A Master's Companion
Course Notes For
Math 502. Deterministic Mathematical Models
Math 503. Mathematical Statistics
Math 504. Numerical Methods
Math 611. Stochastic Simulation
Math 623. Sparse Representations and Random Sampling
Math 640. Bayesian Statistics
Math 658. Survey Sampling

Prepared by Sean Wilson

E-mail address: sdw62@georgetown.edu

Contents

Part 1

Background material

CHAPTER 1

Analysis

1.1. Upper bounds and suprema

This section is drawn from *Analysis: with an introduction to proof* (4th ed.) by Steven R. Lay.

THEOREM 1.1. *Let $x, y \in \mathbb{R}$ such that $x \leq y + \epsilon$ for every $\epsilon > 0$. Then $x \leq y$.*

PROOF. [proof goes here] □

DEFINITION 1.2. Let S be a subset of $\mathbb{R}$. If there exists a real number m such that $m \geq s$ for all $s \in S$, then m is called an *upper bound* for S, and we say that S is bounded above. If $m \leq s$ for all $s \in S$, then m is a *lower bound* for S and S is bounded below. The set S is said to be *bounded* if it is bounded above and bounded below.

If an upper bound m for S is a member of S, then m is called the *maximum* (or largest element) of S, and we write

$$m = \max S.$$

Similarly, if a lower bound of S is a member of S, then it is called the *minimum* (or least element) of S, denoted by $\min S$.

A set may have upper or lower bounds, or it may have neither. If m is an upper bound for S, then any number greater than m is also an upper bound. While a set may have many upper and lower bounds, if it has a maximum or a minimum, then those values are unique. Thus we speak of *an* upper bound and *the* maximum.

DEFINITION 1.3. Let S be a nonempty subset of $\mathbb{R}$. If S is bounded above, then the least upper bound of S is called its *supremum* and is denoted by $\sup S$. Thus $m = \sup S$ iff

(a) $m \geq s$, for all $s \in S$, and
(b) if $m' < m$, then there exists $s' \in S$ such that $s' > m'$.

If S is bounded below, then the greatest lower bound of S is called its *infimum* and is denoted by $\inf S$.

DEFINITION 1.4 (completeness axiom). Every nonempty subset S of $\mathbb{R}$ that is bounded above has a least upper bound. That is, $\sup S$ exists and is a real number.

THEOREM 1.5. *Given nonempty subsets A and B or $\mathbb{R}$, let C denote the set*

$$C = \{x + y : x \in A \text{ and } y \in B\}.$$

If A and B have suprema, then C has a supremum and

$$\sup C = \sup A + \sup B.$$

PROOF. Let $\sup A = a$ and $\sup B = b$. If $z \in C$, then $z = x + y$ for some $x \in A$ and $y \in B$. Thus $z = x + y \leq a + b$, so $a + b$ is an upper bound for C. By the completeness axiom, C has at least an upper bound, say $\sup C = c$. We must show that $c = a + b$. Since c is the *least* upper bound for C, we have $c \leq a + b$.

To see that $a + b \leq c$, choose any $\epsilon > 0$. Since $a = \sup A$, $a - \epsilon$ is not an upper bound for A, and there must exist $x \in A$ such that $a - \epsilon < x$. Similarly, since $b = \sup B$, there exists $y \in B$ such that $b - \epsilon < y$. Combining these inequalities, we have

$$a + b - 2\epsilon < x + y \leq c.$$

That is, $a+b < c+2\epsilon$ for every $\epsilon > 0$. Thus, by theorem 1.1, $a+b \leq c$. Finally, since $c \leq a+b$ and $c \geq a+b$, we conclude that $c = a+b$. □

CHAPTER 2

Probability theory

THEOREM 2.1. *If P is a probability function, then*

a. $P(A) = \sum_{i=1}^{\infty} P(A \cap C_i)$ *for any partition* $C_1, C_2, \ldots$;
b. $P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i)$ *for any sets* $A_1, A_2, \ldots$.

THEOREM 2.2 (Bayes' Rule). *Let $A_1, A_2, \ldots$ be a partition of the sample space, and let B be any set. Then, for each $i = 1, 2, \ldots$,*

$$P(A_i|B) = \frac{P(B|A_i)\,P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)\,P(A_j)}.$$

PROOF. From the definition of conditional probability, we have

$$\begin{aligned} P(A_i|B) &= \frac{P(A_i \cap B)}{P(B)} \\ \text{(conditional probability)} \qquad &= \frac{P(B|A_i)\,P(A_i)}{P(B)} \\ \text{(the } A_i \text{ partition the sample space)} \qquad &= \frac{P(B|A_i)\,P(A_i)}{\sum_{j=1}^{\infty} P(B \cap A_j)} \\ \text{(conditional probability)} \qquad &= \frac{P(B|A_i)}{\sum_{j=1}^{\infty} P(B|A_j)\,P(A_j)}. \end{aligned}$$

□

DEFINITION 2.3. The *cumulative distribution function* or *cdf* of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P_X(X \leq x), \quad \text{for all } x.$$

THEOREM 2.4. *The function $F(x)$ is a cdf if and only if the following three conditions hold:*

(a) $\lim_{x \to -\infty} F(x) = 0$ *and* $\lim_{x \to \infty} F(x) = 1$.
(b) $F(x)$ *is a nondecreasing function of* x.
(c) $F(x)$ *is right-continuous; that is, for every number* x_0, $\lim_{x \downarrow x_0} F(x) = F(x_0)$.

(This is Theorem 1.5.3 from Casella & Berger.)

PROOF. To prove necessity, we can write F in terms of the probability function. We have

$$\begin{aligned} \lim_{x \to -\infty} F(x) &= \lim_{x \to -\infty} P(\{X \leq x\}) \\ &= \end{aligned}$$

FINISH PROOF □

CHAPTER 3

Transformations and expectations

3.1. Distributions of functions of a random variable

When transformations are made, it is important to keep track of the sample spaces of the random variables; otherwise, much confusion can arise. When the transformation is from X to $Y = g(X)$, it is most convenient to use

$$\mathcal{X} = \{x : f_X(x) > 0\} \qquad \text{and} \qquad \mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}. \tag{3.1.1}$$

THEOREM 3.1. *Let X have cdf $F_X(x)$, let $Y = g(X)$, and let $\mathcal{X}$ and $\mathcal{Y}$ be defined as in (3.1.1).*

(1) *If g is an increasing function on $\mathcal{X}$, $F_Y(y) = F_X\left(g^{-1}(y)\right)$ for $y \in \mathcal{Y}$.*
(2) *If g is a decreasing function on $\mathcal{X}$ and X is a continuous random variable, $F_Y(y) = 1 - F_X\left(g^{-1}(y)\right)$ for $y \in \mathcal{Y}$.*

(This is Theorem 2.1.3 from Casella & Berger.)

PROOF. g is a monotone, i.e., it maps each x to a single y, and each y comes from at most one x. In this case that g is increasing, we have

$$\begin{aligned}\{x \in \mathcal{X} : g(x) \le y\} &= \left\{x \in \mathcal{X} : g^{-1}(g(x)) \le g^{-1}(y)\right\} \\ &= \left\{x \in \mathcal{X} : x \le g^{-1}(y)\right\},\end{aligned}$$

and the cdf of Y is

$$\begin{aligned}F_Y(y) &= P(\{Y \le y\}) \\ &= P(\{g(X) \le y\}) \\ &= P(\{x \in \mathcal{X} : g(x) \le y\}) \\ &= P\left(\left\{x \in \mathcal{X} : x \le g^{-1}(y)\right\}\right) \\ &= \int_{-\infty}^{g^{-1}(y)} f_X(x)\,\mathrm{d}x \\ &= F_X\left(g^{-1}(y)\right).\end{aligned}$$

In the case that g is decreasing, we have

$$\begin{aligned}\{x \in \mathcal{X} : g(x) \le y\} &= \left\{x \in \mathcal{X} : g^{-1}(g(x)) \ge g^{-1}(y)\right\} \\ &= \left\{x \in \mathcal{X} : x \ge g^{-1}(y)\right\},\end{aligned}$$

and the cdf of Y is

$$\begin{aligned}F_Y(y) &= P(\{Y \le y\}) \\ &= P(\{g(X) \le y\}) \\ &= P(\{x \in \mathcal{X} : g(x) \le y\}) \\ &= P\left(\left\{x \in \mathcal{X} : x \ge g^{-1}(y)\right\}\right) \\ &= \int_{g^{-1}(y)}^{\infty} f_X(x)\,\mathrm{d}x \\ &= 1 - F_X\left(g^{-1}(y)\right). \qquad \text{(continuity of } X\text{)}\end{aligned}$$

□

THEOREM 3.2. *Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let $\mathcal{X} = \{x : f_X(x) > 0\}$ and let $\mathcal{Y} = \{y : y = g(x), x \in \mathcal{X}\}$. Suppose that $f_X(x)$ is continuous on $\mathcal{X}$ and that $g^{-1}(y)$ has a continuous derivative on $\mathcal{Y}$. Then the pdf of Y is given by*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\dfrac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)\right|, & y \in \mathcal{Y} \\ 0, & \text{otherwise} \end{cases}.$$

(This is Theorem 2.1.5 from Casella & Berger.)

PROOF. From theorem 3.1 and applying the chain rule, we have

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) = f_X\left(g^{-1}(y)\right)\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$$

in the case that g is increasing and

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y}F_Y(y) = 0 - f_X\left(g^{-1}(y)\right)\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y) = -f_X\left(g^{-1}(y)\right)\frac{\mathrm{d}}{\mathrm{d}y}g^{-1}(y)$$

in the case that g is decreasing, which can be expressed concisely as in the theorem. □

We will look at F_X^{-1}, the inverse of the cdf F_X. If F_X is strictly increasing, then F_X^{-1} is well defined by

$$F_X^{-1}(y) = x \implies F_X(x) = y. \tag{3.1.2}$$

However, if F_X is constant on some interval, then F_X^{-1} is not well defined by (3.1.2). Any x satisfying $x_1 \le x \le x_2$ satisfies $F_X(x) = y$. This problem is avoided by defining $F_X^{-1}(y)$ for $0 < y < 1$ by

$$F_X^{-1}(y) = \inf\{x : F_X(x) \ge y\},$$

a definition that agrees with (3.1.2) when F_X is nonconstant and provides an F_X^{-1} that is single-valued even when F_X is not strictly increasing. Using this definition, for some interval (x_1, x_2) on which F_X is constant, we have $F_X^{-1}(y) = x_1$. At the endpoints of the range of y, $F_X^{-1}(y)$ can also be defined. $F_X^{-1}(1) = \infty$ if $F_X(x) < 1$ for all x and, for any F_X, $F_X^{-1}(0) = -\infty$.

THEOREM 3.3 (Probability integral transformation). *Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(X)$. Then Y is uniformly distributed on $(0,1)$, that is, $P(\{Y \le y\}) = y$, $0 < y < 1$. (This is Theorem 2.1.10 from Casella & Berger; the following proof is given there.)*

PROOF. For $Y = F_X(X)$ we have, for $0 < y < 1$,

$$\begin{aligned} P(\{Y \le y\}) &= P(\{F_X(X) \le y\}) \\ (F_X^{-1} \text{ is increasing}) \quad &= P\left(\left\{F_X^{-1}[F_X(X)] \le F_X^{-1}(y)\right\}\right) \\ (\text{see paragraph below}) \quad &= P\left(\left\{X \le F_X^{-1}(y)\right\}\right) \\ (\text{definition of } F_X) \quad &= F_X\left(F_X^{-1}(y)\right) \\ (\text{continuity of } F_X) \quad &= y. \end{aligned}$$

At the endpoints we have $P(\{Y \le y\}) = 1$ for $y \ge 1$ and $P(\{Y \le y\}) = 0$ for $y \le 0$, showing that Y has a uniform distribution.

The reasoning behind the equality

$$P\left(\left\{F_X^{-1}(F_X(X)) \le F_X^{-1}(y)\right\}\right) = P\left(\left\{X \le F_X^{-1}(y)\right\}\right)$$

is somewhat subtle and deserves additional attention. If F_X is strictly increasing, then it is true that $F_X^{-1}(F_X(x)) = x$. However, if F_X is flat, it may be that $F_X^{-1}(F_X(x)) \ne x$. Suppose F_X contains an interval (x_1, x_2) on which F_X is constant, and let $x \in [x_1, x_2]$. Then $F_X^{-1}(F_X(x)) = x_1$ for any x in this interval. Even in this case, though, the probability equality holds, since $P(\{X \le x\}) = P(\{X \le x_1\})$ for any $x \in [x_1, x_2]$. The flat cdf denotes a region of 0 probability ($P(\{x_1 < X \le x\}) = F_X(x) - F_X(x_1) = 0$). □

3.2. Expected values

THEOREM 3.4. *Let X be a random variable and let a, b, and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,*

(a) $\mathrm{E}[ag_1(X) + bg_2(X) + c] = a\,\mathrm{E}[g_1(X)] + b\,\mathrm{E}[g_2(X)] + c$.
(b) *If $g_1(x) \geq 0$ for all x, then $\mathrm{E}[g_1(X)] \geq 0$.*
(c) *If $g_1(x) \geq g_2(x)$ for all x, then $\mathrm{E}[g_1(X)] \geq \mathrm{E}[g_2(X)]$.*
(d) *If $a \leq g_1(x) \leq b$ for all x, then $a \leq \mathrm{E}[g_1(X)] \leq b$.*

(This is Theorem 2.2.5 from Casella & Berger; the following proof is given there.)

PROOF. [proof goes here] □

3.3. Moments and moment generating functions

DEFINITION 3.5. For each integer n, the nth *moment* of X or ($F_X(x)$), μ'_n, is

$$\mu'_n = \mathrm{E}[X^n].$$

The nth *central moment* of X, μ_n, is

$$\mu_n = \mathrm{E}[(X-\mu)^n],$$

where $\mu = \mu'_1 = \mathrm{E}[X]$.

DEFINITION 3.6. Let X be a random variable with cdf F_X. The *moment generating function* (mgf) of X (or F_X), denoted by $M_X(t)$, is

$$M_X(t) = \mathrm{E}\left[\mathrm{e}^{tX}\right],$$

provided that the expectation exists for t in some neighborhood of 0. That is, there is an $h > 0$ such that, for all t in $-h < t < h$, $\mathrm{E}\left[\mathrm{e}^{tX}\right]$ exists. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

More explicitly, we can write the mgf of X as

$$M_X(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}x \qquad \text{if } X \text{ is continuous,}$$

or

$$M_X(t) = \sum_x \mathrm{e}^{tx} P(\{X = x\}) \qquad \text{if } X \text{ is discrete.}$$

THEOREM 3.7. *If X has mgf $M_X(t)$, then*

$$\mathrm{E}[X^n] = M_X^{(n)}(0),$$

where we define

$$M_X^{(n)}(0) = \frac{\mathrm{d}^n}{\mathrm{d}\,t^n} M_X(t)\Big|_{t=0}.$$

That is, the nth moment is equal to the nth derivative of $M_X(t)$ evaluated at $t = 0$. (This is Theorem 2.3.7 from Casella & Berger; the following proof is given there.)

