D} X$.

Proof. Suppose that $X_n \xrightarrow{P} X$. We show that for all $\varepsilon > 0$,

$$F_X(a-\varepsilon) \leq \lim_{n\to\infty} F_n(a) \leq F_X(a+\varepsilon)$$

for all $a \in \mathbb{R}$ such that F_X is continuous at a. This directly implies the result we need.

Fix $a \in \mathbb{R}$ such that F_X is continuous at a. Denote F_n to be the cdf of X_n. Note that for each $n \in \mathbb{N}$,

$$\begin{aligned}
F_n(a) &= \Pr(X_n \leq a, X \leq a+\varepsilon) + \Pr(X_n \leq a, X > a+\varepsilon) \\
&= \Pr(X_n \leq a | X \leq a+\varepsilon) \Pr(X \leq a+\varepsilon) + \Pr(X_n \leq a, X > a+\varepsilon) \\
&\leq \Pr(X \leq a+\varepsilon) + \Pr(X_n \leq a, X > a+\varepsilon) \\
&= \Pr(X \leq a+\varepsilon) + \Pr(X_n \leq a, X - \varepsilon > a) \\
&\leq \Pr(X \leq a+\varepsilon) + \Pr(X - \varepsilon > X_n) \text{ since } a \geq X_n \\
&\leq \Pr(X \leq a+\varepsilon) + \Pr(X_n - X < -\varepsilon) + \Pr(X_n - X > \varepsilon) \\
&= \Pr(X \leq a+\varepsilon) + \Pr(|X_n - X| > \varepsilon).
\end{aligned}$$

Thus

$$\Pr(X \leq a+\varepsilon) \geq F_n(a) - \Pr(|X_n - X| > \varepsilon). \tag{4.1}$$

Similarly,

$$\begin{aligned}
F_X(a-\varepsilon) &= \Pr(X \leq a-\varepsilon) \\
&= \Pr(X \leq a-\varepsilon, X_n \leq a) + \Pr(X \leq a-\varepsilon, X_n > a) \\
&\leq \Pr(X \leq a-\varepsilon | X_n \leq a) \Pr(X_n \leq a) + \Pr(X \leq a-\varepsilon, X_n > a) \\
&\leq \Pr(X_n \leq a) + \Pr(X + \varepsilon \leq a, a < X_n) \\
&\leq \Pr(X_n \leq a) + \Pr(X + \varepsilon < X_n) \\
&\leq \Pr(X_n \leq a) + \Pr(X - X_n > \varepsilon) + \Pr(X - X_n < -\varepsilon) \\
&= \Pr(X_n \leq a) + \Pr(|X - X_n| > \varepsilon)
\end{aligned}$$

which shows that

$$F_X(a-\varepsilon) \leq F_n(a) - \Pr(|X_n - X| > \varepsilon). \tag{4.2}$$

By Def. 4.2.1, for a $\varepsilon > 0$, $\Pr(|X_n - X| > \varepsilon)$ can be made arbitrarily small. Hence, as $n \to \infty$, (4.1) becomes

$$\begin{aligned}
&\lim_{n\to\infty} F_n(a) - \Pr(|X_n - X| > \varepsilon) \\
&= \lim_{n\to\infty} F_n(a) - \lim_{n\to\infty} \Pr(|X_n - X| > \varepsilon) \\
&= \lim_{n\to\infty} F_n(a) \leq \Pr(X \leq a+\varepsilon) = F_X(a+\varepsilon).
\end{aligned}$$

Similarly inequality (4.2) yields

$$\lim_{n\to\infty} F_n(a) \geq F_X(a-\varepsilon),$$

which, together, yields the desired result:

$$F_X(a-\varepsilon) \leq \lim_{n\to\infty} F_n(a) \leq F_X(a+\varepsilon).$$

This completes the proof. ■

Theorem 4.2.3 The sequence $(X_n)_n$ converges stochastically to a constant c if and only if $(X_n)_n$ converges in probability to the degenerate distribution at c, i.e. for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr(|X_n - c| < \varepsilon) = 1.$$

Proof. ($\Rightarrow$) Suppose $X_n \xrightarrow{D} X$ where X is the degenerate distribution at c, then the pointwise limit of $F_n(x)$, which are the cdf's of X_n's, is $F(x) = \mathbf{1}_{x \geq c}$.

Let $\varepsilon > 0$, then

$$\begin{aligned}
\Pr(|X_n - c| < \varepsilon) &= \Pr(X_n - c < \varepsilon) + \Pr(X_n - c > -\varepsilon) \\
&= \Pr(X_n < \varepsilon + c) + \Pr(X_n > c - \varepsilon) \\
&= \Pr(X_n < \varepsilon + c) + 1 - \Pr(X_n \leq c - \varepsilon).
\end{aligned}$$

Take $n \to \infty$, then $\Pr(X_n < \varepsilon + c)$ converges to $\Pr(X < \varepsilon + c)$ and $\Pr(X_n \leq c - \varepsilon)$ converges to $\Pr(X \leq c - \varepsilon)$ by the pointwise convergence of F_n to F_X. Since $c - \varepsilon < c$, we have

$$\Pr(X_n \leq c - \varepsilon) \to 0.$$

Moreover, since $\varepsilon > 0$ is arbitrary,

$$\Pr(X_n < c + \varepsilon) \to 0.$$

So $\Pr(|X_n - c| < \varepsilon) \to 1$ and this completes the proof.

($\Leftarrow$) Suppose conversely that $X_n \xrightarrow{P} X$ where X is a degenerate distribution at c. Apply Thm. 4.2.2 directly and we get $X_n \xrightarrow{D} X$ as required. ■

Remark 4.2.4 For a constant c, we observe that

$$\begin{aligned}
&X_n \xrightarrow{P} c \\
\Leftrightarrow &X_n \xrightarrow{D} c \\
\Leftrightarrow &\lim_{n\to\infty} \Pr(|X_n - c| \geq \varepsilon) = 0 \\
\Leftrightarrow &\lim_{n\to\infty} \Pr(|X_n - c| < \varepsilon) = 1.
\end{aligned}$$

4.3 Weak Law of Large Numbers

Theorem 4.3.1 — Weak Law of Large Numbers. Suppose $(X_n)_n$ are independently and identically distributed with each $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$. Define the sequence $(\bar{X}_n)_n$ where each

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

Then $\bar{X}_n \xrightarrow{P} \mu$.

Proof. Let $\varepsilon > 0$.

By construction, each $\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$. Denote $\sigma_n = \frac{\sigma}{\sqrt{n}}$, then let $k_n = \frac{\varepsilon}{\sigma_n}$. By Chebyshev's Inequality,

$$\Pr(|\bar{X}_n - \mu| \geq k_n\sigma_n) = \Pr\left(|\bar{X}_n - \mu| \geq \frac{\varepsilon}{\sigma_n}\sigma_n\right) \leq k_n^2 = \frac{\sigma_n^2}{\varepsilon^2}.$$

Thus

$$\Pr(|\bar{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \text{ since } \sigma_n^2 = \frac{\sigma^2}{n}$$

and

$$0 \leq \lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| \geq \varepsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$$

giving, by the Squeeze Theorem,

$$\lim_{n\to\infty} \Pr(|\bar{X}_n - \mu| \geq \varepsilon) = 0.$$

By definition, $\bar{X}_n \xrightarrow{P} \mu$. ■

4.4 MGF Technique for Limiting Distributions

Theorem 4.4.1 Let $(X_n)_n$ be a sequence of rv's with MGF's $(M_n(t))_n$. Suppose X is a rv with MGF $M(t)$. If there exists $h > 0$ such that

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

for all $t \in (-h, h)$, then

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

Proof. We omit the proof. ■

Remark 4.4.2 We next prove the Central Limit Theorem using the above result. It is not the most robust proof because it assumes that each rv in the sequence has a well-defined MGF.

Theorem 4.4.3 — Central Limit Theorem. Let $(X_i)_i$ be a sequence of independently, identically distributed rv's with each $E(X_i) = \mu$, $\text{Var}(X_i) = \sigma^2$ which are both finite, then

$$\frac{n\bar{X} - n\mu}{\sqrt{n}\sigma} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \xrightarrow{D} Z \sim N(0,1)$$

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

Proof. Since all X_i's have a well-defined MGF, let $m_i : (-h_i.h_i) \to \mathbb{R}$ be the MGF of $X_i - \mu$. By Remark 1.9.3 and Proposition 1.9.4, for each i,

$$\begin{aligned} m_i(0) &= 1 \\ m_i'(0) &= E(X_i - \mu) = 0 \\ m_i''(0) &= E((X_i - \mu)^2) = \sigma^2. \end{aligned}$$

Now, for $t \in (-h_i, h_i)$, the MacLaurin series of $m_i(t)$ is

$$m_i(t) = m(0) + m'(0)t + \frac{m''(\xi)t^2}{2!}$$

for some $\xi \in [0,t]$ or $\xi \in [t,0]$. Re-write:

$$m_i(t) = 1 + \frac{m''(\xi)t^2}{2} + \frac{\sigma^2 t^2}{2} - \frac{\sigma^2 t^2}{2} = 1 + \frac{\sigma^2 t^2}{2} + \frac{m''(\xi) - \sigma^2}{2} t^2.$$

Now define

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{n\bar{X} - n\mu}{\sqrt{n}\sigma} = \frac{n(\sum_{i=1}^n X_i/n) - \sum_{i=1}^n \mu}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma}.$$

Hence the MGF of Z, by Thm. 1.9.7 and Thm. 3.3.1, is

$$\begin{aligned} M_{Z_n}(t) &= \prod_{i=1}^{n} m_i\left(\frac{t}{\sqrt{n}\sigma}\right) \\ &= \left(1 + \frac{\sigma^2\left(\frac{t}{\sqrt{n}\sigma}\right)^2}{2} + \frac{m''(\xi) - \sigma^2}{2}\left(\frac{t}{\sqrt{n}\sigma}\right)^2\right)^n \\ &= \left(1 + \frac{t^2}{2n} + \frac{m''(\xi) - \sigma^2}{2n\sigma^2} t^2\right)^n, |\xi| < \left|\frac{t}{\sqrt{n}\sigma}\right| \\ &= \left(1 + \frac{t^2}{2n} + \frac{(m''(\xi) - \sigma^2)t^2/2\sigma^2}{n}\right)^n, |\xi| < \left|\frac{t}{\sqrt{n}\sigma}\right|. \end{aligned}$$

As $n \to \infty$, $\left|\frac{t}{\sqrt{n}\sigma}\right| \to 0$, and thus $\xi \to 0$. This means that as $n \to \infty$,

$$m''(\xi) \to m''(0) = \sigma^2 \quad (m'' \text{ continuous})$$

and consequently as $n \to \infty$,

$$(m''(\xi) - \sigma^2)t^2/2\sigma^2 \to 0.$$

By Remark 4.1.2(2),

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + \frac{(m''(\xi) - \sigma^2)t^2/2\sigma^2}{n}\right)^n = e^{\frac{t^2}{2}}, t \in \mathbb{R}.$$

Thus by Thm. 4.4.1, $Z_n \xrightarrow{D} Z \sim N(0,1)$ since $e^{\frac{t^2}{2}}, t \in \mathbb{R}$ is the MGF of $N(0,1)$. ■

Remark 4.4.4 Below we present several small results, which lead to the so-called δ-method, which helps determine the limiting distribution of a function of a sequence of rv's. The function needs to have some nice properties.

Proposition 4.4.5 If $X_n \xrightarrow{P} a$ for some $a \in \mathbb{R}$, i.e. $X_n \xrightarrow{P} X$ where X is a degenerate distribution at a, and $g : \mathbb{R} \to \mathbb{R}$ is a continuous function at a, then $g(X_n) \xrightarrow{P} g(a)$.

Proof. Let $\varepsilon > 0$. Because g is continuous at a, there exists $\delta > 0$ such that $|x-a| < \delta$ implies $|g(x) - g(a)| < \varepsilon$. This further implies that the event $(|X-a| < \delta) \subseteq (|g(X_n) - g(a)| < \varepsilon$ for each n. Therefore

$$\Pr(|g(X_n) - g(a)| < \varepsilon) \geq \Pr(|X-a| < \delta)$$

and taking limit yields

$$\lim_{n\to\infty} \Pr(|g(X_n) - g(a)| < \varepsilon) \geq \lim_{n\to\infty} \Pr(|X-a| < \delta).$$

Now $X_n \xrightarrow{P} X$, so by definition

$$\lim_{n\to\infty} \Pr(|g(X_n) - g(a)| < \varepsilon) \geq \lim_{n\to\infty} \Pr(|X-a| < \delta) = 1.$$

But of course, all probabilities are ≤ 1, so

$$\lim_{n\to\infty} \Pr(|g(X_n) - g(a) < \varepsilon) \leq 1.$$

In summary,

$$\lim_{n\to\infty} \Pr(|g(X_n) - g(a) < \varepsilon) \leq 1 \leq \lim_{n\to\infty} \Pr(|g(X_n) - g(a) < \varepsilon).$$

The Squeeze Theorem yields

$$\lim_{n\to\infty} \Pr(|g(X_n) - g(a) < \varepsilon) = 1$$

and so $g(X_n) \xrightarrow{P} g(a)$. ■

Proposition 4.4.6 If $X_n \xrightarrow{P} a$, $Y_n \xrightarrow{P} b$, $a, b \in \mathbb{R}$, and $g : \mathbb{R}^2 \to \mathbb{R}$ is continuous at (a,b), then

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

Proof. We omit the proof. ■

Theorem 4.4.7 — Slutsky's Theorem. If $X_n \xrightarrow{P} c$ for some $c \in \mathbb{R}$ and $Y_n \xrightarrow{P} Y$, then

1. $X_n + Y_n \xrightarrow{D} c + Y$.
2. $X_n Y_n \xrightarrow{D} cY$.
3. $\frac{Y_n}{c} \xrightarrow{D} \frac{Y}{c}$ if $c \neq 0$.

Proof. We omit the proof. ■

Theorem 4.4.8 — The δ-method. Suppose $(X_i)_i$ is a sequence of rv's with

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

for some $b > 0$, $a \in \mathbb{R}$. Suppose $g : \mathbb{R} \to \mathbb{R}$ is a function such that g is differentiable at a with

$g'(a) \neq 0$, then

$$n^b(g(X_i) - g(a)) \xrightarrow{D} g'(a)X.$$

Proof. Since $b > 0$, we have $\lim_{n\to\infty} n^{-b} = 0$ and hence

$$\lim_{n\to\infty} \Pr(|n^{-b} - 0| \geq \varepsilon) = 0$$

i.e. $n^{-b} \xrightarrow{P} 0$.

Since $n^b(X_i - a) \xrightarrow{D} X$ and $n^{-b} \xrightarrow{P} 0$, by Slutsky's Theorem,

$$X_i - a = n^{-b}n^b(X_i - a) \xrightarrow{D} 0.$$

Now, by Thm. 4.2.3, we have $X_i - a \xrightarrow{P} 0$. Use Taylor expansion of $g(X_i)$ around a to yield

$$g(X_i) = g(a) + \frac{g'(\xi)(X_i - a)}{1!}$$

for some $\xi \in [a, X_i]$ or $\xi \in [X_i, a]$.

Since $X_i \xrightarrow{P} a$ as $i \to \infty$, $\xi \to a$ and hence $\xi \xrightarrow{P} a$. But g' is continuous at a, so

$$g'(\xi) \xrightarrow{P} g'(a).$$

In combination with the fact that $n^b(X_i - a) \xrightarrow{D} X$, use Slutsky's Theorem again to yield

$$g'(\xi)n^b(X_i - a) \xrightarrow{D} g'(a)X.$$

However $g'(\xi)(X_i - a) = g(X_i) - g(a)$. Substitution yields

$$n^b(g(X_i) - g(a)) \xrightarrow{D} g'(a)X$$

as required. ■

Corollary 4.4.9 If $(X_i)_i$ is a sequence of rv's with $E(X_i) = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$, and $g : \mathbb{R} \to \mathbb{R}$ is differentiable at μ with $g'(\mu) \neq 0$, then

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{D} X \sim N(0, (g'(\mu))^2\sigma^2).$$

Proof. By the Central Limit Theorem we have

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{D} Y \sim N(0, \sigma^2).$$

By Thm. 4.4.8 we have

$$\sqrt{n}(g(\bar{X}_n) - g(\mu))/g'(\mu) \xrightarrow{D} Y \sim N(0, \sigma^2)$$

i.e.

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{D} Z \sim N(0, (g'(\mu))^2\sigma^2).$$

■

5. One Parameter Maximum Likelihood Estimation

5.1 Introduction

Definition 5.1.1 Let $X := (X_1, \ldots, X_n)$ be a random vector. A **statistic** $T(X)$ is a function of X that does not depend on any unknown values, i.e. the distribution of each X_i is known.

Remark 5.1.2 A statistic can be calculated explicitly when the rv's are realised. Some examples of statistics are $\bar{X}$ for some known rv X, the sample variance S^2, or $\sum_{i=1}^{n} \frac{1}{X_i}$...

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ independently, then $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is not a statistic because σ and μ are not known.

Definition 5.1.3 Suppose θ is an unknown parameter in the distribution of $X := (X_1, \ldots, X_n)$. A statistic $T(X)$ that is used to estimate the value of θ is called an **estimator** of θ, written as $\tilde{\theta}$. An observed value of T, say $t = T(x)$, is an **estimate** of θ, written as $\hat{\theta}$.

Remark 5.1.4 Estimators are random variables, whereas estimates are real numbers.

Definition 5.1.5 An estimator $\tilde{\theta}$ is **unbiased** if $E(\tilde{\theta}) = \theta$.

Definition 5.1.6 An estimator $\tilde{\theta}$ is **consistent** if $\tilde{\theta}_n \xrightarrow{P} \theta$, where each $\tilde{\theta}_n$ is the estimator based on $T(X_1, \ldots, X_n)$, i.e. $\tilde{\theta}_n$ converges in probability to θ as the number of observations approaches infinity.

5.2 Maximum Likelihood Method

Definition 5.2.1 Suppose X is a rv whose distribution has one unknown parameter θ and whose pdf is $f(x;\theta)$. Denote Ω to be the set of all possible values that θ may take. Let x be an observed value of X. The **likelihood function for** θ based on x is

$$\begin{aligned} L : \Omega &\to [0,1] \\ \theta &\mapsto \Pr(X = x; \theta) = f(x;\theta). \end{aligned}$$

If $X := (X_1, \ldots, X_n)$ is a **random sample of size** n from a population which has pdf $f(x;\theta)$, then the likelihood function for θ based on observation $x := (x_1, \ldots, x_n)$ is

$$L(\theta) = \Pr(X_1 = x_1, \ldots, X_n = x_n; \theta) = \prod_{i=1}^{n} f(x_i;\theta), \theta \in \Omega.$$

Definition 5.2.2 Let $L : \Omega \to [0,1]$ be the likehood function based on x, which is a realised value of rv X. The element in Ω such that it maximises L is the **maximum likelihood estimate** of θ, denoted by $\hat{\theta} = \hat{\theta}(x)$. The corresponding estimator is the **maximum likelihood estimator**, denoted by $\tilde{\theta}(X)$.

We use **MLE** as an abbreviation for both maximum likelihood estimate and maximum likelihood estimator when the context is understood.

Remark 5.2.3 The maximum likelihood estimator enjoys an invariance property under certain conditions: if $\hat{\theta}$ is the MLE for θ, then $g(\hat{\theta})$ is the MLE for $g(\theta)$ where $g : \Omega \to \mathbb{R}$ is a function that satisfies certain properties.

This is known as **Zehna's Theorem**.

5.3 Score and Information Functions

Definition 5.3.1 Suppose X is a rv with observation x, and X has an unknown parameter θ. The **score function of** θ based on x is

$$S : \Omega \to \mathbb{R}$$
$$\theta \mapsto \frac{d}{d\theta} \log(L(\theta))$$

where L is the likelihood function.

For convenience, we use ℓ to denote $\log(L)$.

Definition 5.3.2 Suppose X is a rv with observation x, and X has an unknown parameter θ. The **information of** θ based on x is

$$I : \Omega \to \mathbb{R}$$
$$\theta \mapsto -\frac{d^2}{d\theta^2}\ell(\theta).$$

Let $\hat{\theta}$ be an estimate for θ. $I(\hat{\theta})$ is the **observed information**.

Remark 5.3.3 Note that the X in the above two definitions can be random vectors, i.e. $X = (X_1, \ldots, X_n)$, X_i are real-valued.

Information measures the curvature of the likelihood function on Ω. Generally, when the sample size increases, more information is found.

Information depends on the data collected, so it is a function of the random vector X. This gives rise to the following information.

Definition 5.3.4 Suppose X is a rv with observation x, and X has an unknown parameter θ. The **Fisher information function of** θ based on x is

$$J : \Omega \to \mathbb{R}$$
$$\theta \mapsto E(I(\theta;X) = E\left(-\frac{\partial^2}{\partial\theta^2}\ell(\theta;X)\right)$$

where $X = (X_1, \ldots, X_n)$ is the random sample.

Remark 5.3.5 It follows from properties of logrithm that if $X_1, \ldots, X_n$ have pdf $f(x;\theta)$, then

$$J(\theta) = nE\left(-\frac{\partial^2}{\partial\theta^2}\log(f(x;\theta))\right).$$

■ **Example 5.3.6** Suppose $X_1, \ldots, X_n \sim \text{Bin}(1,p)$ independently, i.e. each X_i follows the Bernoulli distribution with parameter $p \in [0,1]$. Note that $\Omega = [0,1]$ since p can only take a number between 0 and 1.

If $x = (x_1, \ldots, x_n)$ is a sample of size n, the likelihood function of p based on x is

$$\begin{aligned} L : [0,1] &\to [0,1] \\ p &\mapsto \prod_{i=1}^{n} f(x_i;p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} \end{aligned}$$

which yields

$$\ell(p;x) = \sum_{i=1}^{n} x_i \log(p) + \sum_{i=1}^{n}(1-x_i)\log(1-p).$$

Setting $\frac{\partial \ell}{\partial p} = 0$ yields

$$0 = \frac{n\bar{x}}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = \frac{n\bar{x}(1-p) - np + p\sum_{i=1}^{n} x_i}{p(1-p)} = \frac{n\bar{x} - np}{p(1-p)}.$$

Hence the maximum likelihood estimator of p is $\tilde{p} = \bar{X}$, and the maximum likelihood estimate of p given sample x is $\hat{p} = \bar{x}$.

Now

$$-\frac{\partial^2}{\partial p^2}\ell(p;x) = -\frac{\partial}{\partial p}\left(\frac{n\bar{x}}{p} - \frac{n}{1-p} + \frac{n\bar{x}}{1-p}\right) = \frac{1}{p^2}n\bar{x} + \frac{1}{(1-p)^2}\sum_{i=1}^{n}(1-x_i)$$

and so

$$J(p) = \frac{1}{p^2}nE(\bar{X}) + \frac{n}{(1-p)^2}E(1-\bar{X}) = \frac{n}{p} + \frac{n}{1-p} = \frac{n}{p(1-p)}.$$

Note that the variance of $\tilde{p} = \bar{X}$ is $\frac{1}{n}p(1-p) = \frac{1}{J(p)}$.

This means that for the Bernoulli distribution, the variance of the MLE and the Fisher information are reciprocals of each other. ■

■ **Example 5.3.7** Suppose $X_1, \ldots, X_n \sim \text{Poisson}(\theta)$ independently and $x = (x_1, \ldots, x_n)$ is a sample of size n, then the likelihood function based on x is

$$L(\theta;x) = \prod_{i=1}^{n} \frac{\theta^{x_i}e^{-\theta}}{x_i!} = \frac{\theta^{\sum_{i=1}^{n} x_i}e^{-n\theta}}{\prod_{i=1}^{n} x_i!}, \theta \in \mathbb{R}^+$$

and the log likelihood function is

$$\ell(\theta;x) = \sum_{i=1}^{n} x_i \log(\theta) - n\theta - \log\left(\sum_{i=1}^{n} x_i!\right), \theta \in \mathbb{R}^+.$$

Set

$$0 = \frac{\partial \ell}{\partial \theta} = \frac{n\bar{x}}{\theta} - n$$

and this gives $\hat{\theta}_{ML} = \bar{x}$, $\tilde{\theta}_{ML} = \tilde{X}$. Furthermore

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{n\bar{x}}{\theta^2}.$$

Hence

$$J(\theta) = E\left(\frac{n\bar{X}}{\theta^2}\right) = \frac{n}{\theta^2}\theta = \frac{n}{\theta}$$

and

$$\mathrm{Var}(\tilde{\theta}) = \mathrm{Var}(\bar{X}) = \frac{\theta}{n}$$

which again yields

$$\mathrm{Var}(\tilde{\theta}) = J(\theta)^{-1}.$$

■

Remark 5.3.8 In general, $\mathrm{Var}(\tilde{\theta}_{ML}) = J(\theta)^{-1}$ is false. This property makes the Poisson and Bernoulli distributions rather special.

■ **Example 5.3.9** Suppose $X_1, \ldots, X_n \sim \mathrm{Uniform}(0, \theta)$ independently, then the likelihood function based on observation $x = (x_1, \ldots, x_n)$ is

$$L(\theta; x) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbf{1}_{x_i \leq \theta} = \frac{1}{\theta^n} \prod_{i=1}^{n} \mathbf{1}_{x_i \leq \theta}, \theta \in \mathbb{R}^+.$$

where $L(\theta; x) \neq 0$ if and only if $\max\{x_i : 1 \leq i \leq n\} \leq \theta$. Hence

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x) = \max\{x_i : 1 \leq i \leq n\}.$$

■

■ **Example 5.3.10** Suppose $X_1, \ldots, X_n \sim \mathrm{Uniform}(\theta, \theta + 1)$ independently, and $x = (x_1, \ldots, x_n)$ is an observation, then the likelihood function based on x is

$$L(\theta; x) = \prod_{i=1}^{n} \mathbf{1}_{\theta \leq x_i \leq \theta + 1} = \mathbf{1}_{\theta \leq \min\{x_1 : 1 \leq i \leq n\}} \mathbf{1}_{\max\{x_i : 1 \leq i \leq n\} \leq \theta + 1}.$$

Denoting $x_{(1)} = \min\{x_i : 1 \leq i \leq n\}$ and $x_{(n)} = \max\{x_i : 1 \leq i \leq n\}$, we have $L(\theta; x) \neq 0$ if and only if $\theta \leq x_{(1)}$ and $\theta \geq x_{(n)} - 1$. Thus

$$\hat{\theta} = \arg\max_{\theta} L(\theta; x) = [x_{(n)} - 1, x_{(1)}].$$

i.e. $\hat{\theta}$ takes on an uncountably many different possible values. ■

Remark 5.3.11 Compared to Def. 5.3.4, there is a more general definition for Fisher information. Def. 5.3.4 is equivalent to the more general definition under certain regularity conditions (roughly six of them). The only one that needs checking in this course is that the support of the rv does not depend on the unknown parameter(s) θ.