PROOF. Assuming that we can differentiate under the integral sign, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\,t} M_X(t) &= \frac{\mathrm{d}}{\mathrm{d}\,t} \int_{-\infty}^{\infty} \mathrm{e}^{tx} f_X(x)\,\mathrm{d}\,x \\
&= \int_{-\infty}^{\infty} \left(\frac{\mathrm{d}}{\mathrm{d}\,t}\mathrm{e}^{tx}\right) f_X(x)\,\mathrm{d}\,x \\
&= \int_{-\infty}^{\infty} \left(x\mathrm{e}^{tx}\right) f_X(x)\,\mathrm{d}\,x \\
&= \mathrm{E}\left[X\mathrm{e}^{tX}\right].
\end{aligned}$$

Thus,

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

Noting that

$$\frac{\mathrm{d}^n}{\mathrm{d}\,t^n}\mathrm{e}^{tx} = \frac{\mathrm{d}^{n-1}}{\mathrm{d}\,t^{n-1}}\left[\frac{\mathrm{d}}{\mathrm{d}\,t}\mathrm{e}^{tx}\right] = \frac{\mathrm{d}^{n-2}}{\mathrm{d}\,t^{n-2}}\left[\frac{\mathrm{d}}{\mathrm{d}\,t}x\mathrm{e}^{tx}\right] = \frac{\mathrm{d}^{n-2}}{\mathrm{d}\,t^{n-2}}x^2\mathrm{e}^{tx} = \frac{\mathrm{d}}{\mathrm{d}\,t}x^{n-1}\mathrm{e}^{tx} = x^n\mathrm{e}^{tx},$$

we can establish that

$$\begin{aligned}
\frac{\mathrm{d}^n}{\mathrm{d}\,t^n}M_X(t)\Big|_{t=0} &= \frac{\mathrm{d}^n}{\mathrm{d}\,t^n}\int_{-\infty}^{\infty}\mathrm{e}^{tx}f_X(x)\,\mathrm{d}\,x\Big|_{t=0} \\
&= \int_{-\infty}^{\infty}\left(\frac{\mathrm{d}^n}{\mathrm{d}\,t^n}\mathrm{e}^{tx}\right)f_X(x)\,\mathrm{d}\,x\Big|_{t=0} \\
&= \int_{-\infty}^{\infty}\left(x^n\mathrm{e}^{tx}\right)f_X(x)\,\mathrm{d}\,x\Big|_{t=0} \\
&= \mathrm{E}\left[X^n\mathrm{e}^{tX}\right]\Big|_{t=0} \\
&= \mathrm{E}\left[X^n\right].
\end{aligned}$$

□

LEMMA 3.8. *Let $a_1, a_2, \ldots$ be a sequence of numbers converging to a, that is, $\lim_{n\to\infty} a_n = a$. Then*

$$\lim_{n\to\infty}\left(1+\frac{a_n}{n}\right)^n = \mathrm{e}^a.$$

(This is Lemma 2.3.14 from Casella & Berger.)

THEOREM 3.9. *For any constants a and b, the mgf of the random variable $aX+b$ is given by*

$$M_{aX+b}(t) = \mathrm{e}^{bt}M_X(at).$$

(This is Theorem 2.3.15 from Casella & Berger; the following proof is given there.)

PROOF. By definition,

$$M_{aX+b}(t) = \mathrm{E}\left[\mathrm{e}^{(aX+b)t}\right] = \mathrm{E}\left[\mathrm{e}^{(aX)t}\mathrm{e}^{bt}\right] = \mathrm{e}^{bt}\,\mathrm{E}\left[\mathrm{e}^{(at)X}\right] = \mathrm{e}^{bt}M_X(at),$$

proving the theorem. □

CHAPTER 4

Multiple random variables

4.1. Conditional distributions and independence

DEFINITION 4.1. Let (X, Y) be a bivariate random vector with joint pdf or pmf $f(x, y)$ and marginal pdfs or pmfs $f_X(x)$ and $f_Y(y)$. Then X and Y are called *independent random variables* if, for every $x \in \mathbb{R}$ and $y \in \mathbb{R}$,

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

THEOREM 4.2. *Let X and Y be independent random variables.*

(a) *For any $A \subset \mathbb{R}$ and $B \subset \mathbb{R}$, $P(\{X \in A\} \cap \{Y \in B\}) = P(\{X \in A\}) P(\{Y \in B\})$; that is, the events $\{X \in A\}$ and $\{Y \in B\}$ are independent events.*

(b) *Let $g(x)$ be a function only of x and $h(y)$ be a function only of y. Then*

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

(This is Theorem 4.2.10 from Casella & Berger; the following proof is given there.)

PROOF. For continuous random variables, part (b) is proved by noting that

$$\begin{aligned}
\mathrm{E}[g(X) h(Y)] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) h(y) f(x, y) \,\mathrm{d}x \,\mathrm{d}y \\
\text{(independence)} \qquad &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x) h(y) f_X(x) f_Y(y) \,\mathrm{d}x \,\mathrm{d}y \\
&= \int_{-\infty}^{\infty} h(y) f_Y(y) \int_{-\infty}^{\infty} g(x) f_X(x) \,\mathrm{d}x \,\mathrm{d}y \\
&= \left(\int_{-\infty}^{\infty} g(x) f_X(x) \,\mathrm{d}x \right) \left(\int_{-\infty}^{\infty} h(y) f_Y(y) \,\mathrm{d}y \right) \\
&= \mathrm{E}[g(X)] \mathrm{E}[h(Y)].
\end{aligned}$$

The result for discrete random variables is proved by replacing integrals by sums. Part (a) can be proved by series of steps similar to those above or by the following argument. Let $g(x)$ be the indicator function of the set A. Let $h(y)$ be the indicator function of the set B. Note that $g(x) h(y)$ is the indicator function of the set $C \subset \mathbb{R}^2$ defined by $C = \{(x, y) : x \in A, y \in B\}$. Thus using the expectation equality just proved, we have

$$\begin{aligned}
P(\{X \in A\} \cap \{Y \in B\}) &= P(\{(X, Y) \in C\}) \\
&= \mathrm{E}[g(X) h(Y)] \\
&= \mathrm{E}[g(X)] \mathrm{E}[h(Y)] \\
&= P(\{X \in A\}) P(\{Y \in B\}).
\end{aligned}$$

□

THEOREM 4.3. *Let X and Y be independent random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Then the moment generating function of the random variable $Z = X + Y$ is given by*

$$M_Z(t) = M_X(t) M_Y(t).$$

(This is Theorem 4.2.12 from Casella & Berger; the following proof is given there.)

PROOF. Using the definition of the mgf and theorem 4.2, we have

$$M_Z(t) = \mathrm{E}\left[\mathrm{e}^{tZ}\right] = \mathrm{E}\left[\mathrm{e}^{t(X+Y)}\right] = \mathrm{E}\left[\mathrm{e}^{tX}\mathrm{e}^{tY}\right] = \mathrm{E}\left[\mathrm{e}^{tX}\right]\mathrm{E}\left[\mathrm{e}^{tY}\right] = M_X(t)\,M_Y(t).$$

□

4.2. Covariance and correlation

DEFINITION 4.4. The *covariance* of X and Y is the number defined by

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right].$$

DEFINITION 4.5. The *correlation* of X and Y is the number defined by

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}.$$

The value ρ_{XY} is also called the *correlation coefficient*.

THEOREM 4.6. *If X and Y are independent random variables, then $\mathrm{Cov}(X,Y) = 0$ and $\rho_{XY} = 0$. (This is Theorem 4.5.5 from Casella & Berger; the following proof is given there.)*

PROOF. Since X and Y are independent, we have $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$. Thus

$$\begin{aligned}\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] &= \mathrm{E}\left[XY - \mu_Y X - \mu_X Y + \mu_X\mu_Y\right]\\ &= \mathrm{E}[XY] - \mu_Y\,\mathrm{E}[X] - \mu_X\,\mathrm{E}[Y] + \mu_X\mu_Y\\ &= \mathrm{E}[X]\,\mathrm{E}[Y] - \mathrm{E}[Y]\,\mathrm{E}[X] - \mathrm{E}[X]\,\mathrm{E}[Y] + \mathrm{E}[X]\,\mathrm{E}[Y]\\ &= 0\end{aligned}$$

and

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y} = \frac{0}{\sigma_X\sigma_Y} = 0.$$

□

THEOREM 4.7. *If X and Y are any two random variables and a and b are any two constants, then*

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

If X and Y are independent random variables, then

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

(This is Theorem 4.5.6 from Casella & Berger; the following proof is given there.)

PROOF. We have

$$\begin{aligned}\mathrm{Var}(aX+bY) &= \mathrm{E}\left[((aX+bY) - \mathrm{E}[aX+bY])^2\right]\\ &= \mathrm{E}\left[(aX+bY-a\,\mathrm{E}[X]-b\,\mathrm{E}[Y])^2\right]\\ &= \mathrm{E}\left[(aX+bY-a\mu_X-b\mu_Y)^2\right]\\ &= \mathrm{E}\left[(a(X-\mu_X)+b(Y-\mu_Y))^2\right]\\ &= \mathrm{E}\left[(a(X-\mu_X))^2 - 2ab(X-\mu_X)(Y-\mu_Y) + (b(Y-\mu_Y))^2\right]\\ &= \mathrm{E}\left[a^2(X-\mu_X)^2\right] - \mathrm{E}\left[2ab(X-\mu_X)(Y-\mu_Y)\right] + \mathrm{E}\left[b^2(Y-\mu_Y)^2\right]\\ &= a^2\,\mathrm{E}\left[(X-\mu_X)^2\right] - 2ab\,\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right] + b^2\,\mathrm{E}\left[(Y-\mu_Y)^2\right]\\ &= a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y) - 2ab\,\mathrm{Cov}(X,Y).\end{aligned}$$

If X and Y are independent, it follows from theorem 4.6 that $\mathrm{Cov}(X,Y) = 0$ and the second equality is immediate from the first. □

THEOREM 4.8. *For any random variables X and Y,*

(a) $-1 \leq \rho_{XY} \leq 1$.
(b) $|\rho_{XY}| = 1$ *if and only if there exist numbers* $a \neq 0$ *and* b *such that* $P(\{Y = aX + b\}) = 1$. *If* $\rho_{XY} = 1$, *then* $a > 0$, *and if* $\rho_{XY} = -1$, *then* $a < 0$.

(This is Theorem 4.5.7 from Casella & Berger; the following proof is given there.)

PROOF. [proof goes here] □

4.3. Multivariate distributions

THEOREM 4.9. *Let* $X_1, \ldots, X_n$ *be mutually independent random variables with mgfs* $M_{X_1}(t), \ldots M_{X_n}(t)$. *Let* $Z = X_1 + \cdots + X_n$. *Then the mgf of* Z *is*

$$M_Z(t) = M_{X_1}(t) \cdot \cdots \cdot M_{X_n}(t).$$

In particular, if $X_1, \ldots, X_n$ *all have the same distribution with mgf* $M_X(t)$, *then*

$$M_Z(t) = (M_X(t))^n.$$

(This is Theorem 4.6.7 from Casella & Berger, which is a generalization of theorem 4.3).

4.4. Inequalities

LEMMA 4.10 (Young's Inequality). *Let* a *and* b *be any positive numbers, and let* p *and* q *be any positive numbers (necessarily greater than 1) satisfying*

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Then

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

with equality if and only if $a^p = b^q$. *(This is Lemma 4.7.1 from Casella & Berger; the following proof is given there).*

PROOF. Fix b, and consider the function

$$g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab.$$

To minimize $g(a)$, differentiate and set equal to 0:

$$\frac{\mathrm{d}}{\mathrm{d}\,a}g(a) = 0 \implies a^{p-1} - b = 0 \implies b = a^{p-1}.$$

We will evaluate the second derivative of $g(a)$ with respect to a at

$$b = a^{p-1} \implies b^{1/(p-1)} = \left(a^{p-1}\right)^{1/(p-1)} \implies a = b^{1/(p-1)}$$

to verify that this is a minimum.

$$\begin{aligned}
\frac{\mathrm{d}^2}{\mathrm{d}\,a^2}\left[g(a)\big|_{a=b^{1/(p-1)}}\right. &= \frac{\mathrm{d}}{\mathrm{d}\,a}\left[a^{p-1} - b\big|_{a=b^{1/(p-1)}}\right. \\
&= \left[(p-1)\,a^{p-2}\big|_{a=b^{1/(p-1)}}\right. \\
&= (p-1)\left(b^{1/(p-1)}\right)^{p-2} \\
&= (p-1)\,b^{(p-2)/(p-1)} \\
&= (p-1)\,b^{(p-1-1)/(p-1)} \\
&= (p-1)\,b^{[(p-1)/(p-1)]-1/(p-1)} \\
&= (p-1)\,b^{1-1/(p-1)}
\end{aligned}$$

We have $p > 1$ and $b > 0$, so that $p - 1 > 0$ and $b^{1-1/(p-1)} > 0$. It follows that $b = a^{p-1}$ is a minimum. We have

$$\frac{1}{p} + \frac{1}{q} = 1 \implies \frac{1}{q} = 1 - \frac{1}{p} = \frac{p-1}{p} \implies q(p-1) = p,$$

so that the value of $g(a)$ at the minimum is

$$\frac{1}{p}a^p + \frac{1}{q}\left(a^{p-1}\right)^q - aa^{p-1} = \frac{1}{p}a^p + \frac{1}{q}a^p - a^p = a^p\left(\frac{1}{p} + \frac{1}{q} - 1\right) = a^p(1-1) = 0.$$

Hence the minimum is 0 and the inequality is established. The domain of $g(a)$ is $\{a : 0 < a < \infty\}$ and we have $p > 1$, so that for some fixed b,

$$g'(a) = a^{p-1} - b$$

is increasing in a. Thus, the minimum we found is unique, so that equality holds only if $a^{p-1} = b$, which is equivalent to

$$a^{p-1} = b \implies a^{p/q} = b \implies \left(a^{p/q}\right)^q = b^q \implies a^p = b^q.$$

□

THEOREM 4.11 (Hölder's Inequality). *Let X and Y be any two random variables, and let p and q satisfy lemma 4.10. Then*

$$|\mathrm{E}[XY]| \le \mathrm{E}[|XY|] \le \left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}.$$

(This is Theorem 4.7.2 from Casella & Berger; the following proof is given there.)