Example 5.3.9 and Example 5.3.10 are instances where this condition does not hold.

Here is the more general definition of Fisher information.

Definition 5.3.12 The **Fisher information** of unknown parameter θ that is in the rv X is

$$J(\theta) = E\left[\left(\frac{\partial}{\partial\theta}\ell(\theta;X)\right)^2\right] = E(S(\theta;X)^2).$$

Proposition 5.3.13 Given a rv X with sample space Ω, the score function of parameter θ has expectation 0: $E(S(\theta;X)) = 0$.

Consequently, the Fisher information of θ as defined in Def. 5.3.12 is the variance of the score function of θ: $J(\theta) = \mathrm{Var}(S(\theta;X))$.

Proof. By Def. 5.2.1, the likelihood function can be re-written as pdf. By the definition of expectation,

$$\begin{aligned} E(S(\theta;X)) &= \int_\Omega f(\theta;x)\frac{\partial}{\partial\theta}\log(L(\theta;x))dx \\ &= \int_\Omega f(\theta;x)\frac{1}{f(\theta;x)}\frac{\partial}{\partial\theta}f(\theta;x)dx \\ &= \int_\Omega \frac{\partial}{\partial\theta}f(\theta;x)dx. \end{aligned}$$

The regularity conditions state that

$$\int_\Omega \frac{\partial}{\partial\theta}f(x;\theta)dx = \frac{\partial}{\partial\theta}\int_\Omega f(\theta;x)dx = \frac{\partial}{\partial\theta}(1) = 0.$$

Finally, $\mathrm{Var}(S(\theta)) = E(S(\theta)^2) - E(S(\theta))^2 = E(S(\theta)^2)$ which coincides with Def. 5.3.12, so Fisher information is equal to the variance of the score information. ■

Remark 5.3.14 The derivation of the above proof requires some assumptions for the pdf f:

1. f is twice differentiable.
2. $\int f(x;\theta)$ can be differentiated twice under the integral sign with respect to θ.

5.4 Invariance Property of the Maximum Likelihood Estimator

Remark 5.4.1 We present Remark 5.2.3 formally here.

Theorem 5.4.2 — Invariance of MLEs/Zehna's Theorem. If $\hat{\theta}$ is the maximum likelihood estimate of a parameter θ that determines the distribution of a rv X, and $g : \mathbb{R} \to \mathbb{R}$ is a one-to-one transformation, then $g(\hat{\theta})$ is the maximum likelihood estimate of $g(\theta)$.

Proof. Denote Ω to be the parameter space. Clearly $\hat{\theta} \in \Omega$. By the definition of MLEs we have

$$\Pr(X = x|\theta = \hat{\theta}) \geq \Pr(X = x|\theta = \theta_0)$$

for all $\theta_0 \in \Omega$. By the one-to-one property of g we have

$$\Pr(X = x|g(\theta) = g(\hat{\theta})) \geq \Pr(X = x|g(\theta) = g(\theta_0))$$

for all $\theta_0 \in \Omega$, so $g(\hat{\theta})$ is the MLE of $g(\theta)$. ■

■ **Example 5.4.3** If $X_1,\ldots,X_n \sim \text{Exp}(\theta)$ independently, then the median of $X_1,\ldots,X_n$ is found by solving for m in the equation

$$\int_{-\infty}^{m} \frac{1}{\theta} e^{-\frac{x}{\theta}} dx = \frac{1}{2}.$$

This yields $m = \log(2)\theta$.
The likelihood function of θ based on observation $x = (x_1,\ldots,x_n)$ is

$$L(\theta;x) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-\frac{x_i}{\theta}} = \frac{1}{\theta^n} \exp\left(\frac{-\sum_{i=1}^{n} x_i}{\theta}\right)$$

and the log likelihood function is

$$\ell(\theta;x) = -n\log\theta - \frac{\sum_{i=1}^{n} x_i}{\theta}$$

with

$$\ell'(\theta;x) = -\frac{n}{\theta} + \frac{\sum_{i=1}^{n} x_i}{\theta^2}.$$

Setting $\ell'(\theta;x) = 0$ yields $1 = \frac{\bar{x}}{\theta}$, so the maximum likelihood estimate for θ given x is $\hat{\theta} = \bar{x}$. By Invariance of MLEs, the MLE for m is

$$\hat{m} = \log(2)\hat{\theta} = \log(2)\bar{x}.$$

■

5.5 Likelihood Intervals

Definition 5.5.1 Suppose $X_1,\ldots,X_n$ have the same pdf $f(x;\theta)$ independently, and the parameter θ has likelihood function L and MLE $\hat{\theta}$ based on observation $x_1,\ldots,x_n$. The **relative likelihood function** of θ is

$$R : \theta \to \frac{L(\theta)}{L(\hat{\theta})}, \theta \in \Omega.$$

Remark 5.5.2 There are estimates other than maximum likelihood estimates. Suppose we have an estimate θ_0 of θ based on observation x, and $R(\theta_0) = L(\theta_0)/L(\hat{\theta}) \leq 0.1$, then the observation x is at least 10 times more likely to be observed if $\theta = \hat{\theta}$ than $\theta = \theta_0$.

A rule of thumb. If $R(\theta_0) \geq 0.5$, then θ_0 is a plausible value of θ given x.

Definition 5.5.3 Fix $p \in [0,1]$. Let θ be a parameter to be estimated. The set of values θ^* for which $R(\theta^*) \geq p$ is a $100p\%$ **likelihood region for** θ. If this set is a subset of $\mathbb{R}$, then it is the **likelihood interval** for θ.

Remark 5.5.4 Suppose Ω is the set of all possible values of θ, a parameter, then a likelihood interval for θ may not be an actual interval if R possesses multiple local minima and maxima.

■ **Example 5.5.5** Suppose $X_1,\ldots,X_n \sim \text{Poisson}(\theta)$ independently, and an observation has $\sum_{i=1}^{100} x_i = 980$, then

$$L(\theta) = \prod_{i=1}^{100} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{\theta^{\sum x_i} e^{-100\theta}}{\prod_{i=1}^{100} x_i!} = \frac{\theta^{980} e^{-100\theta}}{\prod_{i=1}^{100} x_i!}.$$

With $\tilde{\theta} = \bar{X}$, (Example 5.3.7), we have $\hat{\theta} = \frac{980}{100} = 9.8$ and

$$L(\hat{\theta}) = \frac{(9.8^{980})e^{-980}}{\prod_{i=1}^{100} x_i!}$$

and

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})} = \frac{\theta^{980}e^{-100\theta}}{(9.8^{980})e^{-980}} = e^{980-100\theta}\left(\frac{\theta}{9.8}\right)^{980}.$$

Calculating, say, the 50% and 10% likelihood interval can be left numerical methods, usually available in graphing calculators. ∎

5.6 Limiting Distribution of Maximum Likelihood Estimator

Remark 5.6.1 From Example 5.3.6 and Example 5.3.7 we note that as $n \to \infty$, the variance of the maximum likelihood estimator $\text{Var}(\tilde{\theta})$ approaches 0. There are other desirable properties of estimators which are related to their asymptotic behaviour as the size of the random sample approaches infinity.

Theorem 5.6.2 Suppose $X = (X_1, \ldots, X_n)$ is a random sample from a distribution with pdf $f : \mathbb{R} \to [0,1]$. Let $\tilde{\theta}_n = \tilde{\theta}(X_1, \ldots, X_n)$ be the maximum likelihood estimator of θ, then under certain regularity conditions (see Remark 5.6.3), we have

1. (consistency)

$$\tilde{\theta}_n \xrightarrow{P} \theta.$$

2. (asymptotic normality)

$$J(\theta)^{\frac{1}{2}}(\tilde{\theta}_n - \theta) \xrightarrow{D} Z \sim N(0,1).$$

3. (asymptotic distribution of relative likelihood)

$$-2\log(R(\theta;X)) = 2(\ell(\tilde{\theta};X) - \ell(\theta;X)) \xrightarrow{D} W \sim \chi^2_1.$$

Proof. We omit the proof. ∎

Remark 5.6.3 The regularity conditions in the above theorem include the following:

1. The pdf's are distinct with respect to θ, i.e. if $\theta_1 \neq \theta_2$ then $f(x;\theta_1) \neq f(x;\theta_2)$.
2. The pdf's have common support for all θ.
3. The true value of the parameter θ is an interior point of Ω i.e. there exists $\delta > 0$ such that the open ball centred at θ of radius δ is contained in Ω.

Remark 5.6.4 Thm. 5.6.2(1) implies that the MLE is consistent (see Def. 5.1.6) under the regularity conditions above.
Thm. 5.6.2(2) implies, along with the Central Limit Theorem, that

$$\tilde{\theta}_n \xrightarrow{D} N\left(\theta, \frac{1}{J(\theta)}\right)$$

and hence $\text{Var}(\tilde{\theta}_n) \to \frac{1}{J(\theta)} = J(\theta)^{-1}$ and $E(\tilde{\theta}_n) = \theta$. By Def. 5.1.5, the MLE is asymptotically unbiased.

However, $J(\theta)$ is unknown because θ is unknown, but note that

$$\tilde{\theta}_n \xrightarrow{P} \theta \Rightarrow \sqrt{J(\tilde{\theta}_n)} \xrightarrow{P} \sqrt{J(\theta)}$$

and $\tilde{\theta}_n - \theta \xrightarrow{D} N(0, J(\theta)^{-1})$, so by Slutsky's Thm, we have

$$(\tilde{\theta}_n - \theta)\sqrt{J(\tilde{\theta}_n)} \xrightarrow{D} \sqrt{J(\theta)}N(0, J(\theta)^{-1}) = N(0,1).$$

Therefore

$$\tilde{\theta}_n \xrightarrow{D} N(\theta, J(\tilde{\theta}_n)^{-1})$$

and consequently

$$\text{Var}(\tilde{\theta}_n) \to J(\tilde{\theta}_n)^{-1}.$$

We can also use the information function to estimate the variance of the MLE.

Proposition 5.6.5 Let $Y_n = (X_1, \ldots, X_n)$ be a random sample of size n where each X_i has probability function $f(x;\theta)$ and $\tilde{\theta}_n = \tilde{\theta}(Y_n)$ be the MLE based on Y_n, then under certain regularity conditions, we have

$$\text{Var}(\tilde{\theta}_n) \to \frac{1}{I(\hat{\theta}_n)}$$

as $n \to \infty$.

Proof. By the Weak Law of Large Numbers,

$$\frac{1}{n} I(\theta; Y_n) = \frac{-1}{n} \sum_{i=1}^{n} \frac{\partial^2}{\partial \theta^2} \ell(\theta; X_i) \xrightarrow{P} E\left(-\frac{\partial^2}{\partial \theta^2} \ell(\theta; X_i) \right)$$

and consequently, by Thm. 5.6.2 and Proposition 4.4.5,

$$\sqrt{I(\tilde{\theta}_n; Y_n)} \xrightarrow{P} \sqrt{J(\theta; Y_n)}$$

and by Slutsky's Thm,

$$\sqrt{I(\tilde{\theta}_n; Y_n)}(\tilde{\theta}_n - \theta) \xrightarrow{D} \sqrt{J(\theta)}N(0, J(\theta)^{-1})$$

and

$$\sqrt{I(\tilde{\theta}_n; Y_n)}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0,1).$$

Thus

$$\tilde{\theta}_n \xrightarrow{D} N\left(\theta, \frac{1}{I(\tilde{\theta}_n; Y_n)}\right)$$

and

$$\text{Var}(\tilde{\theta}_n) \to \frac{1}{I(\hat{\theta}_n)}$$

as required. ■

■ **Example 5.6.6** Suppose $X_1,\ldots,X_n \sim \text{Weibull}(\theta,2)$ independently, i.e. each X_i has pdf

$$f(x;\theta) = \frac{2}{\theta^2} x e^{-\left(\frac{x}{\theta}\right)^2}, x>0, \theta>0,$$

then the likelihood function for θ is

$$L(\theta) = \prod_{i=1}^n f(x;\theta) = \frac{2^n}{\theta^{2n}} \left(\prod_{i=1}^n x_i\right) e^{-\frac{1}{\theta^2}\sum_{i=1}^n x_i^2}, \theta>0$$

and the log likelihood function is

$$\ell(\theta) = n\log 2 - 2n\log\theta + \sum_{i=1}^n \log x_i - \frac{1}{\theta^2}\sum_{i=1}^n x_i^2, \theta>0.$$

Setting $\ell'(\theta)=0$ yields

$$0 = \frac{-2n}{\theta} + \frac{2}{\theta^3}\sum_{i=1}^n x_i^2.$$

and so this gives the MLE

$$\tilde{\theta} = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}.$$

By Thm. 5.6.2, $\tilde{\theta}$ is consistent. By the Weak Law of Large Numbers

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} E(X_i^2)$$

where $X_i \sim \text{Weibull}(\beta,2)$.

Now for any $X \sim \text{Weibull}(\alpha,\beta)$, $\alpha,\beta>0$,

$$E(X^k) = \int_0^\infty x^k \frac{\beta}{\alpha^\beta} x^{\beta-1} e^{-\left(\frac{x}{\alpha}\right)^\beta} dx.$$

Let $y = \left(\frac{x}{\alpha}\right)^\beta$, and consequently $x = \alpha y^{\frac{1}{\beta}}$ and $\frac{dx}{dy} = \frac{\alpha}{\beta} y^{\frac{1}{\beta}-1}$. Integration by substitution yields

$$\begin{aligned} E(X^k) &= \frac{\beta}{\alpha^\beta}\int_0^\infty \alpha^{k+\beta-1} y^{\frac{k}{\beta}} y^{1-\frac{1}{\beta}} e^{-y} \frac{\alpha}{\beta} y^{\frac{1}{\beta}-1} dy \\ &= \alpha^k \int_0^\infty y^{\frac{k}{\beta}} e^{-y} dy \\ &= \alpha^k \Gamma\left(\frac{k}{\beta}+1\right) \end{aligned}$$

and so

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} \theta^2\Gamma(2) = \theta^2.$$

Hence

$$\tilde{\theta}_n = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n}} \xrightarrow{P} 0$$

by the fact that the square root is a continuous function, and $\tilde{\theta}_n$ is consistent.

We can also verify asymptotic normality. The information function (Def. 5.3.2) of θ is

$$I(\theta) = -\frac{\partial^2}{\partial\theta^2}\ell(\theta) = -\frac{2n}{\theta^2} + \frac{6\sum_{i=1}^n x_i^2}{\theta^4}$$

which gives the Fisher information (Def. 5.3.4) of θ to be

$$J(\theta) = E(I(\theta)) = \frac{-2n}{\theta^2} + \frac{6}{\theta^4}E\left(\sum_{i=1}^n X_i^2\right) = \frac{-2n}{\theta^2} + \frac{6n}{\theta^4}E(X_i^2) = \frac{-2n}{\theta^2} + \frac{6n}{\theta^4}\theta^2 = \frac{4n}{\theta^2}.$$

To show that

$$J(\theta)^{\frac{1}{2}}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0,1)$$

we first note that, by the Central Limit Theorem, we have

$$\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i^2 - E(X_i^2)\right)}{\mathrm{Var}(X_i^2)} \xrightarrow{D} N(0,1)$$

where $\mathrm{Var}(X_i^2) = E(X_i^4) - E(X_i^2)^2 = \theta^4\Gamma(3) - (\theta^2)^2 = \theta^4$ and $E(X_i^2) = \theta^2$. Thus

$$\frac{\sqrt{n}(\tilde{\theta}_n^2 - \theta^2)}{\theta^2} \xrightarrow{D} N(0,1).$$

By the δ method, with $g : a \mapsto \sqrt{a}$ and $a = \theta^2$, we have

$$\frac{\sqrt{n}(\tilde{\theta}_n - \theta)}{\theta^2} \xrightarrow{D} \frac{1}{2\theta}N(0,1)$$

i.e.

$$\frac{2\sqrt{n}}{\theta}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0,1)$$

where $\frac{2\sqrt{n}}{\theta} = J(\theta)^{\frac{1}{2}}$. This proves the asymptotic normality of $\tilde{\theta}_n$.

Finally, we can show that

$$I(\tilde{\theta}_n; X)^{\frac{1}{2}}(\tilde{\theta}_n - \theta) \xrightarrow{D} N(0,1).$$

First observe that

$$\frac{I(\tilde{\theta})}{J(\theta)} = \frac{I\left(\sqrt{\frac{\sum_{i=1}^n X_i^2}{n}}\right)}{4n/\theta^2} = \frac{\theta^2\left(\frac{-2n}{\sum_{i=1}^n X_i^2/n} + \frac{6\sum_{i=1}^n X_i^2}{\left(\sum_{i=1}^n X_i^2/n\right)^2}\right)}{4n} = \frac{\theta^2\frac{4n^2}{\sum_{i=1}^n X_i^2}}{4n} = \theta^2\frac{n}{\sum_{i=1}^n X_i^2} = \theta^2\tilde{\theta}^{-2}.$$

Since $\tilde{\theta} \xrightarrow{P} \theta$ and $g(x) = x^{-2}$ is continuous,

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

Thus

$$\sqrt{I(\tilde{\theta}_n)}(\tilde{\theta}_n - \theta) = \sqrt{\frac{I(\tilde{\theta}_n)}{J(\theta)}}\sqrt{J(\theta)}(\tilde{\theta}_n - \theta) \xrightarrow{D} 1 \cdot N(0,1) = N(0,1)$$

by Slutsky's Theorem as required. ∎

■ **Example 5.6.7** If $X_1,\dots,X_n \sim \text{Uniform}(0,\theta)$ independently, then by Example 5.3.10, the MLE based on random sample of size n is $\tilde{\theta}_n = X_{(n)} = \max\{X_1,\dots,X_n\}$. Although the regularity condition (1) in Remark 5.6.3 does not hold, $\tilde{\theta}_n$ is still consistent. Note that the cdf of $X_{(n)}$ is

$$F_n(x) = \Pr(X_{(n)} \le x) = \begin{cases} 0 & \text{if} \quad x \le 0 \\ \left(\int_0^x \frac{1}{\theta} dt\right)^n = \left(\frac{x}{\theta}\right)^n & \text{if} \quad x \in [0,\theta) \\ 1 & \text{if} \quad x \ge \theta \end{cases}$$

Thus as $n \to \infty$,

$$F_n \to F = \begin{cases} 0 & \text{if} \quad x < \theta \\ 1 & \text{if} \quad x \ge \theta \end{cases}$$

i.e. $X_{(n)} \to \theta$ stochastically and $X_{(n)} \xrightarrow{P} \theta$ by Thm. 4.2.3. Hence $\tilde{\theta}_n = X_{(n)}$ is consistent. ■

5.7 Interval Estimators

Definition 5.7.1 Suppose X is a rv with distribution that depends on θ. Suppose A, B : support$(X) \to \mathbb{R}$ are functions such that $A(y) \le B(y)$ for all y. Let x be the observed data of X, then the random interval

$$(A(X), B(X))$$

is an **interval estimator** for θ, and $(A(x), B(x))$ is an **interval estimate** for θ.

Definition 5.7.2 Suppose $X = (X_1,\dots,X_n)$ is a random vector whose distribution depends on θ. Let $A(X), B(X)$ be statistics. If

$$\Pr(A(X) \le \theta \le B(X)) = p$$

for some $p \in (0,1)$, then $[A(x), B(x)]$, where x is an observed instance of X, is a $100p\%$ **confidence interval** of θ.

Remark 5.7.3 Confidence intervals are not necessarily unique, but having pivotal quantities makes confidence intervals easy to construct.

Definition 5.7.4 Suppose X is a random vector whose distribution depends on θ. The function $Q(X;\theta)$ is **pivotal quantity** if the distribution of Q does not depend on θ.

Definition 5.7.5 If $Y_n = (X_1,\dots,X_n)$ is a random vector with distribution depending on θ, then the function $Q(Y_n;\theta)$ is an **asymptotic pivotal quantity** if the distribution of Q does not depend on θ as $n \to \infty$.

■ **Example 5.7.6** As demonstrated in Thm. 5.6.2 and the proof of Proposition 5.6.5, both

$$Q = (J(\tilde{\theta}_n))^{\frac{1}{2}}(\tilde{\theta}_n - \theta)$$

and

$$Q = (I(\tilde{\theta}_n))^{\frac{1}{2}}(\tilde{\theta}_n - \theta)$$

are asymptotic pivotal quantities as both of them converge to $N(0,1)$ in distribution. ■

■ **Example 5.7.7** Suppose $X = (X_1,\dots,X_n) \sim \text{Exp}(\theta)$, consider the quantity

$$Q(X;\theta) = \frac{2\sum_{i=1}^n X_i}{\theta}.$$

Note that

$$\sum_{i=1}^{n} X_i \sim \Gamma(n,\theta)$$

and has MGF

$$M(t) = (1-\theta t)^{-n}, t < \frac{1}{\theta}.$$

Hence the MGF of Q is

$$M_Q(t) = M\left(\frac{2}{\theta}t\right) = \left(1-\theta\frac{2}{\theta}t\right)^{-n}, \frac{2}{\theta} < \frac{1}{\theta}$$

i.e.

$$M_Q(t) = (1-2t)^{-n}, t < \frac{1}{2}$$

which is the MGF of χ^2_{2n}, hence $Q \sim \chi^2_{2n}$. Thus Q is a pivotal quantity. Its distribution does not depend on θ and is known, since n is known.

Fix $p \in (0,1)$. Let $a,b \in \mathbb{R}$ such that

$$\Pr(Q \le a) = \frac{1-p}{2} = \Pr(Q \ge b).$$

We then have

$$\Pr(a \le Q(X;\theta) \le b) = p$$

i.e.

$$\Pr\left(a \le \frac{2\sum_{i=1}^n X_i}{\theta} \le b\right) = p \Leftrightarrow \Pr\left(\frac{2\sum_{i=1}^n X_i}{b} \le \theta \le \frac{2\sum_{i=1}^n X_i}{a}\right) = p,$$

making

$$\left[\frac{2\sum_{i=1}^n X_i}{b}, \frac{2\sum_{i=1}^n X_i}{a}\right]$$

a $100p\%$ confidence interval estimator for θ. ∎

∎ **Example 5.7.8** Suppose $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$. By Proposition 3.3.8 and Thm. 3.3.10. Fix $p \in (0,1)$. We can construct confidence interval estimators for μ when σ^2 is unknown, and for σ when μ is unknown.

Suppose we are looking for a $100p\%$ confidence interval for μ. We have

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

and (by Thm. 3.3.10)

$$Q := \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Q is a pivotal quantity. Find $a \in \mathbb{R}$ such that $\Pr(Q \leq -a) = \frac{1-p}{2} = \Pr(Q \geq a)$ (the student-t distribution is symmetric, unlike the χ^2 distribution). It follows that

$$\begin{aligned}
&\Pr(-a \leq Q \leq a) = p \\
\Leftrightarrow &\Pr\left(-a \leq \frac{\bar{X}-\mu}{S/\sqrt{n}} \leq a\right) = p \\
\Leftrightarrow &\Pr\left(\bar{X} - \frac{aS}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{aS}{\sqrt{n}}\right) = p
\end{aligned}$$

making

$$\left[\bar{X} - \frac{aS}{\sqrt{n}}, \bar{X} + \frac{aS}{\sqrt{n}}\right]$$

a $100p\%$ confidence interval for μ.

For the confidence interval of σ^2, recall Proposition 3.3.8 and consider the quantity

$$R := \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

which is a pivotal quantity.

Find $a, b \in \mathbb{R}$ such that $\Pr(R \leq a) = \frac{1-p}{2}$ and $\Pr(R \geq b) = \frac{1-p}{2}$, then

$$\begin{aligned}
&\Pr(a \leq R \leq b) = p \\
\Leftrightarrow &\Pr\left(a \leq \frac{(n-1)S^2}{\sigma^2} \leq b\right) = p \\
\Leftrightarrow &\Pr\left(\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}\right) = p
\end{aligned}$$

making

$$\left[\sqrt{\frac{(n-1)S^2}{b}}, \sqrt{\frac{(n-1)S^2}{a}}\right]$$

a $100p\%$ confidence interval estimator for σ. ∎

Theorem 5.7.9 Suppose $X := (X_1, \ldots, X_n)$ is a random vector where each X_i has a distribution depending on parameter θ and has pdf $f : \mathbb{R} \to [0,1]$, and $\tilde{\theta}$ is the MLE of θ depending on X.

If θ is a location parameter, then

$$Q = \tilde{\theta} - \theta$$

is a pivotal quantity. If θ is a scale parameter, then

$$Q = \frac{\tilde{\theta}}{\theta}$$

is a pivotal quantity.

Location and scale parameters are defined in Def. 1.6.1

Proof. Clearly $\tilde{\theta}$ does not depend on θ. Suppose θ is a location parameter, then the cdf of $\tilde{\theta} - \theta$ is

$$\Pr(\tilde{\theta} - \theta \leq a) = \int_{-\infty}^{a} f(x;\theta)dx = \int_{-\infty}^{a} f_0(x-\theta)dx, a \in \mathbb{R}$$

where f_0 does not depend on θ by Def. 1.6.1.

Similarly if θ is a scale parameter, then the cdf of $\frac{\tilde{\theta}}{\theta}$ is

$$\Pr(\tilde{\theta}-\theta \leq a) = \int_{-\infty}^{a} f(x;\theta)dx = \int_{-\infty}^{a} \frac{1}{\theta} f_1\left(\frac{x}{\theta}\right) dx, a \in \mathbb{R}$$

where f_1 does not depend on θ.