PROOF. The first inequality follows from $-|XY| \le XY \le |XY|$ and theorem 3.4. To prove the second inequality, define

$$a = \frac{|X|}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}} \quad \text{and} \quad b = \frac{|Y|}{\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}}.$$

Applying lemma 4.10, we get

$$\begin{aligned}
&\frac{1}{p}a^p + \frac{1}{q}b^q \ge ab \\
\implies &\frac{1}{p}\left(\frac{|X|}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}}\right)^p + \frac{1}{q}\left(\frac{|Y|}{\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}}\right)^q \ge \frac{|X|\,|Y|}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}} \\
\implies &\frac{1}{p}\frac{|X|^p}{\mathrm{E}\left[|X|^p\right]} + \frac{1}{q}\frac{|Y|^q}{\mathrm{E}\left[|Y|^q\right]} \ge \frac{|XY|}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}}.
\end{aligned}$$

Taking the expectation of both sides gives

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{p}\frac{|X|^p}{\mathrm{E}\left[|X|^p\right]} + \frac{1}{q}\frac{|Y|^q}{\mathrm{E}\left[|Y|^q\right]}\right] \ge \mathrm{E}\left[\frac{|XY|}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}}\right] \\
\implies &\frac{1}{p\,\mathrm{E}\left[|X|^p\right]}\mathrm{E}\left[|X|^p\right] + \frac{1}{q\,\mathrm{E}\left[|Y|^q\right]}\mathrm{E}\left[|Y|^q\right] \ge \mathrm{E}\left[\frac{|XY|}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}}\right] \\
&\implies \frac{1}{p} + \frac{1}{q} \ge \frac{1}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}}\mathrm{E}\left[|XY|\right] \\
&\implies 1 \ge \frac{1}{\left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}}\mathrm{E}\left[|XY|\right] \\
&\implies \mathrm{E}\left[|XY|\right] \le \left(\mathrm{E}\left[|X|^p\right]\right)^{1/p}\left(\mathrm{E}\left[|Y|^q\right]\right)^{1/q}.
\end{aligned}$$

□

Perhaps the most famous special case of Hölder's Inequality is that for which $p = q = 2$.

THEOREM 4.12 (Cauchy-Schwarz Inequality). *For any two random variables X and Y,*

$$|\mathrm{E}[XY]| \leq \mathrm{E}[|XY|] \leq \left(\mathrm{E}\left[|X|^2\right]\right)^{1/2} \left(\mathrm{E}\left[|Y|^2\right]\right)^{1/2}.$$

(This is Theorem 4.7.3 from Casella & Berger.)

CHAPTER 5

Properties of a random sample

5.1. Sums of random variables from a random sample

DEFINITION 5.1. The *sample variance* is the statistic defined by

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

The *sample standard deviation* is the statistic defined by $S = \sqrt{S^2}$.

THEOREM 5.2. *Let $x_1, \ldots, x_n$ be any numbers and $\bar{x} = (x_1 + \cdots + x_n)/n$. Then*

(a) $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2$,
(b) $(n-1)s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$.

(This is Theorem 5.2.4 from Casella & Berger; the following proof is given there.)

PROOF. To prove part (a), add and subtract $\bar{x}$ to get

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - a)^2 &= \sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - a)^2 \\
&= \sum_{i=1}^{n}\left[(x_i - \bar{x}) + (\bar{x} - a)\right]\left[(x_i - \bar{x}) + (\bar{x} - a)\right] \\
&= \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 + (x_i - \bar{x})(\bar{x} - a) + (\bar{x} - a)(x_i - \bar{x}) + (\bar{x} - a)^2\right] \\
&= \sum_{i=1}^{n}\left[(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - a) + (\bar{x} - a)^2\right] \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}\left[2(x_i - \bar{x})(\bar{x} - a)\right] + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\sum_{i=1}^{n}\left(x_i\bar{x} - ax_i - \bar{x}^2 + a\bar{x}\right) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\left(\bar{x}\sum_{i=1}^{n}x_i - a\sum_{i=1}^{n}x_i - \sum_{i=1}^{n}\bar{x}^2 + \sum_{i=1}^{n}a\bar{x}\right) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\left(\bar{x}(n\bar{x}) - a(n\bar{x}) - n\bar{x}^2 + na\bar{x}\right) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2\left(n\bar{x}^2 - na\bar{x} - n\bar{x}^2 + na\bar{x}\right) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + 2(0) + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2.
\end{aligned}$$

It is now clear that the right-hand side is minimized at $a = \bar{x}$. To prove part (b), take $a = 0$ in the above, i.e.,

$$\begin{aligned}
&\sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - a)^2 \\
\implies &\sum_{i=1}^{n}(x_i - 0)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}(\bar{x} - 0)^2 \\
\implies &\sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{n}\bar{x}^2 \\
\implies &\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
\implies &\sum_{i=1}^{n}x_i^2 - n\bar{x}^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 .
\end{aligned}$$

The sample variance is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \implies (n-1)\,s^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 ,$$

so the final equality of part (b) holds. □

5.2. Sampling from the normal distribution

THEOREM 5.3. *Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution, and let $\bar{X} = (1/n)\sum_{i=1}^{n} X_i$ and $S^2 = [1/(n-1)]\sum_{i=1}^{n}(X_i - \bar{X})^2$. Then*

1. *$\bar{X}$ and S^2 are independent random variables,*
2. *$\bar{X}$ has a $\mathcal{N}(\mu, \sigma^2/n)$ distribution,*
3. *$(n-1)\,S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom.*

(This is Theorem 5.3.1 from Casella & Berger; the following proof is given there.)

PROOF. [proof goes here] □

DEFINITION 5.4. Let $X_1, \ldots, X_n$ be a random sample from a $\mathcal{N}(\mu, \sigma^2)$ distribution. The quantity $(\bar{X} - \mu)\,/\,(S/\sqrt{n})$ has *Student's t distribution with $n-1$ degrees of freedom.* Equivalently, a random variable T has Student's t distribution with p degrees of freedom, and we write $T \sim t_p$ if it has pdf

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

5.3. Order statistics

The order statistics of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$. The order statistics are random variables that satisfy $X_{(1)} \le \ldots \le X_{(n)}$, and in particular, $X_{(1)} = \min_{1 \le i \le n} X_i$ and $X_{(n)} = \max_{1 \le i \le n} X_i$.

THEOREM 5.5. *Let $X_1, \ldots, X_n$ be a random sample from a discrete distribution with pmf $f_X(x_i) = p_i$, where $x_1 < x_2 < \cdots$ are the possible values of X in ascending order. Define*

$$\begin{aligned}
P_0 &= 0 \\
P_1 &= p_1 \\
P_2 &= p_1 + p_2 \\
&\vdots
\end{aligned}$$

$$P_i = p_1 + p_2 + \cdots + p_i$$
$$\vdots$$

Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics from the sample. Then

$$P\left(\left\{X_{(j)} \leq x_i\right\}\right) = \sum_{k=j}^{n} \binom{n}{k} P_i^k \left(1 - P_i\right)^{n-k}$$

and

$$P\left(\left\{X_{(j)} = x_i\right\}\right) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k \left(1 - P_i\right)^{n-k} - P_{i-1}^k \left(1 - P_{i-1}\right)^{n-k}\right].$$

(This is Theorem 5.4.3 from Casella & Berger.)

PROOF. [proof goes here] □

THEOREM 5.6. *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the pdf of $X_{(j)}$ is*

$$f_{X_{(j)}}(x) = \frac{n!}{(j-1)!\,(n-j)!} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{n-j}.$$

(This is Theorem 5.4.4 from Casella & Berger.)

PROOF. [proof goes here] □

THEOREM 5.7. *Let $X_{(1)}, \ldots, X_{(n)}$ denote the order statistics of a random sample, $X_1, \ldots, X_n$, from a continuous population with cdf $F_X(x)$ and pdf $f_X(x)$. Then the joint pdf of $X_{(i)}$ and $X_{(j)}$, $1 \leq i < j \leq n$, is*

$$f_{X_{(i)},X_{(j)}}(u, v) = \frac{n!}{(i-1)!\,(j-1-i)!\,(n-j)!} f_X(u)\, f_X(v) \left[F_X(u)\right]^{i-1} \left[F_X(v) - F_X(u)\right]^{j-1-i} \left[1 - F_X(v)\right]^{n-j}$$

for $-\infty < u < v < \infty$. (This is Theorem 5.4.6 from Casella & Berger.)

PROOF. [proof goes here] □

5.4. Convergence concepts

THEOREM 5.8 (Strong Law of Large Numbers). *Let $X_1, X_2, \ldots$ be iid random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 < \infty$, and define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Then, for every $\epsilon > 0$,*

$$P\left(\lim_{n\to\infty} \left|\bar{X}_n - \mu\right| < \epsilon\right) = 1;$$

that is, $\bar{X}_n$ converges almost surely to μ.

PROOF. [proof goes here] □

THEOREM 5.9 (Central Limit Theorem). *Let $X_1, X_2, \ldots$ be a sequence of iid random variables whose mgfs exist in a neighborhood of 0 (that is, $M_{X_i}(t)$ exists for $|t| < h$, for some positive h). Let $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2 > 0$. (Both μ and σ^2 are finite since the mgf exists.) Define $\bar{X}_n = (1/n)\sum_{i=1}^{n} X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}\left(\bar{X} - \mu\right)/\sigma$. Then, for any x, $-\infty < x < \infty$,*

$$\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y^2/2}\, \mathrm{d}\, y;$$

that is, $\sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$ has a limiting standard normal distribution. (This is Theorem 5.5.14 from Casella & Berger; the following proof is given there.)

PROOF. Let $Z \sim \mathcal{N}(0,1)$, so that the mgf of Z given by $M_Z(t) = \mathrm{e}^{0\cdot t+\left(1\cdot t^2\right)/2} = \mathrm{e}^{t^2/2}$. We will show that, for $|t| < h$, the mgf of $\sqrt{n}\left(\bar{X}_n - \mu\right)/\sigma$ converges to $\mathrm{e}^{t^2/2}$.
Define $Y_i = (X_i - \mu)/\sigma$, and let $M_Y(t)$ denote the common mgf of the Y_i's, which exists for $|t| < \sigma h$ and is given by theorem 3.9. We have

$$\frac{X_i - \mu}{\sigma} = Y_i \implies X_i - \mu = \sigma Y_i \implies X_i = \sigma Y_i + \mu,$$

so that

$$\begin{aligned}
\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} &= \frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)}{\sigma} \\
&= \frac{\sqrt{n}}{\sigma}\left[\frac{1}{n}\sum_{i=1}^n (\sigma Y_i + \mu) - \mu\right] \\
&= \frac{\sqrt{n}}{\sigma}\left[\frac{1}{n}\left(\sigma\sum_{i=1}^n Y_i + n\mu\right) - \mu\right] \\
&= \frac{\sqrt{n}}{\sigma}\left[\frac{\sigma}{n}\sum_{i=1}^n Y_i + \mu - \mu\right] \\
&= \frac{\sqrt{n}}{\sigma}\left(\frac{\sigma}{n}\sum_{i=1}^n Y_i\right) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i.
\end{aligned}$$

Then, from the properties of mgfs (see theorem 3.9 and theorem 4.9), we have

$$M_{\sqrt{n}\left(\bar{X}_n-\mu\right)/\sigma}(t) = M_{\sum_{i=1}^n Y_i/\sqrt{n}}(t) = M_{\sum_{i=1}^n Y_i}\left(\frac{t}{\sqrt{n}}\right) = \left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n.$$

We now expand $M_Y\left(t/\sqrt{n}\right)$ in a Taylor series (power series) around 0. We have

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=0}^{\infty} M_Y^{(k)}(0)\,\frac{\left(t/\sqrt{n}\right)^k}{k!},$$

where $M_Y^{(k)}(0) = \left(\mathrm{d}^k / \mathrm{d}\, t^k\right) M_Y(t)\,|_{t=0}$. Since the mgfs exist for $|t| < h$, the power series expansion is valid if $t < \sqrt{n}\sigma h$.
We have

$$M_Y^{(0)} = \mathrm{E}\left[Y^0\right] = \mathrm{E}[1] = 1,$$

$$M_Y^{(1)} = \mathrm{E}\left[Y^1\right] = \mathrm{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}\left(\mathrm{E}[X] - \mathrm{E}[\mu]\right) = \frac{1}{\sigma}(\mu - \mu) = 0,$$

and, noting that

$$\mathrm{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 \implies \sigma^2 = \mathrm{E}\left[X^2\right] - \mu^2 \implies \mathrm{E}\left[X^2\right] = \mu^2 + \sigma^2,$$

we have

$$\begin{aligned}
M_Y^{(2)} &= \mathrm{E}\left[Y^2\right] \\
&= \mathrm{E}\left[\left(\frac{X-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{\sigma^2}\mathrm{E}\left[X^2 - 2\mu X + \mu^2\right] \\
&= \frac{1}{\sigma^2}\left(\mathrm{E}\left[X^2\right] - 2\mu\,\mathrm{E}[X] + \mathrm{E}\left[\mu^2\right]\right) \\
&= \frac{1}{\sigma^2}\left(\mu^2 + \sigma^2 - 2\mu^2 + \mu^2\right)
\end{aligned}$$

$$= \frac{1}{\sigma^2}\left(\sigma^2\right)$$
$$= 1.$$

(By construction, the mean and variance of Y are 0 and 1, respectively.) Then, we have

$$\begin{aligned} M_Y\left(\frac{t}{\sqrt{n}}\right) &= \sum_{k=0}^{\infty} M_Y^{(k)}(0)\left(\frac{t/\sqrt{n}}{k!}\right)^k \\ &= 1\frac{(t\sqrt{n})^0}{0!} + 0\frac{(t\sqrt{n})}{1!} + 1\frac{(t\sqrt{n})^2}{2!} + \sum_{k=3}^{\infty} M_Y^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!} \\ &= T_2\left(\frac{t}{\sqrt{n}}\right) + R_2\left(\frac{t}{\sqrt{n}}\right), \end{aligned}$$

where

$$T_2\left(\frac{t}{\sqrt{n}}\right) = 1 + \frac{(t/\sqrt{n})^2}{2!} \quad \text{and} \quad R_2\left(\frac{t}{\sqrt{n}}\right) = \sum_{k=3}^{\infty} M_Y^{(k)}(0)\frac{(t/\sqrt{n})^k}{k!}.$$

We have $n > 0$, so for fixed $t \neq 0$, the quantity $t/\sqrt{n} \to 0$ as $n \to \infty$. Then, noting that $M_Y^{(2)}(0)$ exists, it follows from theorem 5.11 that

$$\lim_{t/\sqrt{n}\to 0} \frac{M_Y(t/\sqrt{n}) - T_2(t/\sqrt{n})}{(t/\sqrt{n} - 0)^2} = 0 \implies \lim_{n\to\infty} \frac{R_2(t/\sqrt{n})}{(t/\sqrt{n})^2} = 0.$$

Since t is fixed, we also have

$$\lim_{n\to\infty} \frac{R_2(t/\sqrt{n})}{(1/\sqrt{n})^2} = \lim_{n\to\infty} nR_2\left(\frac{t}{\sqrt{n}}\right) = 0,$$

and this is also true at $t = 0$ since

$$R_2\left(\frac{0}{\sqrt{n}}\right) = R_2(0) = \sum_{k=3}^{\infty} M_Y^{(k)}(0)\frac{0^k}{k!} = \sum_{k=3}^{\infty} M_Y^{(k)}(0)\cdot 0 = 0.$$

Thus, for any fixed t, we can write

$$\begin{aligned} \lim_{n\to\infty}\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n &= \lim_{n\to\infty}\left[1 + \frac{(t/\sqrt{n})^2}{2} + R_2\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &= \lim_{n\to\infty}\left[1 + \frac{t^2}{2n} + R_2\left(\frac{t}{\sqrt{n}}\right)\right]^n \\ &= \lim_{n\to\infty}\left[1 + \frac{1}{n}\left(\frac{t^2}{2} + nR_2\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n. \end{aligned}$$

Setting $a_n = \left(t^2/2\right) + nR_2(t/\sqrt{n})$, we have

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty}\left[\frac{t^2}{2} + nR_2\left(\frac{t}{\sqrt{n}}\right)\right] = \lim_{n\to\infty}\frac{t^2}{2} + \lim_{n\to\infty} nR_2\left(\frac{t}{\sqrt{n}}\right) = \frac{t^2}{2} + 0 = \frac{t^2}{2},$$

i.e., the sequence $a_1, a_2, \ldots$ converges to $a = t^2/2$ as $n \to \infty$. It follows from lemma 3.8 that

$$\lim_{n\to\infty}\left(M_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n = \lim_{n\to\infty}\left[1 + \frac{1}{n}\left(\frac{t^2}{2} + nR_2\left(\frac{t}{\sqrt{n}}\right)\right)\right]^n = \lim_{n\to\infty}\left[1 + \frac{a_n}{n}\right]^n = \mathrm{e}^a = \mathrm{e}^{t^2/2}.$$

Since $\mathrm{e}^{t^2/2}$ is the mgf of the $\mathcal{N}(0,1)$ distribution, the theorem is proved. □

DEFINITION 5.10. If a function $g(x)$ has derivatives of order r, that is, $g^{(r)}(x) = \frac{\mathrm{d}^r}{\mathrm{d}\,x^r} g(x)$ exists, then for any constant a, the *Taylor polynomial of order r about a* is

$$T_r(x) = \sum_{i=0}^{r} \frac{g^{(i)}(a)}{i!}(x-a)^i.$$

THEOREM 5.11 (Taylor). *If*

$$g^{(r)}(a) = \frac{\mathrm{d}^r}{\mathrm{d}\,x^r} g(x)\Big|_{x=a}$$

exists, then

$$\lim_{x \to a} \frac{g(x) - T_r(x)}{(x-a)^r} = 0.$$

(This is Theorem 5.5.21 from Casella & Berger).