So the distributions of Q does not depend on θ in both cases, making it a pivotal quantity. ■

■ **Example 5.7.10** Suppose $X_1,\ldots,X_n \sim \text{Exp}(\theta)$ independently. Fix $p \in (0,1)$. Note that, defining $f_1(x) = f(x;\theta=1)$ where f is the pdf of each $\text{Exp}(\theta)$, then

$$\frac{1}{\theta} f_1\left(\frac{x}{\theta}\right) = \frac{1}{\theta}\frac{1}{\theta_*} e^{-\frac{x}{\theta}/\theta_*}\Big|_{\theta_*=1} = \frac{1}{\theta} e^{-\frac{x}{\theta}} = f(x).$$

This shows that θ is a scale parameter. By the above theorem, $\tilde{\theta}/\theta$ is a pivotal quantity. By Example 5.4.3, $\tilde{\theta} = \bar{X}$. Now

$$Q := \frac{\tilde{\theta}}{\theta} = \frac{\sum_{i=1}^n X_i}{n\theta}$$

where $\sum_{i=1}^n X_i \sim \Gamma(n,\theta)$. We find the distribution of $\tilde{\theta}/\theta$ by its MGF

$$M_Q(t) = \left(1-\theta\frac{t}{n\theta}\right)^{-n}, \frac{t}{n\theta} < \frac{1}{n}$$

i.e.

$$M_Q(t) = \left(1-\frac{1}{n}t\right)^{-n}, t < \frac{1}{1/n}$$

which shows that $Q \sim \Gamma\left(n, \frac{1}{n}\right)$. Find $a,b \in \mathbb{R}$ such that

$$\Pr(a \leq Q \leq b) = p$$

and the $100p\%$ confidence interval estimator for θ is

$$\left[\frac{\bar{X}}{b}, \frac{\bar{X}}{a}\right].$$

■

■ **Example 5.7.11** Suppose $X_1,\ldots,X_n$ are identically independently distributed with pdf

$$f : x \mapsto e^{-(x-\theta)}\mathbf{1}_{x>\theta}, x \in \mathbb{R}.$$

θ is a location parameter since, for $x > \theta$,

$$f_0(x-\theta) := e^{-(x-\theta-\theta_*)}\mathbf{1}_{x-\theta>\theta_*}|_{\theta_*=0} = f(x).$$

So $\tilde{\theta}-\theta$ is a pivotal quantity. Now the likelihood function of θ given observations $x_1,\ldots,x_n$ is

$$L(\theta) = \prod_{i=1}^n e^{-(x_i-\theta)}\mathbf{1}_{x_i>\theta} = e^{n\theta}e^{-\sum_{i=1}^n x_i}\prod_{i=1}^n \mathbf{1}_{x_i>\theta}$$

which gives $\tilde{\theta} = X_{(1)}$ since $\prod_{i=1}^{n} \mathbf{1}_{x_i>\theta} \neq 0$ if and only if $x_{(1)} = \min_{1\leq i\leq n} x_i > \theta$.

Fix $p \in (0,1)$. To find the $100p\%$ confidence interval of θ, solve for a in the equation

$$\Pr(X_{(1)} - \theta \leq a) = \frac{1-p}{2}$$

where

$$\Pr(X_{(1)} - \theta \leq a) = \Pr(X_{(1)} \leq a+\theta) = 1 - \Pr(X_{(1)} \geq a+\theta) = 1 - \prod_{i=1}^{n} \Pr(X_i \geq a+\theta)$$

and each

$$\begin{aligned}
\Pr(X_i \geq a+\theta) &= \int_{a+\theta}^{\infty} f(x)dx \\
&= \int_{a+\theta}^{\infty} e^{-(x-\theta)} \cdot 1 dx \\
&= e^{\theta} \int_{a+\theta}^{\infty} e^{-x} dx \\
&= e^{\theta}(0 + e^{-(a+\theta)}) \\
&= e^{-a}
\end{aligned}$$

So $\Pr(X_{(1)} - \theta \leq a) = 1 - e^{-na}$ and setting $\Pr(X_{(1)} - \theta \leq a) = \frac{1-p}{2}$ yields

$$a = \frac{\log\left(\frac{1+p}{2}\right)}{-n}.$$

Similarly, solving for b in $\Pr(X_{(1)} - \theta \geq b) = \frac{1-p}{2}$ yields

$$e^{-nb} = \frac{1-p}{2} \Leftrightarrow b = \frac{\log\left(\frac{1-p}{2}\right)}{-n}.$$

Hence the $100p\%$ estimator for the confidence interval of θ is

$$\left[X_{(1)} - b, X_{(1)} + a\right] = \left[X_{(1)} - \frac{\log\left(\frac{1-p}{2}\right)}{-n}, X_{(1)} + \frac{\log\left(\frac{1+p}{2}\right)}{-n}\right].$$

Suppose we have a sample with $n = 20$, $x_{(1)} = 10$, then the 95% confidence interval estimate using the estimator above is

$$\left[10 - \frac{\log(0.025)}{-20}, 10 + \frac{\log(0.975)}{-20}\right] = [9.82, 10.0013].$$

■

■ **Example 5.7.12** Suppose $X_1, \ldots, X_n \sim \text{Poisson}(\theta)$ independently, then by the Central Limit Theorem,

$$\sqrt{n}(\bar{X}_n - \theta) \xrightarrow{D} N(0, \theta).$$

But by the Weak Law of Large Numbers, the continuity of the square root function, and Proposition 4.4.5,

$$\sqrt{\bar{X}_n} \xrightarrow{P} \sqrt{\theta}.$$

Thus by Slutsky's Theorem,

$$\frac{\sqrt{n}(\bar{X}_n-\theta)}{\sqrt{\bar{X}_n}} \xrightarrow{D} \frac{1}{\sqrt{\theta}}N(0,\theta) = N(0,1)$$

which shows the left-hand side to be an asymptotic pivotal quantity.

This can be used to construct an approximate confidence interval for θ. This estimate will become more and more accurate as n, the sample size, increases. ∎

5.8 Approximate Confidence Intervals

Remark 5.8.1 Def. 5.7.5 and Example 5.7.12 provide motivation to treat approximate confidence interval estimators in detail. In particular, besides using the Central Limit Theorem like we did in Example 5.7.12, we may exploit the asymptotic properties of the maximum likelihood estimator $\tilde{\theta}$.

Proposition 5.8.2 Suppose $X_1,\ldots,X_n$ form an identically, independently distributed sample of size n such that the distribution of each X_i depends on parameter θ. Let $\tilde{\theta}_n$ be the maximum likelihood estimator of θ based on n random sample points. Fix $p \in (0,1)$, then

$$\left[\tilde{\theta}_n - \frac{a}{\sqrt{J(\tilde{\theta}_n)}}, \tilde{\theta}_n + \frac{a}{\sqrt{J(\tilde{\theta}_n)}}\right]$$

is an approximate $100p\%$ confidence interval estimator for θ for some $a > 0$ ($J(\tilde{\theta})$ is Fisher information).

Proof. By the result of Remark 5.6.4 we have

$$\sqrt{J(\tilde{\theta}_n)}(\tilde{\theta}_n-\theta) \xrightarrow{D} N(0,1).$$

Since $N(0,1)$ is a symmetric distribution around 0, we may find $a > 0$ such that

$$\Pr(-a \le \sqrt{J(\tilde{\theta}_n)}(\tilde{\theta}_n-\theta) \le a) = p$$

which implies

$$\Pr\left(\frac{-a}{\sqrt{J(\tilde{\theta}_n)}} \le \tilde{\theta}_n-\theta \le \frac{a}{\sqrt{J(\tilde{\theta}_n)}}\right) = p \Leftrightarrow \Pr\left(\tilde{\theta}_n - \frac{a}{\sqrt{J(\tilde{\theta}_n)}} \le \theta \le \tilde{\theta}_n + \frac{a}{\sqrt{J(\tilde{\theta}_n)}}\right) = p$$

which completes the proof. ∎

Corollary 5.8.3 Suppose $X_1,\ldots X_n$ is a random sample that has parameter θ. Let $\tilde{\theta}_n$ denote the MLE of θ based on a sample of size n. Fix $p \in (0,1)$, then

$$\left[\tilde{\theta}_n - \frac{a}{\sqrt{I(\tilde{\theta}_n)}}, \tilde{\theta}_n + \frac{a}{\sqrt{I(\tilde{\theta}_n)}}\right]$$

is an approximate $100p\%$ confidence interval for θ for some $a > 0$ ($I(\tilde{\theta}_n)$ is the information function).

Proof. In the proof of Proposition 5.6.5, we have

$$I(\tilde{\theta}_n;X)(\tilde{\theta}_n-\theta) \xrightarrow{D} N(0,1).$$

The arguments analogous to the proof of Proposition 5.8.2 complete the proof. ■

Remark 5.8.4 Another method to construct approximate CI's is to make use of Thm. 5.6.2(3) and the relative likelihood interval. The following theorem and example illustrate the asymptotic relationship between relative likelihood intervals and confidence intervals as the random sample size n increases.

Theorem 5.8.5 Suppose $X_1,\ldots,X_n$ is a random sample that has distribution parameter θ. Fix $p \in (0,1)$. Suppose $a \in \mathbb{R}$ satisfies the equation

$$p = 2\Pr(Z \leq a) - 1$$

where $Z \sim N(0,1)$, then the interval

$$\left\{\theta \in \Omega : R(\theta) = \frac{L(\theta)}{L(\tilde{\theta})} \geq e^{-\frac{a^2}{2}}\right\}$$

is an approximate $100p\%$ confidence interval for θ.

Proof. By Theorem 5.6.2(3) we have

$$-\log(R(\theta)) \xrightarrow{D} \chi^2_1.$$

Denote W to be the random variable following the χ^2_1 distribution. We have

$$\Pr(R(\theta) \geq e^{-\frac{-a^2}{2}}) = \Pr\left(\log(R(\theta)) \geq \frac{-a^2}{2}\right) = \Pr(-2\log(R(\theta)) \leq a^2) \approx \Pr(W \leq a^2).$$

We know that the square of the standard normal random variable Z is distributed according to the random variable $W \sim \chi^2_1$, so by this fact and the symmetry of the standard normal distribution,

$$\Pr(W \leq a^2) = \Pr(-a \leq Z \leq a) = 2\Pr(Z \leq a) - 1.$$

This in turn implies

$$\Pr(R(\theta) \geq e^{-\frac{a^2}{2}}) \approx 2\Pr(Z \leq a) - 1 = p$$

making the set

$$\{\theta \in \Omega : R(\theta) \geq e^{-\frac{a^2}{2}}\}$$

an apprximate $100p\%$ confidence interval for θ. ■

■ **Example 5.8.6** Fix $p \in (0,1)$. This is how to find a $100p\%$ confidence interval for a distribution θ. Solve for a in the equation

$$p = 2\Pr(Z \leq a) - 1$$

i.e. $a = \Phi^{-1}\left(\frac{1+p}{2}\right)$ where Φ is the distribution function of the standard normal distribution. Then calculate

$$q := e^{-\frac{a^2}{2}}$$

to get a $100q\%$ relative likelihood interval for θ. By Thm. 5.8.5, the $100q\%$ approximate likelihood interval of θ is the $100p\%$ approximate confidence interval of θ.

To see this numerically, suppose $p = 0.95$, then $a = \Phi^{-1}(1.95/2) \approx 1.96$ and consequently

$$e^{-\frac{a^2}{2}} \approx 0.15.$$

Thus a 15% relative likelihood interval is an apprixmate 95% confidence interval. ∎

6. Multi-Parameter Maximum Likelihood Estimation

6.1 Likelihood and Related Functions

Remark 6.1.1 The methods in Chapter 5 naturally extends to the estimation of more than one parameter in a distribution. Below are some definitions that are analogous to the ones in Chapter 5.

Definition 6.1.2 Suppose X is a random vector with a distribution that depends on $\theta \in \Omega \subseteq \mathbb{R}^n$ for some $n \geq 1$. Suppose x is an observed value for X, then the **likelihood function** of θ based on x is

$$\begin{aligned} L_x : \Omega &\to [0,1] \\ (\theta_1, \ldots, \theta_n) =: \theta &\mapsto f(x;\theta) \end{aligned}$$

where f is the pdf of X.

The **maximum likelihood estimator**, **maximum likelihood estimate**, and the **log likelihood function** are defined in the analogous fashion as before (Def. 5.2.2).

Definition 6.1.3 Suppose X is a rv that has a distribution that depends on $(\theta_1, \ldots, \theta_n) =: \theta \in \Omega \subseteq \mathbb{R}^n$, then the **score vector of** θ is

$$S(\theta) = \left(\frac{\partial \ell(\theta)}{\partial \theta_1}, \ldots, \frac{\partial \ell(\theta)}{\partial \theta_n} \right)^T \in \mathbb{R}^n$$

where $\ell : \Omega \to \mathbb{R}$ is the log likelihood function of θ.

Remark 6.1.4 The MLE of multi-dimensional parameter θ is obtained by solving $\frac{\partial \ell(\theta)}{\partial \theta_i} = 0$ for each θ_i, $1 \leq i \leq n$. This is usually done with numerical methods, since each partial derivative may not be linear.

Definition 6.1.5 Suppose X has a distribution that depends on $\theta = (\theta_1, \ldots, \theta_n) \in \Omega \subseteq \mathbb{R}^n$, then

the **information matrix of** θ, $I(\theta) \in M_{n\times n}$, is given by

$$[I(\theta)]_{ij} = -\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta), 1 \le i \le j \le n.$$

$I(\hat{\theta})$, where $\hat{\theta}$ is the MLE of θ, is the **observed information matrix**.

Remark 6.1.6 It is clear from the definition that the information matrix of a multi-dimensional parameter is symmetric.

Definition 6.1.7 Suppose X has a distribution that depends on $\theta = (\theta_1, \ldots, \theta_n) \in \Omega \subseteq \mathbb{R}^n$, then the **expected information matrix**, or the **Fisher information matrix** of θ, $J(\theta) \in M_{n\times n}$, is given by

$$[J(\theta)]_{ij} = E\left(-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta;X)\right), 1 \le i \le j \le n.$$

Definition 6.1.8 Fix $p \in (0,1)$. Suppose $\theta \in \Omega \subseteq \mathbb{R}^n$ is a parameter for a distribution of rv X and R is the relative likelihood function for θ, then the $100p\%$ **likelihood region for** θ is

$$\left\{\theta \in \mathbb{R}^n : R(\theta) = \frac{L(\theta)}{L(\tilde{\theta})} \ge p\right\}.$$

Remark 6.1.9 When finding the maximum of the likelihood function, the analogous version of the second-derivative test in the 1-dimensional case (for estimating an unknown parameter) is checking that the matrix of the second derivatives H, the Hessian matrix, is negative definite when calculated at the MLE $\hat{\theta}$, i.e. for all $0 \neq a \in \mathbb{R}^n$,

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

where

$$[H(\hat{\theta})]_{ij} = \frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\hat{\theta}).$$

An easy way to check this is to verify that $\det(H) < 0$.