CHAPTER 6

Linear algebra

THEOREM 6.1 (Fredholm Alternative). *Let* $\mathbf{A} \in \mathbf{M}_{m,n}(\mathbb{R})$*, let* $\mathbf{x} \in \mathbb{R}^n$*, and let* $\mathbf{b} \in \mathbb{R}^m$*. Then, there are two mutually exclusive possibilities:*

(1) *The system* $\mathbf{Ax} = \mathbf{b}$ *has a unique solution* $\mathbf{x}$ *for each* $\mathbf{b}$*. In particular, the system has the solution* $\mathbf{x} = \mathbf{0}$ *for* $\mathbf{b} = \mathbf{0}$*.*
(2) *The homogeneous equation* $\mathbf{Ax} = \mathbf{0}$ *has exactly* p *linearly independent solutions* $\{\mathbf{x}_i\}_{i=1}^p$ *for some* $p \geq 1$*.*

DEFINITION 6.2. Suppose that $\mathbf{A} \in \mathbf{M}_{m,p}(\mathbb{R})$, and suppose that $\mathbf{B} \in \mathbf{M}_{p,n}(\mathbb{R})$. Then, the *Wronskian* of $\mathbf{A}$ and $\mathbf{B}$ is $\langle \mathbf{A}, \mathbf{B} \rangle \coloneqq \mathbf{AB} - \mathbf{BA}$.

DEFINITION 6.3. The *range* of a matrix $\mathbf{A}$, denoted $\text{range}(\mathbf{A})$, is the space spanned by the columns of $\mathbf{A}$.

DEFINITION 6.4. The *null space* of a matrix $\mathbf{A} \in \mathbf{M}_{m,n}(\mathbb{R})$, denoted $\text{null}(\mathbf{A})$, is the set of vectors $\mathbf{x} \in \mathbb{R}^n$ that satisfy $\mathbf{Ax} = \mathbf{0}$.

THEOREM 6.5 (Invertible Matrix Theorem). *Let* $\mathbf{A} \in \mathbf{M}_{m,m}(\mathbb{R})$*. Then, the following are equivalent:*

(1) $\mathbf{A}^{-1}$ *exists.*
(2) $\mathit{rank}(\mathbf{A}) = m$*.*
(3) $\mathit{range}(\mathbf{A}) = \mathbb{R}^m$*.*
(4) $\mathit{null}(\mathbf{A}) = \mathbf{0}$*.*

THEOREM 6.6 (Spectral Theorem). *A non-degenerate matrix* $\mathbf{A} \in \mathbf{M}_{m,m}(\mathbb{R})$ *has a decomposition of the form* $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^{-1}$*, provided that* $\mathbf{X}^{-1} \in \mathbf{M}_{m,m}(\mathbb{R})$ *exists, and where* $\mathbf{\Lambda}$ *is a diagonal matrix whose entries are the eigenvalues of* $\mathbf{A}$*.*

PROOF. [proof goes here] □

DEFINITION 6.7. A *unitary matrix* $\mathbf{U} \in \mathbf{M}_{m,m}(\mathbb{R})$ has the property $\mathbf{U}^{-1} = \mathbf{U}^{\mathsf{H}}$, where $\mathbf{A}^{\mathsf{H}}$ denotes the Hermitian conjugate (conjugate transpose) of $\mathbf{A}$.

THEOREM 6.8 (Unitary Decomposition). *A symmetric matrix* $\mathbf{A} \in \mathbf{M}_{m,m}(\mathbb{R})$ *admits the unitary diagonalization* $\mathbf{A} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{-1}$*, where* $\mathbf{Q} \in \mathbf{M}_{m,m}(\mathbb{R})$ *is unitary.*

Part 2

Mathematical statistics

CHAPTER 7

Common families of distributions

7.1. Exponential families

A family of pdfs (or pmfs) indexed by a parameter $\boldsymbol{\theta}$ is called a k-parameter exponential family if it can be expressed as

$$f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \exp\left\{\sum_{j=1}^{k} \omega_j(\boldsymbol{\theta})\, t_j(x)\right\}$$

where $h(x) \geq 0$, $c(\boldsymbol{\theta}) \geq 0$, $t_1(x), \ldots, t_k(x)$ are real-valued functions of x, and $\omega_1(\boldsymbol{\theta}), \ldots, \omega_k(\boldsymbol{\theta})$ are real-valued functions of the possibly vector-valued parameter $\boldsymbol{\theta}$. I.e., $f(x|\boldsymbol{\theta})$ can be expressed in three parts: a part that depends only on the random variable(s), a part that depends only on the parameter(s), and a part that depends on both the random variable(s) and the parameter(s). Most of the parametric models you have studied in Math-501 are exponential families, e.g., normal, gamma, beta, binomial, negative binomial, Poisson, and multinomial. The uniform distribution is not an exponential family (see example 7.6 below).

EXAMPLE 7.1 (Logistic regression). For $Y_1, Y_2, \ldots, Y_n$, let $Y_i \sim \text{Bernoulli}(p)$, i.e.,

$$Y_i = \begin{cases} 0, & \text{if no event} \\ 1, & \text{if event.} \end{cases}$$

Then the logistic regression model is

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$$

where $\log(p/(1-p))$ is called the logit link.

EXAMPLE 7.2 (Binomial random variables). Let $X \sim \mathcal{B}(n, p)$, where $p \in (0, 1)$. Recall that X represents the number of successes in n i.i.d. Bernoulli trials and its pmf is given by

$$f(x|p) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x = 0, 1, \ldots, n$ and $f(x|p) = 0$ otherwise. Express $f(x|p)$ in exponential family form.

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} \\ &= \binom{n}{x} p^x (1-p)^n (1-p)^{-x} \\ &= \binom{n}{x} (1-p)^n \left(\frac{p^x}{(1-p)^x}\right) \\ &= \binom{n}{x} (1-p)^n \left(\frac{p}{1-p}\right)^x \\ &= \binom{n}{x} (1-p)^n \exp\left\{\log\left(\frac{p}{1-p}\right)^x\right\} \end{aligned}$$

$$= \underbrace{\binom{n}{x}}_{h(x)} \underbrace{(1-p)^n}_{c(p)} \exp\left\{ \underbrace{x}_{t_1(x)} \underbrace{\log\left(\frac{p}{1-p}\right)}_{\omega_1(p)} \right\}$$

EXAMPLE 7.3 (Poisson random variables). Let $X \sim \text{Poisson}(\lambda)$, where $\lambda > 0$. Recall that X represents the frequency with which a specified event occurs given some fixed dimension, such as space or time, and its pmf is given by

$$f(x|\lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}$$

for $x = 0, 1, 2, \ldots$ and $f(x|\lambda) = 0$ otherwise. Express $f(x|\lambda)$ in exponential family form.

$$f(x|\lambda) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!} = \frac{1}{x!}\mathrm{e}^{-\lambda}\exp\{\log(\lambda^x)\} = \frac{1}{x!}\mathrm{e}^{-\lambda}\exp\{x\log\lambda\}$$

Then, we have $h(x) = 1/x!$, $c(\lambda) = \mathrm{e}^{-\lambda}$, $t_1(x) = x$, and $\omega_1(\lambda) = \log\lambda$. In a Poisson regression, we have $\log(\lambda) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$.

EXAMPLE 7.4 (Normal random variables). Let $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma > 0$. A pdf for X is given by

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

for $x \in \mathbb{R}$. Express $f(x|\mu,\sigma^2)$ in exponential family form.

Suppose σ is known.

$$\begin{aligned} f(x|\mu) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{x^2-2\mu x+\mu^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}\exp\left\{-\frac{-2\mu x}{2\sigma^2}\right\} \\ &= \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{x^2}{2\sigma^2}\right\}}_{h(x)}\underbrace{\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}}_{c(\mu)}\exp\left\{\underbrace{\frac{\mu}{\sigma^2}}_{\omega_1(\mu)}\cdot\underbrace{x}_{t_1(x)}\right\} \end{aligned}$$

Suppose σ is unknown.

$$\begin{aligned} f(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}}\left(\sigma^2\right)^{-1/2}\exp\left\{-\frac{x^2-2\mu x+\mu^2}{2\sigma^2}\right\} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left\{\log\left(\sigma^2\right)^{-1/2}\right\}\exp\left\{-\frac{x^2-2\mu x}{2\sigma^2}\right\}\exp\left\{-\frac{\mu^2}{2\sigma^2}\right\} \\ &= \underbrace{\frac{1}{\sqrt{2\pi}}}_{h(x)}\underbrace{\exp\left\{-\frac{\mu^2}{2\sigma^2}-\frac{1}{2}\log\sigma^2\right\}}_{c(\mu,\sigma^2)}\exp\left\{\underbrace{\frac{1}{\sigma^2}}_{\omega_1(\mu,\sigma^2)}\cdot\underbrace{\left(-\frac{x^2}{2}\right)}_{t_1(x)}+\underbrace{\frac{\mu}{\sigma^2}}_{\omega_2(\mu,\sigma^2)}\cdot\underbrace{x}_{t_2(x)}\right\} \end{aligned}$$

Thus, in the case that σ is unknown, $f(x|\mu,\sigma^2)$ is a two-parameter exponential family, i.e., we have $k = 2$ for $\sum_{j=1}^{k}\omega_j(\theta)\,t_j(x)$.

DEFINITION 7.5. The *indicator function* of a set A, most often denoted by $I_A(x)$, is the function

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}.$$

EXAMPLE 7.6 (Uniform random variables). Let $X \sim \mathcal{U}(0,\theta)$, where $\theta > 0$. A pdf for X is given by

$$f(x|\theta) = \frac{1}{\theta - 0} = \frac{1}{\theta}$$

for $0 < x < \theta$. Express $f(x|\theta)$ in exponential family form, if possible.
Let $A = \{x : x \in (0,\theta)\}$ and let I_A be the indicator function of A, i.e.,

$$I_A(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{if } x \notin A \end{cases}.$$

Then, we can write $f(x|\theta)$ as

$$f(x|\theta) = \frac{1}{\theta} I_A(x) = \frac{1}{\theta} I_{(0,\theta)}(x).$$

Notice that $I_{(0,\theta)}(x)$ is not a function of x exclusively, not a function of θ exclusively, and cannot be written as an exponential. Because the entire pdf must be incorporated into $h(x)$, $c(\theta)$, $t_j(x)$, and $\omega_j(\theta)$, it follows that the familiy of pdfs given by $f(x|\theta)$ is not an exponential family.

EXAMPLE 7.7 (Three-parameter exponential family distribution). Consider the family of distributions with densities

$$f(x|\theta) = \frac{2}{\Gamma(1/4)} \exp\left[-(x-\theta)^4\right]$$

for $x \in \mathbb{R}$. Express $f(x|\theta)$ in exponential family form.
Recall that the binomial theorem states that

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k},$$

so we have

$$\begin{aligned}
f(x|\theta) &= \frac{2}{\Gamma(1/4)} \exp\left[-(x-\theta)^4\right] \\
&= \frac{2}{\Gamma(1/4)} \exp\left\{-\sum_{k=0}^{4} \binom{4}{k} x^k (-\theta)^{4-k}\right\} \\
&= \frac{2}{\Gamma(1/4)} \exp\left\{-\left[\binom{4}{0} x^0 (-\theta)^4 + \binom{4}{1} x (-\theta)^3 + \binom{4}{2} x^2 (-\theta)^2 + \binom{4}{3} x^3 (-\theta) + \binom{4}{4} x^4 (-\theta)^0\right]\right\} \\
&= \frac{2}{\Gamma(1/4)} \exp\left\{-\left[1 \cdot 1 \cdot \theta^4 - 4x\theta^3 + 6x^2\theta^2 - 4x^3\theta + 1 \cdot x^4 \cdot 1\right]\right\} \\
&= \underbrace{\frac{2}{\Gamma(1/4)} \exp\left\{-x^4\right\}}_{h(x)} \underbrace{\exp\left\{-\theta^4\right\}}_{c(\theta)} \exp\left\{\underbrace{4x^3}_{t_1(x)}\underbrace{\theta}_{\omega_1(\theta)} \underbrace{-6x^2}_{t_2(x)}\underbrace{\theta^2}_{\omega_2(\theta)} + \underbrace{4x}_{t_3(x)}\underbrace{\theta^3}_{\omega_3(\theta)}\right\}.
\end{aligned}$$

THEOREM 7.8. *Random samples from k-parameter exponential families have joint distributions which are k-parameter exponential families.*