Since H is just the negative of the observed information matrix $I(\hat{\theta})$, $I(\hat{\theta})$ must be positive definite. A sufficient, but not necessary condition, is

$$\det(I(\hat{\theta})) > 0$$

or all the eigenvalues of $I(\hat{\theta})$ is positive.

Below, we state the extension of Thm. 5.6.2, Remark 5.6.4 and Proposition 5.6.5.

Theorem 6.1.10 Suppose $\theta \in \Omega \subseteq \mathbb{R}^m$ is a parameter for a random distribution that governs n-dimensional random vector $X := (X_1, \ldots, X_n)$. Let $\tilde{\theta}_n$ be the MLE of θ given an observed random sample x of size n, then under certain regularity conditions

1. (consistency) $\tilde{\theta}_n$ is consistent:

$$\tilde{\theta}_n \xrightarrow{P} \theta.$$

2. (Asymptotic normality)

$$(\tilde{\theta}_n - \theta)[J(\theta)]^{\frac{1}{2}} \xrightarrow{D} MVN(0, I_m)$$

where $0 \in \mathbb{R}^m$, I_m is the identity matrix in $M_{m\times m}$, and MVN is the Multi-Variate Normal distribution defined for X (see Remark 2.10.2).

3. $-2\log(R(\theta;X)) \xrightarrow{D} \chi^2_m$.

Proof. We omit the proof. ■

Proposition 6.1.11 Let $\theta \in \Omega \in \mathbb{R}^m$ be a parameter vector for a distribution that determines $X = (X_1,\ldots,X_n)$, then as $n \to \infty$,

$$\lim_{n\to\infty} \text{Var}(\tilde{\theta}_n) = J(\theta)^{-1}$$

where $\tilde{\theta}_n$ is the MLE of θ, $\text{Var}(\tilde{\theta}_n)$ is the $m \times m$ variance-covariance matrix of $\tilde{\theta}_n$, and $J(\theta)^{-1}$ is the inverse of the Fisher information matrix of θ.

Proof. We omit the proof. ■

■ **Example 6.1.12** Suppose we are dealing with $\theta \in \Omega \subseteq \mathbb{R}^2$, then

$$\lim_{n\to\infty} \begin{bmatrix} \text{Var}(\tilde{\theta}_{1,n}) & \text{Cov}(\tilde{\theta}_{1,n},\tilde{\theta}_{2,n}) \\ \text{Cov}(\tilde{\theta}_{1,n},\tilde{\theta}_{2,n}) & \text{Var}(\tilde{\theta}_{2,n}) \end{bmatrix} = J(\theta)^{-1}.$$

Note that the 2-dimensional random vector $\tilde{\theta}_n := (\tilde{\theta}_{1,n},\tilde{\theta}_{2,n})$ is the MLE of θ given a random sample of size n. ■

Proposition 6.1.13 Let $\theta \in \Omega \in \mathbb{R}^m$ be a parameter vector for a distribution that determines $X = (X_1,\ldots,X_n)$, then as $n \to \infty$,

$$(\tilde{\theta}_n - \theta)I(\theta) \xrightarrow{D} MVN(0, I_m)$$

and

$$\lim_{n\to\infty} \text{Var}(\tilde{\theta}_n) = I(\hat{\theta}_n)^{-1}$$

where $\tilde{\theta}_n$ is the MLE of θ, $\text{Var}(\tilde{\theta}_n)$ is the $m \times m$ variance-covariance matrix of $\tilde{\theta}_n$, and $I(\hat{\theta})^{-1}$ is the inverse of the observed information matrix of θ.

Proof. We omit the proof. ■

Theorem 6.1.14 If $\hat{\theta} \in \mathbb{R}^m$ is the MLE of $\theta \in \Omega \subseteq \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^m$ is a one-to-one function, then $g(\hat{\theta})$ is the MLE of $g(\theta)$.

Proof. We omit the proof. ■

6.2 An Example That Does Not Require Numerical Methods

■ **Example 6.2.1** Suppose $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ indenpendently. This is a distribution with 2 unknown parameters. The pdf for each X_i is

$$f(x;\mu,\sigma^2) = (2\pi)^{-\frac{1}{2}}(\sigma^2)^{-\frac{1}{2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, x \in \mathbb{R}$$

and the likelihood function given observations $x_1, \ldots, x_n$ is

$$L : \mathbb{R}^2 \to [0,1]$$
$$(\mu, \sigma) \mapsto \prod_{i=1}^{n} f(x; \mu, \sigma^2) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

with the log likelihood function being

$$\ell : \mathbb{R}^2 \to \mathbb{R}$$
$$(\mu, \sigma) \mapsto \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

For the purpose of maximizing the (log) likelihood function, we take first and second partial derivative of ℓ with respect to μ:

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

and

$$\frac{\partial^2 \ell}{\partial \mu^2} = \frac{-n}{\sigma^2}.$$

We also take partial derivative with respect to σ^2:

$$\frac{\partial \ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2(\sigma^2)^2},$$

$$\frac{\partial^2 \ell}{\partial (\sigma^2)^2} = \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{(\sigma^2)^3}.$$

Finally

$$\frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} = \frac{-2\sum_{i=1}^{n}(x_i-\mu)}{2(\sigma^2)^2} = \frac{-n}{(\sigma^2)^2}(\bar{x}-\mu).$$

Solving $\frac{\partial \ell}{\partial \mu} = 0$ and $\frac{\partial \ell}{\partial \sigma^2} = 0$ simultaneously yields

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

To ensure that these are indeed maximum likelihood estimates, we find the Hessian matrix with respect to ℓ:

$$H = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \mu^2} & \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ell}{\partial \mu \partial \sigma^2} & \frac{\partial^2}{\partial (\sigma^2)^2} \end{bmatrix} = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{n}{(\sigma^2)^2}(\bar{x}-\mu) \\ -\frac{n}{(\sigma^2)^2}(\bar{x}-\mu) & \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{(\sigma^2)^3} \end{bmatrix}$$

which is negative definite when we substitute in $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$. Moreover, at $\hat{\mu}$ and $\hat{\sigma}^2$, the observed information matrix is simply the negative of the Hessian matrix with the substitution:

$$I(\hat{\mu}, \hat{\sigma}^2) = -H|_{\mu=\hat{\mu}, \sigma=\hat{\sigma}} = \begin{bmatrix} \frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2(\hat{\sigma})^2} \end{bmatrix}$$

Consequently the Fisher information matrix of the unknown parameters (μ, σ^2) is

$$J(\mu,\sigma^2) = E(I(\tilde{\mu},\tilde{\sigma}^2)) = \begin{bmatrix} E\left(\frac{n}{\sigma^2}\right) & E\left(\frac{n}{\sigma^4}(\bar{X}-\mu)\right) \\ E\left(\frac{n}{\sigma^4}(\bar{X}-\mu)\right) & E\left(\frac{-n}{2\sigma^4} + \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^6}\right). \end{bmatrix}$$

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

$$\begin{aligned} E\left(\frac{-n}{2\sigma^4} + \frac{\sum_{i=1}^n (X_i-\mu)^2}{\sigma^6}\right) &= \frac{-n}{2\sigma^4} + \frac{1}{\sigma^6}E\left(\sum_{i=1}^n (X_i^2 - 2X_i\mu + \mu^2)\right) \\ &= \frac{-n}{\sigma^4} + \frac{1}{\sigma^6}\left(\sum_{i=1}^n E(X_i^2) - E(2n\bar{X}\mu) - E(n\mu^2)\right) \\ &= \frac{-n}{2\sigma^4} + \frac{1}{\sigma^6}\left(\sum_{i=1}^n (\sigma^2+\mu^2) - 2n\mu^2 + n\mu^2\right) \\ &= \frac{n}{2\sigma^4}. \end{aligned}$$

Hence the Fisher information matrix is

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

and its inverse is

$$J(\mu,\sigma^2)^{-1} = \frac{2\sigma^6}{n^2}\begin{bmatrix} \frac{n}{2\sigma^4} & 0 \\ 0 & \frac{n}{\sigma^2} \end{bmatrix} = \begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2\sigma^4}{n}. \end{bmatrix}$$

We can verify Remark 6.1.12 by calculating the variance-covariance matrix of $\tilde{\mu}$ and $\tilde{\sigma}^2$. We have

$$\mathrm{Var}(\tilde{\mu}) = \mathrm{Var}(\bar{X}) = \frac{1}{n}\mathrm{Var}(X_i) = \frac{1}{n}\sigma^2$$

and

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

since $\bar{X} \perp \sum_{i=1}^n (X_i-\bar{X})^2$ by Lemma 3.3.7. Finally

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

where $S^2 = \frac{\sum_{i=1}^n (X_i-\bar{X})^2}{n-1}$ is the sample variance.

Since each $X_i \sim N(\mu,\sigma^2)$, we recall the conclusion of Proposition 3.3.8:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

and hence

$$\mathrm{Var}\left(\frac{(n-1)S^2}{\sigma^2}\right) = 2(n-1) = \frac{1}{\sigma^4}\mathrm{Var}((n-1)S^2)$$

which yields

$$\frac{1}{n^2}\operatorname{Var}((n-1)S^2) = \frac{1}{n^2}2(n-1)\sigma^4 = \operatorname{Var}(\tilde{\sigma}^2).$$

So the variance-covariance matrix of $\tilde{\mu}$ and $\tilde{\sigma}^2$ is

$$\begin{bmatrix} \frac{\sigma^2}{n} & 0 \\ 0 & \frac{2(n-1)\sigma^4}{n^2} \end{bmatrix}$$

which, when $n \to \infty$, converges to

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

■

7. Hypothesis Testing

7.1 Hypothesis Testing Introduction

Definition 7.1.1 A **statistical hypothesis** is a statement about the parameters of the distribution of a random variable that governs a population.

Definition 7.1.2 A **test of hypothesis** if the procedure to check the validity of a statistical hypothesis based on observed data from a population.

Remark 7.1.3 A hypothesis is usually formulated as

$$H_0 : \theta \in \Omega_0$$

where $\Omega_0 \subseteq \Omega$ and a distribution with pdf $f(x;\theta)$ is proposed for the rv that governs a population.

The alternative hypothesis to H_0 is

$$H_a : \theta \notin \Omega_0$$

or

$$H_a : \theta \in \Omega_0^c := \mathbb{R}^m \setminus \Omega_0$$

where m is the number of unknown parameters in the parameter vector θ.

Definition 7.1.4 A **test statistic** or **discrepancy measure** is a pivotal quantity based on unknown parameter vector θ that evaluates the consistency of observed data under H_0.

Definition 7.1.5 Given a population, a set of observed data, a test statistic, and a hypothesis H_0, the **p-value** is the probability of observing the observed data or something more extreme if H_0 is true.

Definition 7.1.6 Given a population and a hypothesis H_0, the **Type I error** of a hypothesis test is the probability of rejecting H_0 when H_0 is in fact true.

The **Type II error**, denoted β, is the probability of accepting H_0 when H_0 is in fact false.
The **power** of a hypothesis test is the probability of rejecting H_0 when it is indeed false.

Remark 7.1.7 Type I error of a test is, by definition, equal to the p-value of the test.

β is generally harder to compute than the p-value. β may also depend on θ and the size of the observed data, or the sample size.

Moreover, it is immediate from the definition above that the power of a test is $1-\beta$.

Definition 7.1.8 A **simple hypothesis test** is a hypothesis test with hypothesis $H_0 : \theta = \theta_0$ for some $\theta_0 \in \Omega$, i.e. if H_0 is true, then the population distribution is fully known.

■ **Example 7.1.9** Suppose $X_1, \ldots, X_{25} \sim N(\mu, 1)$ is a random sample from a population. Let $\mu_0 = 0$. Let $H_0 : \mu = \mu_0$ be a hypothesis. Suppose an observation has $\bar{x} = 0.5$. By definition, H_0 is a simple hypothesis.

A possible test statistic is

$$Z := \frac{\bar{X} - \mu_0}{1/\sqrt{25}}$$

which follows the standard normal distribution.

Given the known values from the observed sample, we have $z = (0.5-0)/(1/5) = 2.5$. The p-value is therefore

$$p = \Pr(-2.5 \leq Z \leq 2.5) = \Pr(|Z| \leq 2.5) = 2(1 - \Pr(Z \leq 2.5)) = 0.012.$$

which means the probability of observing the observed data or something more extreme is 0.012. ■

Remark 7.1.10 A test statistic like the one in the above example isn't always obviously to find, especially if we assume a more complicated distribution than the normal distribution for the population. Likelihood methods provide a way to find test statistics systematically.