PROOF. Suppose that a random variable X has a pdf $f(x|\theta)$, and that $X_1, X_2, \ldots, X_n$ is a random sample from a population having the distribution of X. It follows that the X_i's are independent and identically distributed, and that each X_i has the same cdf as X, and therefore that $f(x|\theta)$ is a pdf for each X_i. Then, the joint pdf of the X_i's is given by

$$\begin{aligned}
\text{(independence)} \qquad f(x_1, x_2, \ldots, x_n|\theta) &= \prod_{i=1}^{n} f(x_i|\theta) \\
\text{(f is part of an exponential family)} \qquad &= \prod_{i=1}^{n} \left[h(x_i)\, c(\theta) \exp\left\{\sum_{j=1}^{k} t_j(x_i)\, \omega_j(\theta)\right\}\right]
\end{aligned}$$

$$(\mathrm{e}^x \cdot \mathrm{e}^y = \mathrm{e}^{x+y}) \qquad = \left[\prod_{i=1}^{n} h(x_i)\right] [c(\theta)]^n \exp\left\{\sum_{j=1}^{k}\sum_{i=1}^{n} t_j(x_i)\,\omega_j(\theta)\right\}$$

Then, let

$$h^*(x) = \prod_{i=1}^{n} h(x_i) \quad \text{and} \quad c^*(\theta) = [c(\theta)]^n,$$

so that we have

$$f(x_1, x_2, \ldots, x_n|\theta) = \left[\prod_{i=1}^{n} h(x_i)\right] [c(\theta)]^n \exp\left\{\sum_{j=1}^{k}\sum_{i=1}^{n} t_j(x_i)\,\omega_j(\theta)\right\}$$
$$= h^*(x)\, c^*(\theta) \exp\left\{\sum_{j=1}^{k}\left(\omega_j(\theta)\sum_{i=1}^{n} t_j(x_i)\right)\right\}.$$

Now, let

$$T_j(x) = \sum_{i=1}^{n} t_j(x_i),$$

so that

$$f(x_1, x_2, \ldots, x_n|\theta) = h^*(x)\, c^*(\theta) \exp\left\{\sum_{j=1}^{k} \omega_j(\theta)\, T_j(x)\right\}.$$

Thus, the joint pdf $f(x_1, x_2, \ldots, x_n|\theta)$ is a k-parameter exponential family. □

7.1.1. Natural parameters. An exponential family is sometimes reparametrized as

$$f(x|\boldsymbol{\eta}) = h(x)\, c^*(\boldsymbol{\eta}) \exp\left(\sum_{j=1}^{k} \eta_j t_j(x)\right),$$

where the natural parameters are defined by $\eta_j = \omega_j(\theta)$ and the natural parameter space is

$$\left\{\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k) : \int h(x) \exp\left\{\sum_{j=1}^{k} \eta_j t_j(x)\right\} \mathrm{d}\,x < \infty\right\}$$

so that

$$c^*(\boldsymbol{\eta}) = \frac{1}{\int h(x) \exp\left\{\sum_{j=1}^{k} \eta_j t_j(x)\,\mathrm{d}\,x\right\}},$$

which ensures that the pdf integrates to 1.

EXAMPLE 7.9 (Binomial random variables). Let $X \sim \text{Binomial}(n, p)$. From example 7.2, the pmf of X can be written as

$$f(x|p) = \binom{n}{x} (1-p)^n \exp\left\{x \log\left(\frac{p}{1-p}\right)\right\},$$

where $k = 1$ and

$$\omega_1(p) = \log\frac{p}{1-p}.$$

Then, let $\eta = \omega_1(p)$, so that

$$\eta = \log\frac{p}{1-p} \implies \mathrm{e}^\eta = \frac{p}{1-p} \implies p = \mathrm{e}^\eta(1-p) = \mathrm{e}^\eta - \mathrm{e}^\eta p \implies \mathrm{e}^\eta = p(1+\mathrm{e}^\eta) \implies p = \frac{\mathrm{e}^\eta}{1+\mathrm{e}^\eta}.$$

Then, we have

$$c(p) = (1-p)^n \implies c(\eta) = \left(1 - \frac{\mathrm{e}^\eta}{1+\mathrm{e}^\eta}\right)^n = \left(\frac{1}{1+\mathrm{e}^\eta}\right)^n$$

and

$$f(x|\eta) = \binom{n}{x}\left(\frac{1}{1+\mathrm{e}^{\eta}}\right)^{n}\exp(x\eta).$$

EXAMPLE 7.10 (Poisson random variables). Let $X \sim \text{Poisson}(\lambda)$. From example 7.3, the pmf of X can be written as

$$f(x|\lambda) = \frac{1}{x!}\mathrm{e}^{-\lambda}\exp\{x\log\lambda\},$$

where $k = 1$ and

$$\omega_1(\lambda) = \log\lambda.$$

Then, let $\eta = \omega_1(\lambda)$, so that

$$\eta = \log\lambda \implies \mathrm{e}^{\eta} = \exp(\log\lambda) \implies \mathrm{e}^{\eta} = \lambda.$$

Then, we have

$$c(\lambda) = \mathrm{e}^{-\lambda} \implies c(\eta) = \exp(-\mathrm{e}^{\eta})$$

and

$$f(x|\eta) = \frac{1}{x!}\exp(-\mathrm{e}^{\eta})\exp(x\eta).$$

EXAMPLE 7.11 (Bernoulli random variables). Let $X \sim \text{Bernoulli}(p)$, i.e., $X \sim \text{Binomial}(1,p)$. From example 7.9, we have

$$f(x|\eta) = \binom{n}{x}\left(\frac{1}{1+\mathrm{e}^{\eta}}\right)^{n}\exp(x\eta).$$

With $n = 1$, we have

$$f(x|\eta) = \binom{1}{x}\left(\frac{1}{1+\mathrm{e}^{\eta}}\right)\exp(x\eta) = 1\cdot\left(\frac{1}{1+\mathrm{e}^{\eta}}\right)\exp(x\eta) = \frac{1}{1+\mathrm{e}^{\eta}}\exp(x\eta).$$

EXAMPLE 7.12 (Normal random variables). Let $X \sim \mathcal{N}(\mu,\sigma^2)$, where $\sigma > 0$ and σ is unknown. From example 7.4, we have

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2\right\}\exp\left\{-x^2\frac{1}{2\sigma^2} + x\frac{\mu}{\sigma^2}\right\}.$$

Then, let $\eta_1 = \omega_1(\mu,\sigma^2)$, so that

$$\eta_1 = \frac{1}{\sigma^2} \implies \sigma^2\eta_1 = 1 \implies \sigma^2 = \frac{1}{\eta_1}$$

and let $\eta_2 = \omega_2(\mu,\sigma^2)$, so that

$$\eta_2 = \frac{\mu}{\sigma^2} \implies \mu = \sigma^2\eta_2 = \frac{\eta_2}{\eta_1}.$$

Then, we have

$$\begin{aligned}
c(\mu,\sigma^2) &= \exp\left\{-\frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log\sigma^2\right\} \\
c^*(\eta_1,\eta_2) &= \exp\left\{-\frac{\left(\frac{\eta_2}{\eta_1}\right)^2}{2\left(\frac{1}{\eta_1}\right)} - \frac{1}{2}\log\frac{1}{\eta_1}\right\} \\
&= \exp\left\{-\frac{\frac{\eta_2^2}{\eta_1^2}}{\frac{2}{\eta_1}} - \frac{1}{2}\log\frac{1}{\eta_1}\right\} \\
&= \exp\left\{-\frac{\eta_2^2}{2\eta_1} + \log\left(\frac{1}{\eta_1}\right)^{-1/2}\right\} \\
&= \exp\left\{-\frac{\eta_2^2}{2\eta_1} + \log\left((\eta_1)^{-1}\right)^{-1/2}\right\}
\end{aligned}$$

$$= \exp\left\{-\frac{\eta_2^2}{2\eta_1} + \log\sqrt{\eta_1}\right\}$$

and

$$f(x|\eta_1,\eta_2) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{\eta_2^2}{2\eta_1} + \log\sqrt{\eta_1}\right\}\exp\left\{-\frac{\eta_1 x^2}{2} + \eta_2 x\right\}.$$

THEOREM 7.13. *Let X have density in an exponential family. Then,*

(1) $\mathrm{E}[t_j(X)] = -\dfrac{\partial}{\partial\eta_j}\log c^*(\eta)$,

(2) $\mathrm{Var}(t_j(X)) = -\dfrac{\partial^2}{\partial\eta_j^2}\log c^*(\eta)$,

and the moment-generating function for $(X_1,\ldots,X_k)$ is

$$M_{(X_1,\ldots,X_k)}(s_1,\ldots,s_k) = \mathrm{E}\left[\mathrm{e}^{\sum_{j=1}^k s_j X_j}\right].$$

PROOF. We begin with the pdf of an exponential family, i.e.,

$$\begin{aligned}
&\text{(definition of a pdf)} && 1 = \int f(x|\theta)\,\mathrm{d}x\\
&\text{(f is in an exponential family)} && = \int h(x)\,c(\theta)\exp\left(\sum_{i=1}^k \omega_i(\theta)\,t_i(x)\right)\mathrm{d}\,x\\
&\text{(natural parameterization)} && = \int h(x)\,c^*(\eta)\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\mathrm{d}\,x.
\end{aligned}$$

Taking the derivative of both sides with repect to η_j gives

$$\begin{aligned}
\frac{\partial}{\partial\eta_j}1 &= \frac{\partial}{\partial\eta_j}\int h(x)\,c^*(\eta)\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\mathrm{d}\,x\\
\implies 0 &= \int \frac{\partial}{\partial\eta_j}\left[h(x)\,c^*(\eta)\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\right]\mathrm{d}\,x\\
&= \int\left[h(x)\left[\frac{\partial}{\partial\eta_j}(c^*(\eta))\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right) + c^*(\eta)\frac{\partial}{\partial\eta_j}\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\right]\right]\mathrm{d}\,x\\
&= \int h(x)\frac{\partial}{\partial\eta_j}(c^*(\eta))\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\mathrm{d}\,x\\
&\quad + \int h(x)\,c^*(\eta)\frac{\partial}{\partial\eta_j}\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\mathrm{d}\,x\\
&= \int h(x)\frac{\partial}{\partial\eta_j}(c^*(\eta))\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\mathrm{d}\,x\\
&\quad + \int h(x)\,c^*(\eta)\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\left(\sum_{i=1}^k \frac{\partial}{\partial\eta_j}\eta_i t_i(x)\right)\mathrm{d}\,x\\
&= \int h(x)\frac{\partial}{\partial\eta_j}(c^*(\eta))\exp\left(\sum_{i=1}^k \eta_i t_i(x)\right)\mathrm{d}\,x + \mathrm{E}\left[\sum_{i=1}^k \frac{\partial}{\partial\eta_j}\eta_i t_i(X)\right],
\end{aligned}$$

where the final equality follows from the definition of expected value. Observe that for some differentiable function $g(x)$, we have

$$g'(x) = \frac{g(x)}{g(x)}g'(x) = g(x)\frac{\mathrm{d}}{\mathrm{d}\,x}\log(g(x)),$$

which leads to

$$\begin{aligned}
0 &= \int h(x)\, c^*(\eta) \frac{\partial}{\partial \eta_j} \log\left(c^*(\eta)\right) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x + \mathrm{E}\left[\sum_{i=1}^{k} \frac{\partial}{\partial \eta_j} \eta_i t_i(X)\right] \\
&= \frac{\partial}{\partial \eta_j}\left(\log c^*(\eta)\right) \int h(x)\, c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x + \mathrm{E}\left[\sum_{i=1}^{k} \frac{\partial}{\partial \eta_j} \eta_i t_i(X)\right] \\
&= \frac{\partial}{\partial \eta_j}\left(\log c^*(\eta)\right) \cdot 1 + \mathrm{E}\left[\sum_{i=1}^{k} \frac{\partial}{\partial \eta_j} \eta_i t_i(X)\right],
\end{aligned}$$

where the final equality follows from the fact that the integral of a pdf over its range of positivity is equal to 1. Then,

$$\begin{aligned}
-\frac{\partial}{\partial \eta_j} \log c^*(\eta) &= \mathrm{E}\left[\frac{\partial}{\partial \eta_j} \eta_1 t_1(X) + \ldots + \frac{\partial}{\partial \eta_j} \eta_j t_j(X) + \ldots + \frac{\partial}{\partial \eta_j} \eta_k t_k(X)\right] \\
&= \mathrm{E}\left[0 \cdot t_1(X) + \ldots + 1 \cdot t_j(X) + \ldots + 0 \cdot t_k(X)\right] \\
&= \mathrm{E}\left[t_j(X)\right],
\end{aligned}$$

proving the first claim. Then,

$$\begin{aligned}
-\frac{\partial^2}{\partial \eta_j^2} \log c^*(\eta) &= \frac{\partial}{\partial \eta_j}\left(-\frac{\partial}{\partial \eta_j} \log c^*(\eta)\right) \\
&= \frac{\partial}{\partial \eta_j} \mathrm{E}\left[t_j(X)\right] \\
&= \frac{\partial}{\partial \eta_j} \int t_j(x)\, h(x)\, c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x \\
&= \int t_j(x)\, h(x) \frac{\partial}{\partial \eta_j} c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x \\
&= \int t_j(x)\, h(x) \left[\frac{\partial}{\partial \eta_j}\left(c^*(\eta)\right) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) + c^*(\eta) \frac{\partial}{\partial \eta_j} \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right)\right] \mathrm{d}\, x \\
&= \int t_j(x)\, h(x) \frac{\partial}{\partial \eta_j}\left(c^*(\eta)\right) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x \\
&\quad + \int t_j(x)\, h(x)\, c^*(\eta) \frac{\partial}{\partial \eta_j} \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x.
\end{aligned}$$

The first summand becomes

$$\begin{aligned}
&\int t_j(x)\, h(x)\, c^*(\eta) \frac{\partial}{\partial \eta_j} \log\left(c^*(\eta)\right) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x \\
&= \frac{\partial}{\partial \eta_j} \log\left(c^*(\eta)\right) \int t_j(x)\, h(x)\, c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x \\
&= \frac{\partial}{\partial \eta_j} \log\left(c^*(\eta)\right) \mathrm{E}\left[t_j(X)\right] \\
&= \left(-\mathrm{E}\left[t_j(X)\right]\right) \mathrm{E}\left[t_j(X)\right] \\
&= -\left(\mathrm{E}\left[t_j(X)\right]\right)^2,
\end{aligned}$$

where the penultimate equality follows from the first part of the proof. The second summand becomes

$$\int t_j(x)\, h(x)\, c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \left(\sum_{i=1}^{k} \frac{\partial}{\partial \eta_j} \eta_i t_i(x)\right) \mathrm{d}\, x$$

$$
\begin{aligned}
&= \int t_j(x)\, h(x)\, c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \left(\frac{\partial}{\partial \eta_1} n_1 t_1(x) + \cdots + \frac{\partial}{\partial \eta_k} n_k t_k(x)\right) \mathrm{d}\, x \\
&= \int t_j(x)\, h(x)\, c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) (0 + \cdots + 1 \cdot t_j(x) + \cdots + 0)\, \mathrm{d}\, x \\
&= \int (t_j(x))^2\, h(x)\, c^*(\eta) \exp\left(\sum_{i=1}^{k} \eta_i t_i(x)\right) \mathrm{d}\, x \\
&= \mathrm{E}\left[(t_j(X))^2\right].
\end{aligned}
$$

For some random variable Y with defined second central moment, we have

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] \\
&= \mathrm{E}\left[Y^2 - 2Y\,\mathrm{E}[Y] + (\mathrm{E}[Y])^2\right] \\
&= \mathrm{E}\left[Y^2\right] - 2\,\mathrm{E}[Y\,\mathrm{E}[Y]] + \mathrm{E}\left[(\mathrm{E}[Y])^2\right] \\
(\mathrm{E}[Y] \text{ is constant}) \qquad &= \mathrm{E}\left[Y^2\right] - 2\,\mathrm{E}[Y]\,\mathrm{E}[Y] + (\mathrm{E}[Y])^2 \\
&= \mathrm{E}\left[Y^2\right] - 2(\mathrm{E}[Y])^2 + (\mathrm{E}[Y])^2 \\
&= \mathrm{E}\left[Y^2\right] - (\mathrm{E}[Y])^2.
\end{aligned}
$$

It follows that

$$
-\frac{\partial^2}{\partial \eta_j^2} \log c^*(\eta) = -(\mathrm{E}[t_j(X)])^2 + \mathrm{E}\left[(t_j(X))^2\right] = \mathrm{Var}(t_j(X)),
$$

proving the second claim. □

EXAMPLE 7.14 (Expected value of a binomial random variable). Let $X \sim \mathrm{Binomial}(n, p)$. We will find the expected value of X by applying theorem 7.13. From example 7.9, the pmf of X is given by

$$
f(x|\eta) = \binom{n}{x} \left(\frac{1}{1+\mathrm{e}^{\eta}}\right)^n \exp(x\eta),
$$

where $\eta = \log(p/(1-p))$, so that $p = 1/(1+\mathrm{e}^{\eta})$. From the general form of a natural parameterization, we have $k = 1$, $t(x) = x$, and $c^*(\eta) = (1/(1+\mathrm{e}^{\eta}))^n$. Then, we have

$$
\begin{aligned}
\mathrm{E}[X] &= \mathrm{E}[t(X)] \\
&= -\frac{\partial}{\partial \eta} \log \left(\frac{1}{1+\mathrm{e}^{\eta}}\right)^n \\
&= -\frac{\partial}{\partial \eta} \log (1+\mathrm{e}^{\eta})^{-n} \\
&= -\frac{\partial}{\partial \eta} \left(-n \log (1+\mathrm{e}^{\eta})\right) \\
&= n \frac{\partial}{\partial \eta} \log (1+\mathrm{e}^{\eta}) \\
&= n \frac{\mathrm{e}^{\eta}}{1+\mathrm{e}^{\eta}} \\
&= np.
\end{aligned}
$$

THEOREM 7.15. *If X has a k-parameter exponential family distribution indexed by the natural parameters, then for any η on the interior of the natural parameter space, the mgf of $(t_1(X), \ldots, t_k(X))$ exists and is given by*

$$
M_{(t_1(X),\ldots,t_k(X))}(s_1, \ldots, s_k) = \frac{c^*(\eta)}{c^*(\eta + s)}
$$

where $\eta + s$ is the vector $(\eta_1 + s_1, \ldots, \eta_k + s_k)$.

PROOF. Suppose that X is a k-parameter exponential family distribution indexed by the natural parameters. Then, from section §7.1.1, it has a pdf given by

$$f(x|\eta) = h(x)\, c^*(\eta) \exp\left\{\sum_{j=1}^{k} \eta_j t_j(x)\right\}.$$

It follows from theorem 7.13 that

$$M_{(t_1(X),\ldots,t_k(X))}(s_1,\ldots,s_k) = \mathrm{E}\left[\mathrm{e}^{\sum_{j=1}^{k} s_j t_j(X)}\right],$$

with X_i replaced by $t_i(X)$. Then, we have

$$\begin{aligned}
\mathrm{E}\left[\mathrm{e}^{\sum_{j=1}^{k} s_j t_j(X)}\right] &= \int \exp\left\{\sum_{j=1}^{k} s_j t_j(x)\right\} h(x)\, c^*(\eta) \exp\left\{\sum_{j=1}^{k} \eta_j t_j(x)\right\} \mathrm{d}\, x \\
&= \int h(x)\, c^*(\eta) \exp\left\{\sum_{j=1}^{k} s_j t_j(x) + \sum_{j=1}^{k} \eta_j t_j(x)\right\} \mathrm{d}\, x \\
&= \int h(x)\, c^*(\eta) \exp\left\{\sum_{j=1}^{k} (s_j + \eta_j)\, t_j(x)\right\} \mathrm{d}\, x \\
&= \frac{c^*(\eta+s)}{c^*(\eta+s)} \int h(x)\, c^*(\eta) \exp\left\{\sum_{j=1}^{k} (s_j + \eta_j)\, t_j(x)\right\} \mathrm{d}\, x \\
&= \frac{c^*(\eta)}{c^*(\eta+s)} \int h(x)\, c^*(\eta+s) \exp\left\{\sum_{j=1}^{k} (s_j + \eta_j)\, t_j(x)\right\} \mathrm{d}\, x \\
&= \frac{c^*(\eta)}{c^*(\eta+s)} \cdot \int f(x|\eta+s)\, \mathrm{d}\, x \\
&= \frac{c^*(\eta)}{c^*(\eta+s)} \cdot 1 \\
&= \frac{c^*(\eta)}{c^*(\eta+s)},
\end{aligned}$$

establishing the claim. □

DEFINITION 7.16. A *curved exponential family* is a family of densities of the form given in section §7.1 for which the dimension of the vector $\boldsymbol{\theta}$ is equal to $d < k$, where k is the number of terms in the sum in the exponent. If $d = k$, the family is a *full exponential family.*

7.2. Location and scale families

Location families, scale families, and location-scale families are constructed by specifying a single pdf, $f(x)$, called the standard pdf for the family. Then, all other pdfs in the family are generated by transforming the standard pdf in a prescribed way.

7.2.1. Location families.

DEFINITION 7.17. Let $f(x)$ be any pdf. Then, the family of pdfs indexed by μ, $f(x-\mu)$, is called the *location family* with respect to the standard pdf f, and μ is called the l*ocation parameter.*

E.g., $f(x) \sim \mathcal{N}(0,1^2)$, $\mathcal{N}(\mu,1^2)$ is a location family. The location parameter μ simply shifts the pdf $f(x)$ so that the shape of the graph is unchanged but the point on the graph that was above $x = 0$ under $f(x)$ is above $x = \mu$ for $f(x-\mu)$, thus

$$P(\{-1 \le X \le 2 | X \sim f(x)\}) = P(\{\mu - 1 \le X \le \mu + 2 | X \sim f(x-\mu)\}).$$

Figure 7.2.1 shows the normal distribution with $\sigma^2 = 1^2$ and $\mu = 0$, $\mu = 2$, and $\mu = -2$ in red, blue, and green, respectively.

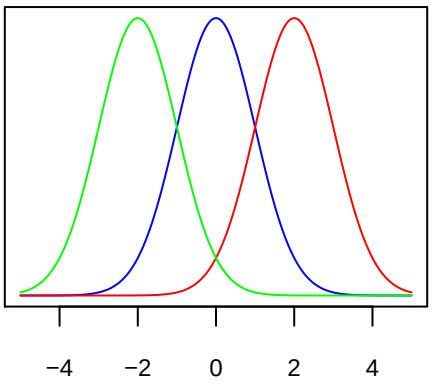


FIGURE 7.2.1. example of a normal location family

7.2.2. Scale families.

DEFINITION 7.18. Let $f(x)$ be any pdf. Then, for any $\sigma > 0$, the family of pdfs $(1/\sigma) f(x/\sigma)$, indexed by the parameter σ, is called the *scale family* with standard pdf $f(x)$ and σ is called the *scale parameter* of the family.

E.g., $f(x) \sim \mathcal{N}(0, 1^2)$, $\mathcal{N}(0, \sigma^2)$ is a scale family. The effect of introducing the scale parameter σ is either to stretch ($\sigma > 1$) or to contract ($\sigma < 1$) the graph of $f(x)$ while still maintaining the same basic shape of the graph. Figure 7.2.2 shows the normal distribution with $\mu = 0$ and $\sigma^2 = 1^2$, $\sigma^2 = 0.75^2$, and $\sigma^2 = 1.5^2$ in red, blue, and green, respectively.

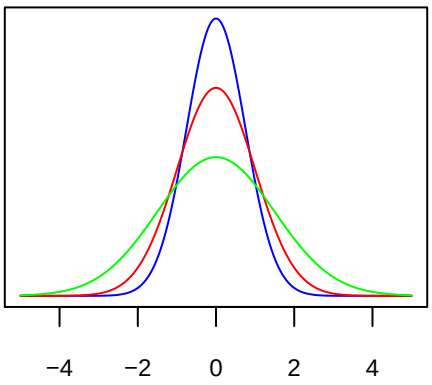


FIGURE 7.2.2. example of a normal scale family

7.2.3. Location-scale families. Let $f(x)$ be any pdf. Then, for any $-\infty < \mu < \infty$ and any $\sigma > 0$, the family of pdfs $(1/\sigma) f((x-\mu)/\sigma)$ is called the location-scale family with standard pdf $f(x)$. μ is called the location parameter and σ is called the scale parameter. E.g., $f(x) \sim \mathcal{N}(0, 1^2)$, $\mathcal{N}(\mu, \sigma^2)$ is a location-scale family. The effect of introducing both the location and scale parameters is to stretch/contract the graph with the scale parameter and shift the graph with the location parameter.

THEOREM 7.19. *Let $f(x)$ be any pdf. Let μ be any real number, and let σ be any positive real number. Then X is a random variable with pdf $(1/\sigma) f((x-\mu)/\sigma)$ if and only if there exists a random variable Z with pdf $f(z)$ and $X = \sigma Z + \mu$. (This is Theorem 3.5.6 from Casella & Berger.)*

PROOF. Suppose that $X \sim (1/\sigma) f((x-\mu)/\sigma)$. Then, the cdf of X is given by

$$\begin{aligned}
F_X(x) &= P(\{X \le x\}) \\
&= \int_{-\infty}^{x} f(x)\,\mathrm{d}\,x \\
&= \int_{-\infty}^{x} \frac{1}{\sigma} f\left(\frac{t-\mu}{\sigma}\right)\mathrm{d}\,t \\
(z = (t-\mu)/\sigma \implies \mathrm{d}\,t = \sigma\,\mathrm{d}\,z) \qquad &= \int_{-\infty}^{(x-\mu)/\sigma} \frac{1}{\sigma} f(z)\,\sigma\,\mathrm{d}\,z \\
&= \int_{-\infty}^{(x-\mu)/\sigma} f(z)\,\mathrm{d}\,z \\
&= P\left(\left\{Z \le \frac{x-\mu}{\sigma}\right\}\right) \\
(z = (x-\mu)/\sigma) \qquad &= P(\{Z \le z\}) \\
&= F_Z(z).
\end{aligned}$$

Thus, X will be a random variable with the specified pdf if and only if $Z = (X-\mu)/\sigma$ is also a random variable. □

THEOREM 7.20. *Let Z be a random variable with pdf $f(z)$. Suppose $\mathrm{E}[Z]$ and $\mathrm{Var}(Z)$ exist. If X is a random variable with pdf $(1/\sigma) f((x-\mu)/\sigma)$, then $\mathrm{E}[X] = \sigma\,\mathrm{E}[Z] + \mu$, and $\mathrm{Var}(X) = \sigma^2\,\mathrm{Var}(Z)$. (This is Theorem 3.5.7 from Casella and Berger; the version in the lecture slides is a slight restatement.)*

PROOF. By theorem 7.19, there is a random variable Z^* with pdf $f(z)$ and $X = \sigma Z^* + \mu$. So

$$\begin{aligned}
\mathrm{E}[X] &= \mathrm{E}[\sigma Z^* + \mu] \\
\text{(linearity)} \qquad &= \mathrm{E}[\sigma Z^*] + \mathrm{E}[\mu] \\
(\mu \text{ is constant}) \qquad &= \sigma\,\mathrm{E}[Z^*] + \mu \\
\text{(definition of expected value)} \qquad &= \sigma \int z^* \cdot f(z^*)\,\mathrm{d}\,z^* + \mu \\
(Z^* \text{ has the same pdf as } Z) \qquad &= \sigma \int z \cdot f(z)\,\mathrm{d}\,z + \mu \\
\text{(definition of expected value)} \qquad &= \sigma\,\mathrm{E}[Z] + \mu.
\end{aligned}$$

Then,

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(\sigma Z^* + \mu) \\
\text{(definition of variance)} \qquad &= \mathrm{E}\left[((\sigma Z^* + \mu) - \mathrm{E}[\sigma Z^* + \mu])^2\right] \\
\text{(linearity)} \qquad &= \mathrm{E}\left[(\sigma Z^* + \mu - (\sigma\,\mathrm{E}[Z^*] + \mu))^2\right] \\
&= \mathrm{E}\left[(\sigma(Z^* - \mathrm{E}[Z^*]))^2\right] \\
&= \mathrm{E}\left[\sigma^2 (Z^* - \mathrm{E}[Z^*])^2\right] \\
\text{(linearity)} \qquad &= \sigma^2\,\mathrm{E}\left[(Z^* - \mathrm{E}[Z^*])^2\right] \\
\text{(definition of variance)} \qquad &= \sigma^2\,\mathrm{Var}(Z^*) \\
(Z^* \text{ has the same pdf as } Z) \qquad &= \sigma^2\,\mathrm{Var}(Z),
\end{aligned}$$

and the result has been shown. □

CHAPTER 8

Principles of data reduction

8.1. Sufficiency

The concept of sufficiency attempts to find a statistic $T(X_1, \ldots, X_n)$ that contains all the information in the sample about the model parameter θ.

DEFINITION 8.1. A *statistic* is a function $T(X)$ of the data, such as mean, variance, max, or min. $T(X)$ is a random variable.

DEFINITION 8.2. An *estimate* is a statistic that is intended to be close to a parameter.

8.1.1. Data reduction. Any statistic $T(X)$ defines a form of data reduction or data summary. An investigator who uses only the observed value of the statistic rather than the entire observed sample $\mathbf{X}$ will treat as equal two samples $\mathbf{X}$ and $\mathbf{Y}$ that satisfy $T(x) = T(y)$ even though the actual sample values may be different in some ways. Data reduction in terms of a particular statistic can be thought of as a partition of the sample space $\mathcal{X}$ of $X_1, \ldots, X_n$. $T(X)$ reduces the data by partitioning the sample space into sets A_t, $t \in \mathcal{T}$, defined by $A_t = \{x : T(x) = t\}$ where $\mathcal{T} = \{t : t = T(x)\}$ for some $x \in \mathcal{X}$.

EXAMPLE 8.3 (Waiting time). Suppose that X represents waiting time, e.g., for a bus. We first collect data $x_1, \ldots, x_n$. Suppose that we wish to find the mean waiting time $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$, i.e., $T(X) = \bar{x}$. We calculate $\bar{x} = 8.32$. Then, all sets of $x_1, \ldots, x_n$ that give $\bar{x} = 8.32$ are a partition of the sample space. Many points in the sample space have this same mean, and we can consider them as belonging to the set $\{(x_1, \ldots, x_n) : \bar{x} = 8.32\}$, which is also the hyperplane $x_1 + \cdots + x_n = (8.32)\, n$. In this case, the sample mean $\bar{X}$ (or any statistic $T(X)$) partitions the sample space into a collection of sets.