7.2 Likelihood Ratio Tests for Simple Hypotheses

Definition 7.2.1 Given a population, a random sample $X = (X_1, \ldots, X_n)$, the MLE $\tilde{\theta}$, and a simple hypothesis $\theta = \theta_0 \in \Omega \subseteq \mathbb{R}^m$, where θ is a m-dimensional parameter for the distribution that governs the population, the **likelihood statistic** is

$$\Lambda(\theta_0) = -2\log(R(\theta_0; X))$$

where

$$\log(R(\theta_0; X)) = \log\left(\frac{L(\theta_0)}{L(\tilde{\theta})}\right) = \ell(\theta_0) - \ell(\tilde{\theta})$$

and

$$\Lambda(\theta_0) = 2(\ell(\tilde{\theta}) - \ell(\theta_0)).$$

Proposition 7.2.2 Given a population, a random sample $X = (X_1, \ldots, X_n)$, the MLE $\tilde{\theta}$ based on X, and a simple hypothesis $H_0 : \theta = \theta_0$, we have

$$\Lambda(\theta_0) \xrightarrow{D} \chi^2_k, k = n_1 - n_2,$$

where n_1 is the number of unknown parameters in the (assumed) distribution without H_0, and n_2 is the number of unknown parameters in the distribution with H_0.

Proof. We omit the proof. ■

Remark 7.2.3 Proposition 7.2.2 gives a way to compute an approximate p-value of a hypothesis $H_0 : \theta = \theta_0$ as long as the observed relative likelihood function at θ_0, written $\lambda(\theta_0)$, can be evaluated:

$$p = \Pr(W_k \geq \lambda(\theta_0)), W_k \sim \chi^2_k,$$

k as defined in Proposition 7.2.2.

■ **Example 7.2.4** Suppose $X := (X_1, \ldots, X_n)$ is a random sample from a population, and we assume a $N(\mu, 1)$ distribution for the population. Suppose we would like to test the hypothesis $H_0 : \mu = 0$, and we an observed sample of size 25 that has $\bar{x} = 0.5$.

We already know that the MLE of μ is $\tilde{\mu} = \bar{X}$. Consequently the likelihood functions of μ_0 and $\hat{\mu} = \bar{x}$ are

$$L(\mu_0) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}\sum_{i=1}^{n}(x_i - \mu_0)^2\right)$$

and

$$L(\hat{\mu}) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right).$$

This gives the log of the relative likelihood function

$$\begin{aligned}
\log(R(\mu_0)) &= \frac{-1}{2}\sum_{i=1}^{n}(x_i^2 - 2x_i\mu_0 + \mu_0^2) + \frac{1}{2}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{2}\sum_{i=1}^{n}(2x_i\mu_0 - \mu_0^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \frac{1}{2}(2\mu_0 n\bar{x} - n\mu_0^2 - 2n\bar{x}\bar{x} + n\bar{x}^2) \\
&= \frac{n}{2}(-\bar{x}^2 + 2\mu_0\bar{x} - \mu_0^2)
\end{aligned}$$

Thus the observed likelihood statistic in this case is

$$\begin{aligned}
\lambda(\mu_0 = 0) &= n(\bar{x}^2 - 2\mu\bar{x} - \mu_0^2) \\
&= n(\bar{x} - \mu_0)^2 \\
&= 25(0.5 - 0)^2 \\
&= 6.25
\end{aligned}$$

Finally, in this case, the number of unknown parameters without H_0 is 1 (just μ), and the number of unknown parameters with H_0 is 0, so

$$\Lambda(\mu) \xrightarrow{D} \chi^2_1.$$

This gives the p-value to be

$$p = \Pr(W_1 \geq 6.25) = 1 - \Pr(W_1 \leq 6.25) = 0.0124, W_1 \sim \chi^2_1.$$

Finally note that the test statistic $\Lambda(\mu = 0)$ in this case is in fact the square of the test statistic Z in Example 7.1.9. The p-values in these two examples are slightly different because in this example, the χ^2_1 is the asymptotic distribution of the test statistic. ■

Face(i)	1	2	3	4	5	6	**Total**
Frequency (f_i)	16	15	14	20	22	13	100

Table 7.1: Frequency Table

■ **Example 7.2.5** Suppose that the observed frequencies of the 6 faces of a die is as follows: We pose a hypothesis that the die is a fair die. To write this formally, assume

$$f_1,\ldots,f_6 \sim \text{MUL}(100;\theta_1,\ldots,\theta_6).$$

and write the hypothesis as

$$H_0 : \theta_1 = \cdots = \theta_6 = \frac{1}{6}.$$

The maximum likelihood estimates are

$$\hat{\theta}_1 = 16/100$$
$$\hat{\theta}_2 = 15/100$$
$$\hat{\theta}_3 = 14/100$$
$$\hat{\theta}_4 = 20/100$$
$$\hat{\theta}_5 = 22/100$$
$$\hat{\theta}_6 = 13/100$$

i.e. $\hat{\theta}_i = f_i/100$.

The likelihood function of $\theta_1,\ldots,\theta_6$ is

$$L(\theta_1,\ldots,\theta_6) = \frac{100!}{\prod_{i=1}^6 f_i!}\prod_{i=1}^6\left(\frac{1}{6}\right)^{f_i}$$

and the likelihood function of the corresponding MLEs are

$$L(\hat{\theta}_1,\ldots,\hat{\theta}_6) = \frac{100!}{\prod_{i=1}^6 f_i!}\left(\frac{16}{100}\right)^1 6\ldots\left(\frac{13}{100}\right)^{13}$$

which gives

$$\log(R(\theta_1,\ldots,\theta_6)) = \log\left(\frac{\prod_{i=1}^6\left(\frac{1}{6}\right)^{f_i}}{\prod_{i=1}^6\left(\frac{f_i}{100}\right)^{f_i}}\right) = -1.8498.$$

and the observed likelihood test statistic

$$\lambda(\theta_1,\ldots,\theta_6) = -2(-1.8498) = 3.70.$$

Now, the number of unknown parameters without H_0 is 5, and the number of unknown parameters with H_0 is 0, so by Proposition 7.2.2, the p-value is

$$p = \Pr(W_5 \geq 3.70) = 0.59, W_5 \sim \chi^2_5.$$

To interpret: assuming that the die is a fair die, there is roughly a 60% probability of observing the data in the table above. ■

■ **Example 7.2.6** Suppose $X_1,\dots,X_n \sim N(\mu_0,\sigma^2)$ independently where $\mu_0 \in \mathbb{R}$ is a known parameter. Suppose we want to test the hypothesis $H_0 : \sigma = \sigma_0$ for some fixed $\sigma_0 \in \mathbb{R}$.

Recall that the MLE of σ^2 is

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^n (x_i-\mu_0)^2}{n}$$

and so the relative likelihood function evaluated at σ_0 is

$$\begin{aligned} R(\mu_0,\sigma_0^2) &= \frac{L(\mu_0,\sigma_0^2)}{L(\mu_0,\hat{\sigma}^2)} = \frac{(\sigma_0^2)^{\frac{n}{2}}\exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\mu_0)^2\right)}{(\hat{\sigma}^2)^{\frac{n}{2}}\exp\left(\frac{-1}{2\hat{\sigma}^2}\sum_{i=1}^n (x_i-\mu_0)^2\right)} \\ &= \left(\frac{\hat{\sigma}^2}{\sigma_0^2}\right)^{\frac{n}{2}} \exp\left(-\frac{1}{2\sigma_0^2}\sum_{i=1}^n (x_i-\mu_0)^2 + \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^n (x_i-\mu_0)^2\right) \end{aligned}$$

Take the log of this value to get the observed likelihood test statistic $\lambda(\mu_0,\sigma_0^2)$ and find the p-value

$$p = \Pr(W_1 \geq \lambda(\mu_0,\sigma_0^2)), W_1 \sim \chi_1^2,$$

for the hypothesis $\sigma = \sigma_0$. ■

7.3 Likelihood Ratio Tests for Composite Hypotheses

Remark 7.3.1 Suppose $X_1,\dots,X_n$ is a random sample that follows an (assumed) distribution dependent on parameter $\theta \in \Omega \subseteq \mathbb{R}^m$, where Ω is an open set.

Suppose we would like to test the hypothesis $H_0 : \theta \in \Omega_0$, where $\Omega_0 \subseteq \mathbb{R}^k$, $k \leq m$ and Ω_0 is open. This is known as a **composite hypothesis** because it assumes that θ belongs to a region rather than equals a specific value.

The test statistic based on the random sample is

$$\Lambda = -2\log\left(\frac{\max_{\theta\in\Omega_0} L(\theta)}{\max_{\theta\in\Omega} L(\theta)}\right) = 2(\ell(\tilde{\theta}) - \max_{\theta\in\Omega_0}\ell(\theta))$$

where $\tilde{\theta}$ is the MLE based on the random sample.

The p-value of the hypothesis can be obtained by

$$p \approx \Pr(W \geq -2\log(\theta_0))$$

where $W \sim \chi^2_{m-k}$.

■ **Example 7.3.2** Suppose $X_1,\dots,X_n \sim \text{Exp}(\theta_1)$ independently, and $Y_1,\dots,Y_n \sim \text{Exp}(\theta_2)$ independently (indepedent from the X_i's as well). Suppose we have observed data of size 10, $\sum_{i=1}^{10} x_i = 15$, $\sum_{i=1}^{10} y_i = 20$, $\bar{x} = 1.5$, and $\bar{y} = 2$.

We can use the likelihood statistic to test the hypothesis $H_0 : \theta_1 = \theta_2$.

First, the joint likelihood function is

$$L(\theta_1,\theta_2) = \prod_{i=1}^n \frac{1}{\theta_1} e^{-\frac{x_i}{\theta_1}} \frac{1}{\theta_2} e^{-\frac{y_i}{\theta_2}} = \frac{1}{\theta_1^n\theta_2^n}\exp\left(-\frac{1}{\theta_1}\sum_{i=1}^n x_i - \frac{1}{\theta_2}\sum_{i=1}^n y_i\right)$$

with the log likelihood function

$$\ell(\theta_1,\theta_2) = -n\log(\theta_1) - n\log(\theta_2) - \frac{1}{\theta_1}\sum_{i=1}^n x_i - \frac{1}{\theta_2}\sum_{i=1}^n y_i.$$

We have

$$\frac{\partial \ell}{\partial \theta_1} = \frac{-n}{\theta_1} + \frac{1}{\theta_1^2}\sum_{i=1}^{n} x_i, \frac{\partial \ell}{\partial \theta_2} = \frac{-n}{\theta_2} + \frac{1}{\theta_2^2}\sum_{i=1}^{n} y_i$$

which yields the MLEs $\tilde{\theta}_1 = \bar{X}$ and $\tilde{\theta}_2 = \bar{Y}$. Moreover, under H_0, $\theta_1 = \theta_2$. As a consequence

$$L(\theta_1, \theta_2) = \frac{1}{\theta_1^{2n}} e^{-\frac{1}{\theta_1}(\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i)}$$

while

$$L(\hat{\theta}_1, \hat{\theta}_2) = \frac{1}{\bar{x}^n \bar{y}^n} e^{-\frac{1}{\bar{x}} n\bar{x} - \frac{1}{\bar{y}} n\bar{y}} = (\bar{x}\bar{y})^{-n} e^{-2n}.$$

To maximize $L(\theta_1, \theta_1)$, get its log:

$$\ell(\theta_1, \theta_1) = -2n\log(\theta_1) - \frac{1}{\theta_1}(n\bar{x} + n\bar{y})$$

and its derivative

$$\ell' = \frac{-2n}{\theta_1} + \frac{n}{\theta_1^2}(\bar{x} + \bar{y})$$

and consequently the MLEs under H_0 are $\tilde{\theta}_1 = \tilde{\theta}_2 = \frac{\bar{X}+\bar{Y}}{2}$. It follows that

$$\max\{L(\theta_1, \theta_1) : \theta_1 \in \Omega\} = \left(\frac{\bar{x}+\bar{y}}{2}\right)^{-2n} e^{-\frac{2}{\bar{x}+\bar{y}} n(\bar{x}+\bar{y})} = \left(\frac{\bar{x}+\bar{y}}{2}\right)^{-2n} e^{-2n}.$$

Hence the observed likelihood statistic under H_0 is

$$\begin{aligned}
\lambda(\theta_1, \theta_2) &= -2\log\left(\frac{\max\{L(\theta, \theta) : \theta \in \Omega\}}{(\bar{x}\bar{y})^{-n} e^{-2n}}\right) \\
&= -2\log\left(\frac{\left(\frac{\bar{x}+\bar{y}}{2}\right)^{-2n} e^{-2n}}{(\bar{x}\bar{y})^{-n} e^{-2n}}\right) \\
&= -2\log\left(\frac{\left(\frac{\bar{x}+\bar{y}}{2}\right)^{-2n}}{(\bar{x}\bar{y})^{-n}}\right) \\
&= -2\log\left(\frac{(3.5/2)^{-20}}{(2 \cdot 3.5)^{-10}}\right) \\
&= 0.41.
\end{aligned}$$

By Remark 7.3.1, the p-value of H_0 is given by

$$p = \Pr(W_1 \geq 0.41) = 0.52, W_1 \sim \chi^2_{1-0} = \chi^2_1.$$

■

8. Additional Exapmles

8.1 Estimation

■ **Example 8.1.1** Suppose $X_1, X_2 \sim \text{MUL}(n, \theta_1, \theta_2)$ for some $\theta_1, \theta_2 \in [0,1]$, then

$$\Pr(X_1 = x_1, X_2 = x_2) = \frac{n!}{x_1! x_2! (n - x_1 - x_2)!} \theta_1^{x_1} \theta_2^{x_2} (1 - \theta_1 - \theta_2)^{n - x_1 - x_2}, 0 \leq x_1 + x_2 \leq n, x_1, x_2 \geq 0.$$