EXAMPLE 8.4 (Sampling from a binomial distribution). Let $X \sim \text{Binomial}(3, p)$, where $p \in (0, 1)$. Then, X represents the number of successes in 3 trials of a random experiment. Suppose that we draw a sample of size 2, so that the sample space consists of

$$\{(0,0), (0,1), (0,2), (0,3), (1,1), (1,2), (1,3), (2,2), (2,3), (3,3)\}.$$

Let $T(\mathbf{X})$ be the maximum of the sample, i.e., $T(\mathbf{X}) = \max(X_1, X_2)$. Then, $T(\mathbf{X})$ reduces the data by partitioning the sample space into sets A_t with $\mathcal{T} = \{0, 1, 2, 3\}$, i.e.,

$$\begin{aligned} A_0 &= \{(0,0)\} \\ A_1 &= \{(0,1), (1,2)\} \\ A_2 &= \{(0,2), (1,2), (2,2)\} \\ A_3 &= \{(0,3), (1,3), (2,3), (3,3)\}. \end{aligned}$$

8.1.2. Sufficiency principle.

DEFINITION 8.5. A statistic $T(\mathbf{X})$ is a *sufficient statistic* for θ if the conditional distribution of the sample $\mathbf{X}$ given the value of $T(\mathbf{X})$ does not depend on θ, i.e., if $P(\mathbf{X}|T(\mathbf{X}) = T(\mathbf{x}))$ does not depend on θ.

A sufficient statistic for a parameter θ contains all information that is in the data about θ. I.e., given the value of T, the sufficient statistic, we can gain no more knowledge about θ from knowing more about the probability distribution of $X_1, \ldots, X_n$. If $\mathbf{X}$ and $\mathbf{X}$ are two samples and $T(\mathbf{X}) = T(\mathbf{Y})$, then inference about θ should be the same whether $\mathbf{X} = \mathbf{x}$ is observed or $\mathbf{Y} = \mathbf{y}$ is observed. Note that $T(\mathbf{X})$ must be a one-to-one function to be a sufficient statistic.

THEOREM 8.6. *If $p(\mathbf{x}|\theta)$ is the joint pdf or pmf of $\mathbf{X}$ and $q(T(\mathbf{x})|\theta)$ is the pdf or pmf of $T(\mathbf{X})$, then $T(\mathbf{X})$ is a sufficient statistic for θ if, for every $\mathbf{x}$ in the sample space, the ratio $p(\mathbf{x}|\theta)/q(T(\mathbf{x})|\theta))$ is constant as a function of θ. (This is Theorem 6.2.2 from Casella & Berger; the following proof is given there.)*

PROOF. By definition, if $T(\mathbf{X})$ is a sufficient statistic for θ, then $P(\mathbf{X}|T(\mathbf{X}))$ does not depend on θ .

$$P(\mathbf{X}|T(\mathbf{X})) = P(\{\mathbf{X}=\mathbf{x}\} \mid \{T(\mathbf{X})=T(\mathbf{x})\}) \tag{8.1.1}$$

$$= \frac{P(\left(\{X_1=x_1\}\cap\cdots\cap\{X_n=x_n\}\right)\cap\{T(\mathbf{X})=T(\mathbf{x})\})}{P(\{T(\mathbf{X})=T(\mathbf{x})\})} \tag{8.1.2}$$

$$= \frac{P(\{X_1=x_1\}\cap\cdots\cap\{X_n=x_n\})}{P(\{T(\mathbf{X})=T(\mathbf{x})\})} \tag{8.1.3}$$

$$= \frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x})|\theta))} \tag{8.1.4}$$

(1) is simply an expansion of the notation. (2) follows from the definition of conditional probability. (3) follows because $\{\mathbf{X}=\mathbf{x}\}$ is a subset of $\{T(\mathbf{X})=T(\mathbf{x})\}$, and if we have two sets A and B such that $A\subset B$, then $A\cap B=A$, so that $P(A\cap B)=P(A)$. (4) follows from setting $p(\mathbf{x}|\theta)$ equal to the numerator, which is the joint pdf of $\mathbf{X}$, and setting $q(T(\mathbf{X}|\theta))$ equal to the denominator, which is the pdf of $T(\mathbf{X})$. Then, if $p(\mathbf{x}|\theta)/q(T(\mathbf{x}|\theta))$ is constant as a function of θ, i.e., does not depend on θ, then $T(\mathbf{X})$ is a sufficient statistic for θ. □

EXAMPLE 8.7 (Sampling from a binomial distribution). Let X be as in example 8.4. If $T(\mathbf{X})$ is a sufficient statistic for p, show it, and if not, explain why not. We have $\mathbf{X}=X_1,X_2$, where the X_i's are independent and identically distributed because $\mathbf{X}$ is a random sample (by assumption). Then, each X_i has the same pmf as X, which is given by

$$p_X(x) = P(\{X=x\}) = \binom{3}{x} p^x (1-p)^{3-x}$$

where $x\in\{0,1,2,3\}$ and $p_X(x)=0$ otherwise, so that the joint pmf of $\mathbf{X}$ is given by

$$p_{\mathbf{X}}(\mathbf{x}) = P(\{\mathbf{X}=\mathbf{x}\}) = P(\{X_1=x_1\}\cap\{X_2=x_2\}) = P(\{X_1=x_1\})\cdot P(\{X_2=x_2\}) = \prod_{i=1}^{2} p_X(x_i)$$

Each X_i also has the same cdf as X, which is given by

$$F_X(t) = P(\{X\le t\}) = \sum_{k=0}^{t} p_X(k).$$

We have $T(\mathbf{X})=\max(X_1,X_2)$, which is just the order statistic $X_{(2)}$. For some maximum value $T(\mathbf{x})=t$, we have

$$\begin{aligned} P\left(\left\{X_{(2)}\le t\right\}\right) &= P(\{X_1\le t\}\cap\{X_2\le t\}) \\ &= P(\{X_1\le t\})\cdot P(\{X_2\le t\}) \\ &= \prod_{i=1}^{2} P(\{X_i\le t\}) \\ &= \prod_{i=1}^{2} F_X(t) \\ &= [F_X(t)]^2 \end{aligned}$$

so that the pmf of $X_{(2)}$ is given by

$$\begin{aligned} p_{X_{(2)}}(t) &= P\left(\left\{X_{(2)}=t\right\}\right) \\ &= P\left(\left\{X_{(2)}\le t\right\}\right) - P\left(\left\{X_{(2)}\le t-1\right\}\right) \\ &= F_{X_{(2)}}(t) - F_{X_{(2)}}(t-1) \\ &= [F_X(t)]^2 - [F_X(t-1)]^2 \end{aligned}$$

$$
\begin{aligned}
&= \left[\sum_{k=0}^{t} p_X(k)\right]^2 - \left[\sum_{k=0}^{t-1} p_X(k)\right]^2 \\
&= \left[\sum_{k=0}^{t} p_X(k)\right]^2 - \left[\sum_{k=0}^{t} p_X(k) - p_X(t)\right]^2 \\
&= \left[\sum_{k=0}^{t} p_X(k)\right]^2 - \left[\left(\sum_{k=0}^{t} p_X(k)\right)^2 - 2p_X(t)\sum_{k=0}^{t} p_X(k) + [p_X(t)]^2\right] \\
&= 2p_X(t)\sum_{k=0}^{t} p_X(k) - [p_X(t)]^2 \\
&= p_X(t)\left[2\sum_{k=0}^{t} p_X(k) - p_X(t)\right].
\end{aligned}
$$

Then, we have

$$
\begin{aligned}
P(\mathbf{X}|T(\mathbf{X})) &= \frac{P(\{\mathbf{X}=\mathbf{x}\}\cap\{T(\mathbf{X})=T(\mathbf{x})\})}{P(\{T(\mathbf{X})=T(\mathbf{x})\})} \\
&= \frac{P(\{\mathbf{X}=\mathbf{x}\})}{P(\{T(\mathbf{X})=T(\mathbf{x})\})} \\
&= \frac{\prod_{i=1}^{2} p_X(x_i)}{p_X(t)\left[2\sum_{k=0}^{t} p_X(k) - p_X(t)\right]} \\
&= \frac{\left[\binom{3}{x_1}p^{x_1}(1-p)^{3-x_1}\right]\left[\binom{3}{x_2}p^{x_2}(1-p)^{3-x_2}\right]}{\left[\binom{3}{t}p^{t}(1-p)^{3-t}\right]\left[2\sum_{k=0}^{t}\left(\binom{3}{k}p^{k}(1-p)^{3-k}\right) - \binom{3}{t}p^{t}(1-p)^{3-t}\right]}.
\end{aligned}
$$

Clearly, this expression depends on p, so $T(\mathbf{X}) = \max(\mathbf{X})$ is not a sufficient statistic for p.

EXAMPLE 8.8 (Sampling from a uniform distribution). Let $X \sim \mathcal{U}(0,\theta)$, and let $\mathbf{X} = X_1,\ldots,X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pdf of X. Let $T(\mathbf{X}) = \max(X_1,\ldots,X_n)$. Show that $T(\mathbf{X})$ is a sufficient statistic for θ.

From example 7.6, The pdf of X is given by

$$
f_X(x) = \frac{1}{\theta} I_{(0,\theta)}(x),
$$

so that the joint pdf of $\mathbf{X}$ is given by

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}) &= P(\{X_1=x_1\}\cap\cdots\cap\{X_n=x_n\}) \\
&= P(\{X_1=x_1\})\cdot\ldots\cdot P(\{X_n=x_n\}) \\
&= \prod_{i=1}^{n} f_X(x_i) \\
&= \left(\frac{1}{\theta}\right)^n \prod_{i=1}^{n} I_{(0,\theta)}(x_i) \\
&= \frac{1}{\theta^n}\left[I_{(0,\theta)}(x_1)\cdot I_{(0,\theta)}(x_2)\cdot\ldots I_{(0,\theta)}(x_n)\right] \\
&= \frac{1}{\theta^n} I_{(0,\theta)}\left(x_{(n)}\right),
\end{aligned}
$$

where the final equality follows from the fact that if $x_{(n)} = \max_{1\le i\le n} x_i \in (0,\theta)$, then $x_{(i)} \in (0,\theta)$, $1 \le i < n$. The cdf of X is given by

$$
F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}\,t
$$

$$
\begin{aligned}
&= \int_{-\infty}^{x} \frac{1}{\theta} I_{(0,\theta)}(t)\,\mathrm{d}\,t \\
&= \int_{-\infty}^{0} 0\,\mathrm{d}\,t + \int_{0}^{x} \frac{1}{\theta}\,\mathrm{d}\,t \\
&= 0 + \left(\frac{1}{\theta}\int_{0}^{x} 1\,\mathrm{d}\,t\right) \\
&= \frac{1}{\theta}\left(t\Big|_{0}^{x}\right) \\
&= \frac{1}{\theta}(x-0) \\
&= \frac{x}{\theta}.
\end{aligned}
$$

We have $T(\mathbf{X}) = \max(X_1, \ldots, X_n)$, which is just the order statistic $X_{(n)}$. By theorem 5.6, we have

$$
\begin{aligned}
f_{X_{(n)}}(t) &= \frac{n!}{(j-1)!\,(n-j)!} f_X(t)\,[F_X(t)]^{j-1}\,[1-F_X(t)]^{n-j} \\
&= \frac{n!}{(n-1)!\,(n-n)!} f_X(t)\,[F_X(t)]^{n-1}\,[1-F_X(t)]^{n-n} \\
&= \frac{n!}{(n-1)!}\left(\frac{1}{\theta} I_{(0,\theta)}(t)\right)\left[\frac{t}{\theta}\right]^{n-1}\left[1-\frac{t}{\theta}\right]^{0} \\
&= n\left(\frac{1}{\theta}\right)\left(\frac{t}{\theta}\right)^{n-1} I_{(0,\theta)}(t) \\
&= n\left(\frac{1}{\theta}\right)\frac{t^{n-1}}{\theta^{n-1}} I_{(0,\theta)}(t) \\
&= nt^{n-1}\left(\frac{1}{\theta}\right)\left(\frac{1}{\theta^{n-1}}\right) I_{(0,\theta)}(t) \\
&= nt^{n-1}\left(\frac{1}{\theta^{n}}\right) I_{(0,\theta)}(t).
\end{aligned}
$$

Let $p(\mathbf{x}|\theta)$ be the joint pdf of $\mathbf{X}$, and let $q(T(\mathbf{X})|\theta)$ be the pdf of $T(\mathbf{X})$. Then, by theorem 8.6, we have

$$
\begin{aligned}
\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{X})|\theta)} &= \frac{(1/\theta^n)\, I_{(0,\theta)}\left(x_{(n)}\right)}{nt^{n-1}\,(1/\theta^n)\, I_{(0,\theta)}(t)} \\
&= \frac{1}{nx^{n-1}} \frac{I_{(0,\theta)}\left(x_{(n)}\right)}{I_{(0,\theta)}(t)} \\
(t = x_{(n)}) \qquad &= \frac{1}{nx^{n-1}} \frac{I_{(0,\theta)}\left(x_{(n)}\right)}{I_{(0,\theta)}\left(x_{(n)}\right)} \\
&= \frac{1}{nx^{n-1}}.
\end{aligned}
$$

This expression does not depend on θ, so it follows that $p(\mathbf{x}|\theta)/q(T(\mathbf{X})|\theta)$ is constant as a function of θ, so that $T(\mathbf{X})$ is a sufficient statistic for θ.

EXAMPLE 8.9 (Sampling from a Poisson distribution). Let $X \sim \text{Poisson}(\lambda)$, and let $\mathbf{X} = X_1, X_2$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pmf of X. Let $T(\mathbf{X}) = X_1 + X_2$. Show that $T(\mathbf{X})$ is a sufficient statistic for λ.