The likelihood function for θ_1, θ_2 given observations x_1, x_2 are actually the same as the probability mass function. Consequently the log likelihood function is

$$\ell(\theta_1, \theta_2) = x_1 \log(\theta_1) + x_2 \log(\theta_2) + (n - x_1 - x_2) \log(1 - \theta_1 - \theta_2).$$

and

$$\frac{\partial \ell}{\partial \theta_i} = \frac{x_i}{\theta_i} - \frac{n - x_1 - x_2}{1 - \theta_1 - \theta_2}, i \in \{1, 2\},$$

$$\frac{\partial^2 \ell}{\partial \theta_i^2} = -\frac{x_i}{\theta_i^2} - \frac{x_1 + x_2 - n}{(1 - \theta_1 - \theta_2)^n}, i \in \{1, 2\},$$

and

$$\frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} = \frac{n - x_1 - x_2}{(1 - \theta_1 - \theta_2)^2}.$$

Solving $\frac{\partial \ell}{\partial \theta_1} = \frac{\partial \ell}{\partial \theta_2} = 0$ simultaneously yields

$$\frac{x_1}{\theta_1} = \frac{x_2}{\theta_2} = \frac{n - x_1 - x_2}{1 - \theta_1 - \theta_2}.$$

This in turn yields $\theta_1 + \theta_2 = \frac{x_1 + x_2}{n}$ and the MLEs

$$\hat{\theta}_1 = \frac{x_1}{n}, \hat{\theta}_2 = \frac{x_2}{n}.$$

Note that by the invariance property of the MLE, the MLE for $\frac{\theta_1}{\theta_1+\theta_2}$ is

$$\frac{\hat{\theta}_1}{\hat{\theta}_1+\hat{\theta}_2}=\frac{x_1}{x_1+x_2}.$$

The information matrix for (θ_1,θ_2) is

$$I(\theta_1,\theta_2)=\begin{bmatrix} \frac{X_1}{\theta_1^2}-\frac{n-X_1-X_2}{(1-\theta_1-\theta_2)^2} & \frac{n-X_1-X_2}{(1-\theta_1-\theta_2)^2} \\ \frac{n-X_1-X_2}{(1-\theta_1-\theta_2)^2} & \frac{X_2}{\theta_2^2}-\frac{n-X_1-X_2}{(1-\theta_1-\theta_2)^2} \end{bmatrix}$$

and the observed information matrix is

$$I(\hat{\theta}_1,\hat{\theta}_2)=\begin{bmatrix} n-\frac{n^2}{n-x_1-x_2} & \frac{n^2}{n-x_1-x_2} \\ \frac{n^2}{n-x_1-x_2} & n-\frac{n^2}{n-x_1-x_2}. \end{bmatrix}$$

The Fisher information matrix is

$$J(\theta_1,\theta_2)=E(I(\theta_1,\theta_2))=\begin{bmatrix} \frac{n}{\theta_1}-\frac{n}{1-\theta_1-\theta_2} & \frac{n}{1-\theta_1-\theta_2} \\ \frac{n}{1-\theta_1-\theta_2} & \frac{n}{\theta_2}-\frac{n}{1-\theta_1-\theta_2} \end{bmatrix}.$$

■

■ **Example 8.1.2** Suppose $p\in[0,1]$ is the portion of Canadians who have blue eyes, and we have a sample of 40, of which 5 people's eyes are blue. Suppose we assume a Bernoulli distribution for the population, which has unknown parameter p, we can then find a 90% confidence interval for p.

The random sample is written as $X_1,\ldots,X_{40}$. Each $\mathrm{Var}(X_1)=p(1-p)$ and $E(X_i)=p$.

By Central Limit Theorem, we have

$$\frac{\bar{X}-p}{\sqrt{\tilde{p}(1-\tilde{p})}/\sqrt{n}}\xrightarrow{D}N(0,1).$$

Thus, for some $a>0$,

$$0.9=\Pr\left(-a\le\frac{\bar{X}-p}{\sqrt{\tilde{p}(1-\tilde{p})/n}}\le a\right)=2\Pr\left(\frac{\bar{X}-p}{\sqrt{\tilde{p}(1-\tilde{p})/n}}\le a\right)-1.$$

An easy derivation yields the MLE for p is $\tilde{p}=\bar{X}$, which in this case is $\frac{5}{40}=\frac{1}{8}$.

Since, apprximately,

$$\frac{\bar{X}-p}{\sqrt{\tilde{p}(1-\tilde{p})/n}}=\frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})/n}}\sim N(0,1),$$

we have $a=\Phi^{-1}(1.9/2)=1.645$, Φ being the distribution function of the standard normal distribution. Now

$$\begin{aligned} 0.9&=\Pr\left(-1.645\le\frac{\frac{1}{8}-p}{\sqrt{\left(\frac{1}{8}\right)\left(\frac{7}{8}\right)/40}}\le 1.645\right)\\ &=\Pr\left(\frac{1}{8}-1.645\sqrt{\frac{7}{64\cdot 40}}\le p\le\frac{1}{8}+1.645\sqrt{\frac{7}{64\cdot 40}}\right)\\ &=\Pr(0.0390\le p\le 0.2110). \end{aligned}$$

Thus the approximate 90% CI for p is $[0.0390,0.2110]$. ■

■ **Example 8.1.3** Suppose $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$. If σ^2 is known, then

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

by the Central Limit Theorem and we can construct the $100p\%$ confidence interval for fixed $p \in (0,1)$ with

$$\bar{X} \pm a\sqrt{\frac{\sigma}{n}}, a = \Phi^{-1}\left(\frac{1+p}{2}\right),$$

where Φ is the distribution function of the standard normal distribution. The details are in Example 5.7.8.

If σ^2 is unknown, then using the fact that

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1},$$

proved in Thm. 3.3.10, we have

$$\left[\bar{X}-b\sqrt{\frac{s}{n}}, \bar{X}+b\sqrt{\frac{s}{n}}\right]$$

to be the $100p\%$ CI for μ, where $b = \Psi_{n-1}^{-1}\left(\frac{1+p}{2}\right)$, Ψ_{n-1} being the distribution function of t_{n-1}.

We can show that as $n \to \infty$, the end points of the CI's in these two cases converge to one point. We first show that $t_{n-1} \xrightarrow{P} N(0,1)$ by invoking Lemma 3.3.9:

$$t_{n-1} = \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}}$$

where $\chi^2_{n-1} = \sum_{i=1}^{n-1} \chi^2_1$ is a sum of distributions. By the Weak Law of Large Numbers,

$$\frac{\sum_{i=1}^{n-1} \chi^2_1}{n-1} \xrightarrow{P} E(\chi^2_1) = 1.$$

Since $Z \xrightarrow{P} Z$, by the continuity of the square root function and Proposition 4.4.6, we have

$$\frac{1}{\sqrt{\chi^2_{n-1}/(n-1)}} \xrightarrow{P} 1, \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}} \xrightarrow{P} Z,$$

and

$$t_{n-1} \xrightarrow{P} Z = N(0,1).$$

Now, S^2 is a consistent estimator of σ^2, so

$$S^2 \xrightarrow{P} \sigma^2$$

and

$$S \xrightarrow{P} \sigma \Rightarrow \frac{S}{\sigma} \xrightarrow{P} 1 \Rightarrow \sqrt{\frac{S}{\sigma}} \xrightarrow{P} 1.$$

Now

$$\frac{b\sqrt{\frac{s}{n}}}{a\sqrt{\frac{\sigma}{n}}} = \frac{b\sqrt{s}}{a\sqrt{\sigma}}$$

and it remains to show that $\lim_{n\to\infty} \frac{b}{a} = 1$. To do this, let $\varepsilon > 0$. Since $t_{n-1} \xrightarrow{P} N(0,1)$, we have

$$\lim_{n\to\infty} \Pr(|b-a| \le \varepsilon) = \lim_{n\to\infty} \Pr(|T-Z| \le \varepsilon) = 1, T \sim t_{n-1}, Z \sim N(0,1)$$

by definition of converging in probability, and hence,

$$\lim_{n\to\infty} \Pr\left(\left|\frac{b}{a} - 1\right| \le \varepsilon\right) = 1$$

and thus

$$\frac{b}{a} \cdot \frac{\sqrt{s}}{\sqrt{\sigma}} \xrightarrow{P} 1.$$

So the end points indeed get closer as $n \to \infty$. This shows that the $100p\%$ confidence interval gets narrower and narrower as the sample size increases. ∎

8.2 Assorted Examples

∎ **Example 8.2.1** Suppose $N \sim \text{Poisson}(\lambda)$ and a population $X_1, \ldots, X_n, \ldots$ are distributed identically and independently of each other and of N such that each

$$E(X_i) = \mu, \text{Var}(X_i) = \sigma^2.$$

Denote $S_N = \sum_{i=1}^{N} X_i$. We can use the Law of Total Expectation and the Law of Total Variance to find $E(S_N)$ and $\text{Var}(S_N)$.

By the Law of Total Expectation, $E(S_N) = E(E(S_N|N))$.

If N is observed, say $N = n$, then

$$E(S_N|N=n) = E\left(\sum_{i=1}^{n} X_i\right) = n\mu.$$

Hence $E(S_N|N) = N\mu$ and thus

$$E(S_N) = E(N\mu) = E(N)\mu = \lambda\mu.$$

For $\text{Var}(S_N)$, we have, by the Law of Total Variance,

$$\text{Var}(S_N) = E(\text{Var}(S_N|N)) + \text{Var}(E(S_N|N)).$$

The easier one out of the two terms above is

$$\text{Var}(E(S_N|N)) = \text{Var}(N\mu) = \mu^2 \text{Var}(N) = \mu^2\lambda.$$

For $E(\text{Var}(S_N|N))$, note that if N is observed at $N = n$, then

$$\text{Var}(S_N|N=n) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} 1^2 \text{Var}(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2.$$

Hence

$$E(\mathrm{Var}(S_N|N)) = E(N\sigma^2) = \sigma^2 E(N) = \sigma^2\lambda$$

and

$$\mathrm{Var}(S_N) = \sigma^2\lambda + \mu^2\lambda.$$

■

■ **Example 8.2.2** Suppose $X_1,\ldots,X_n$ are identically and independently distributed according to the pdf

$$f : x \mapsto e^{-(x-\theta)}, x > \theta > 0.$$

Clearly θ is a location parameter, so we can use $\tilde{\theta} - \theta$ as a pivotal quantity to construct confidence intervals for θ, but first we find $\tilde{\theta}$.

We have the likelihood function

$$L : \theta \mapsto \prod_{i=1}^{n} e^{-x_i+\theta}\mathbf{1}_{x_i\geq\theta}.$$

Thus the MLE for θ is

$$\tilde{\theta} = X_{(1)} = \min\{X_i : 1 \leq i \leq n\}.$$

For $x \in \mathbb{R}$, the cdf of $Q := \tilde{\theta} - \theta = X_{(1)} - \theta$ is

$$\begin{aligned}
F_Q(x) &= \Pr(X_{(1)} \leq \theta + x) = 1 - \Pr(X_{(1)} \geq \theta + x) \\
&= 1 - \Pr(X_1 \geq \theta + x)^n \\
&= 1 - \left(\int_{\theta+x}^{\infty} e^{-t+\theta} dt\right)^n \\
&= 1 - e^{n\theta}\left(-e^{-t}\big|_{t=\theta+x}^{\infty}\right)^n \\
&= 1 - e^{n\theta}(0 + e^{-\theta-x})^n \\
&= 1 - e^{nx}, x \in \mathbb{R}.
\end{aligned}$$

Now, if $p \in (0,1)$, the $100p\%$ confidence interval for θ is constructed by

$$p = \Pr(a \leq \tilde{\theta} - \theta \leq b) = \Pr(\tilde{\theta} - b \leq \theta \leq \tilde{\theta} + a).$$

For convenience, let $a = 0$, since confidence intervals are not unique anyway, and solve for b such that $p = \Pr(\tilde{\theta} - b \leq \theta \leq \tilde{\theta})$. Now

$$\Pr(\tilde{\theta} - b \leq \theta \leq \tilde{\theta}) = \Pr(0 \leq \tilde{\theta} - \theta \leq b) = F_Q(b) - F_Q(0) = 1 - e^{-nb}.$$

Hence, given $p \in (0,1)$, we have

$$b = \frac{\log(1-p)}{-n}$$

and

$$\left[X_{(1)} + \frac{\log(1-p)}{-n}, X_{(1)}\right]$$

to be the $100p\%$ confidence interval of θ. ■

■ **Example 8.2.3** Suppose $X_1, \ldots, X_n, \ldots$ is a sequence of random variables with each $E(X_i) = \mu$ and $\text{Var}(X_i) = \frac{a}{n^p}$ for some fixed $p > 0$.

We can show that $X_n \xrightarrow{P} \mu$, but note that we cannot use the Weak Law of Large Numbers or the Central Limit Theorem, because the variance of the X_i's vary as i changes. Instead, we use the definition of converging in probability (Def. 4.2.1) and Chebyshev's Inequality (Remark 1.8.5).

Let $\varepsilon > 0$, then by Remark 1.8.5 and the Squeeze Theorem,

$$0 \leq \Pr(|X_n - \mu| \leq \varepsilon) \leq \frac{a^2}{n^2 p \varepsilon^2} \to 0.$$

Thus

$$\lim_{n \to \infty} \Pr(|X_n - \mu| \geq \varepsilon) = 0$$

and $X_n \xrightarrow{P} \mu$ by definition. ■

Index

F

G

H

I

J

L

M

N

O

P

R

S

T

V

W

Z