The pmf of X is given by

$$
p_X(x) = \frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}
$$

where $x \in \{0, 1, 2, \ldots\}$ and $p_X(x) = 0$ otherwise, so that the joint pmf of $\mathbf{X}$ is given by

$$
p_{\mathbf{X}}(\mathbf{x}) = P(\{X_1 = x_1\} \cap \{X_2 \cap x_x\})
$$

$$
\begin{aligned}
&= P\left(\{X_2 = x_1\}\right) \cdot P\left(\{X_2 = x_2\}\right) \\
&= p_{X_1}\left(x_1\right) \cdot p_{X_2}\left(x_2\right) \\
&= \left(\frac{\mathrm{e}^{-\lambda}\lambda^{x_1}}{x_1!}\right)\left(\frac{\mathrm{e}^{-\lambda}\lambda^{x_2}}{x_2!}\right) \\
&= \frac{\mathrm{e}^{-2\lambda}\lambda^{x_1+x_2}}{x_1!x_2!}.
\end{aligned}
$$

Let $z = x_1 + x_2$. Then, the pmf of $T\left(\mathbf{X}\right)$ is given by

$$P\left(\{X_1 + X_2 = z\}\right) = \sum_{k=0}^{z} P\left(\{X_1 = k\} \cap \{X_2 = z - k\}\right) \tag{8.1.5}$$

$$= \sum_{k=0}^{z} P\left(\{X_1 = k\}\right) \cdot P\left(\{X_2 = z - k\}\right) \tag{8.1.6}$$

$$= \sum_{k=0}^{z} \left(\frac{\mathrm{e}^{-\lambda}\lambda^{k}}{k!}\right)\left(\frac{\mathrm{e}^{-\lambda}\lambda^{z-k}}{(z-k)!}\right) \tag{8.1.7}$$

$$= \sum_{k=0}^{z} \frac{\mathrm{e}^{-2\lambda}\lambda^{z}}{k!\,(z-k)!} \tag{8.1.8}$$

$$= \mathrm{e}^{-2\lambda}\lambda^{z} \sum_{k=0}^{z} \frac{1}{k!\,(z-k)!} \tag{8.1.9}$$

$$= \frac{\mathrm{e}^{-2\lambda}\lambda^{z}}{z!} \sum_{k=0}^{z} \frac{z!}{k!\,(z-k)!} \tag{8.1.10}$$

$$= \frac{\mathrm{e}^{-2\lambda}\lambda^{z}}{z!} \cdot 2^{z} \tag{8.1.11}$$

$$= \frac{\mathrm{e}^{-2\lambda}\left(2\lambda\right)^{z}}{z!} \tag{8.1.12}$$

$$= \frac{\mathrm{e}^{-2\lambda}\left(2\lambda\right)^{x_1+x_2}}{(x_1+x_2)!} \tag{8.1.13}$$

so that $T\left(\mathbf{X}\right) = X_1 + X_2 \sim P\left(2\lambda\right)$. (1) follows from Proposition 6.18 in Weiss' text, which gives the pmf of the sum of two discrete random variables. (2) follows because X_1 and X_2 are independent. (3) follows because X_1 and X_2 each has the pmf of X. (4) and (5) follow from algebra. (6) follows by multiplying by $1 = z!/z!$ and rearrangement. (7) follows from the binomial theorem, which gives

$$\left(x + y\right)^{n} = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^{k}$$

so that

$$2^{z} = \left(1 + 1\right)^{z} = \sum_{k=0}^{z} \binom{z}{k} 1^{z-k} 1^{k} = \sum_{k=0}^{z} \binom{z}{k} = \sum_{k=0}^{z} \frac{z!}{k!\,(z-k)!}.$$

(8) and (9) follow from algebra. Let $p\left(\mathbf{x}|\lambda\right)$ be the joint pmf of $\mathbf{X}$, and let $q\left(T\left(\mathbf{X}\right)|\lambda\right)$ be the pmf of $T\left(\mathbf{X}\right)$. Then, by theorem 8.6, we have

$$
\begin{aligned}
\frac{p\left(\mathbf{x}|\lambda\right)}{q\left(T\left(\mathbf{X}\right)|\lambda\right)} &= \frac{\frac{\mathrm{e}^{-2\lambda}\lambda^{x_1+x_2}}{x_1!x_2!}}{\frac{\mathrm{e}^{-2\lambda}(2\lambda)^{x_1+x_2}}{(x_1+x_2)!}} \\
&= \frac{\lambda^{x_1+x_2}\left(x_1 + x_2\right)!}{\left(2\lambda\right)^{x_1+x_2} x_1!x_2!} \\
&= \frac{\lambda^{x_1+x_2}\left(x_1 + x_2\right)!}{2^{x_1+x_2}\lambda^{x_1+x_2} x_1!x_2!}
\end{aligned}
$$

$$= \frac{(x_1+x_2)!}{x_1!x_2!}\left(\frac{1}{2}\right)^{x_1+x_2}.$$

Clearly, this expression does not depend on λ, so it follows that $p(\mathbf{x}|\lambda)/q(T(\mathbf{X})|\lambda)$ is constant as a function of λ, so that $T(\mathbf{X})$ is a sufficient statistic for λ.

EXAMPLE 8.10 (Sampling from a Bernoulli distribution). Let $X \sim \text{Bernoulli}(p)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pmf of X. Let $T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$ (this is the MLE $\hat{p}$). Show that $T(\mathbf{X})$ is a sufficient statistic for p.

A Bernoulli random variable is a binomial random variable with $n = 1$, so the pmf of X is given by

$$p_X(x) = \binom{1}{x} p^x (1-p)^{1-x} = p^x (1-p)^{1-x}$$

where $x \in \{0,1\}$ and $p_X(x) = 0$ otherwise. Then, the joint pmf of $\mathbf{X}$ is given by

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) &= P(\{X_1 = x_1\} \cap \cdots \cap \{X_n = x_n\}) \\ &= P(\{X_1 = x_1\}) \cdot \ldots \cdot P(\{X_n = x_n\}) \\ &= \prod_{i=1}^n p_X(x_i) \\ &= \prod_{i=1}^n p^{x_i}(1-p)^{1-x_i} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{\sum_{i=1}^n (1-x_i)} \\ &= p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}. \end{aligned}$$

Proposition 6.20 from Weiss' text states that if $Y_1, \ldots, Y_m$ are independent random variables with $Y_j \sim \mathcal{B}(n_j, p)$ for $1 \le j \le m$, then $Y_1 + \cdots + Y_m \sim \mathcal{B}(n_1 + \cdots n_m, p)$. It follows that

$$\sum_{i=1}^n X_i \sim \mathcal{B}\left(\sum_{i=1}^n 1, p\right) = \mathcal{B}(n,p).$$

We have $T(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^n X_i$, so it follows that $nT(\mathbf{X}) = \sum_{i=1}^n X_i$, so that $nT(\mathbf{X}) \sim \mathcal{B}(n,p)$. Then, the pmf of $nT(\mathbf{X})$ is given by

$$P\left(\left\{nT(\mathbf{X}) = \sum_{i=1}^n x_i\right\}\right) = \binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}.$$

Let $p(\mathbf{x}|p)$ be the joint pmf of $\mathbf{X}$, and let $q(nT(\mathbf{X})|p)$ be the pmf of $nT(\mathbf{X})$. Then, by theorem 8.6, we have

$$\frac{p(\mathbf{x}|p)}{q(nT(\mathbf{X})|p)} = \frac{p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}}{\binom{n}{\sum_{i=1}^n x_i} p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i}} = \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}.$$

Clearly, this expression does not depend on p, so it follows that $nT(\mathbf{X})$ is a sufficient statistic for p, so that $T(\mathbf{X}) = \hat{p}$ is a sufficient statistic for p.

8.1.3. Factorization theorem.

THEOREM 8.11 (Factorization Theorem). *Let $f(\mathbf{x}|\theta)$ be the joint pdf or pmf of a sample $\mathbf{X}$. A statistic $T(\mathbf{X})$ is sufficient for θ if and only if there exist functions $g(t|\theta)$ and $h(\mathbf{x})$ such that $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)\, h(\mathbf{x})$ for all sample points $\mathbf{x}$ and all θ. (This is Theorem 6.2.6 from Casella and Berger; the following proof is given there.)*

PROOF. We will give the proof only for the discrete case.

Suppose $T(\mathbf{X})$ is a sufficient statistic. Choose $g(t|\theta) = P_\theta(\{T(\mathbf{X}) = t\})$ and $h(\mathbf{x}) = P(\{\mathbf{X} = \mathbf{x}|T(\mathbf{X}) = T(\mathbf{x})\})$. Because $T(\mathbf{X})$ is sufficient, the conditional probability defining $h(\mathbf{x})$ does not depend on θ. Thus, this choice of $h(\mathbf{x})$ and $g(t|\theta)$ is legitimate, and for this choice we have

$$f(\mathbf{x}|\theta) = P_\theta(\{\mathbf{X} = \mathbf{x}\}) \tag{8.1.14}$$

$$= P_\theta\left(\{\mathbf{X}=\mathbf{x}\}\cap\{T(\mathbf{X})=T(\mathbf{x})\}\right) \tag{8.1.15}$$

$$= P_\theta\left(\{T(\mathbf{X})=T(\mathbf{x})\}\right)\cdot P\left(\{\mathbf{X}=\mathbf{x}\}\,|\,\{T(\mathbf{X})=T(\mathbf{x})\}\right) \tag{8.1.16}$$

$$= g\left(T(\mathbf{x})\,|\theta\right)h(\mathbf{x}). \tag{8.1.17}$$

(1) follows from the definition of a pmf. (2) follows because $\{\mathbf{X}=\mathbf{x}\}$ is a subset of $\{T(\mathbf{X})=T(\mathbf{x})\}$, and if we have two sets A and B such that $A\subset B$, then $A\cap B=A$, so that $P(A\cap B)=P(A)$. (3) follows from the definition of conditional probability. (4) follows from our definitions of $g(t|\theta)$ and $h(\mathbf{x})$. So factorization has been exhibited. We also see from the last two lines above that $P_\theta\left(\{T(\mathbf{X})=T(\mathbf{x})\}\right)=g\left(T(\mathbf{x})\,|\theta\right)$, so $g\left(T(\mathbf{x})\,|\theta\right)$ is the pmf of $T(\mathbf{X})$.

Now assume the factorization $f(\mathbf{x}|\theta)=g\left(T(\mathbf{x})\,|\theta\right)h(\mathbf{x})$ exists. Let $q(t|\theta)$ be the pmf of $T(\mathbf{X})$. To now show that $T(\mathbf{X})$ is sufficient, we examine the ratio $f(\mathbf{x}|\theta)/q\left(T(\mathbf{x})\,|\theta\right)$. Define $A_{T(\mathbf{x})}=\{\mathbf{y}:T(\mathbf{y})=T(\mathbf{x})\}$. Then

$$\frac{f(\mathbf{x}|\theta)}{q\left(T(\mathbf{x})\,|\theta\right)}=\frac{g\left(T(\mathbf{x})\,|\theta\right)h(\mathbf{x})}{q\left(T(\mathbf{x})\,|\theta\right)} \tag{8.1.18}$$

$$=\frac{g\left(T(\mathbf{x})\,|\theta\right)h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}}g\left(T(\mathbf{y})\,|\theta\right)h(\mathbf{y})} \tag{8.1.19}$$

$$=\frac{g\left(T(\mathbf{x})\,|\theta\right)h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}}g\left(T(\mathbf{x})\,|\theta\right)h(\mathbf{y})} \tag{8.1.20}$$

$$=\frac{g\left(T(\mathbf{x})\,|\theta\right)h(\mathbf{x})}{g\left(T(\mathbf{x})\,|\theta\right)\sum_{A_{T(\mathbf{x})}}h(\mathbf{y})} \tag{8.1.21}$$

$$=\frac{h(\mathbf{x})}{\sum_{A_{T(\mathbf{x})}}h(\mathbf{y})}. \tag{8.1.22}$$

(1) follows from our assumption that the factorization exists. (2) follows from the definition of the pmf of T, which is given by

$$\begin{aligned}
f_{T(\mathbf{X})}\left(T(\mathbf{x})\right)&=P\left(\{T(\mathbf{X})=T(\mathbf{x})\}\right)\\
&=\sum_{\mathbf{y}\in T^{-1}(\{T(\mathbf{x})\})}P\left(\{\mathbf{X}=\mathbf{y}\}\right)\\
&=\sum_{\mathbf{y}\in T^{-1}(\{T(\mathbf{x})\})}f_{\mathbf{X}}(\mathbf{y})\\
&=\sum_{\mathbf{y}\in\{\mathbf{y}:T(\mathbf{y})=T(\mathbf{x})\}}f_{\mathbf{X}}(\mathbf{y})\\
&=\sum_{\mathbf{y}\in A_{T(\mathbf{x})}}g\left(T(\mathbf{y})\,|\theta\right)h(\mathbf{y}).
\end{aligned}$$

(3) and (4) follow from the fact that T is constant on $A_{T(\mathbf{x})}$, i.e., $T(\mathbf{x})=T(\mathbf{y})$. (5) follows from algebra. Because (5) does not depend on θ, by theorem 8.6 $T(\mathbf{X})$ is a sufficient statistic for θ. □

Example 8.12 (Sufficient statistic for a Poisson random variable). Let $X\sim\text{Poisson}(\lambda)$, and let $\mathbf{X}=X_1,\ldots,X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pmf of X, which is given by

$$p_X(x|\lambda)=\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}$$

where $x\in\{0,1,2,\ldots\}$ and $p_X(x)=0$ otherwise, so that the joint pmf of $\mathbf{X}$ is given by

$$\begin{aligned}
p_{\mathbf{X}}(\mathbf{x}|\lambda)&=\prod_{i=1}^{n}p_{X_i}(x_i)\\
&=\prod_{i=1}^{n}\left(\frac{\mathrm{e}^{-\lambda}\lambda^{x_i}}{x_i!}\right)
\end{aligned}$$

$$
\begin{aligned}
&= \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \left(\prod_{i=1}^{n} \mathrm{e}^{-\lambda} \lambda^{x_i}\right) \\
&= \left(\prod_{i=1}^{n} \frac{1}{x_i!}\right) \mathrm{e}^{-n\lambda} \lambda^{x_1+\cdots+x_n} \\
&= \underbrace{\left(\prod_{i=1}^{n} \frac{1}{x_i!}\right)}_{h(\mathbf{x})} \underbrace{\mathrm{e}^{-n\lambda} \lambda^{\sum_{i=1}^{n} x_i}}_{g(T(\mathbf{x})|\lambda)}.
\end{aligned}
$$

It follows from theorem 8.11 that $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for λ.

EXAMPLE 8.13 (Sufficient statistic for a Bernoulli random variable). Consider a sequence of independent Bernoulli random variables $\mathbf{X} = X_1, \ldots, X_n$ where $P(\{X_i = x\}) = \theta^x (1-\theta)^{1-x}$ where $x \in \{0, 1\}$. Then, the joint pmf of the X_i's is given by

$$
\begin{aligned}
p_{(X_1,\ldots,X_n|\theta)}(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \left(\theta^{x_i} (1-\theta)^{1-x_i}\right) \\
&= \theta^{\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} \left[(1-\theta)^1 (1-\theta)^{-x_i}\right] \\
&= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^n (1-\theta)^{-\sum_{i=1}^{n} x_i} \\
&= \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^n \left(\frac{1}{1-\theta}\right)^{\sum_{i=1}^{n} x_i} \\
&= \underbrace{(1-\theta)^n \left(\frac{\theta}{1-\theta}\right)^{\sum_{i=1}^{n} x_i}}_{g(T(\mathbf{x})|\theta)}.
\end{aligned}
$$

We set $h(\mathbf{x}) = 1$. Then, it follows from theorem 8.11 that $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ is a sufficient statistic for θ.

EXAMPLE 8.14 (Normal sufficient statistic). Let $X \sim \mathcal{N}(\mu, \sigma^2)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pdf of X, which is given by

$$
f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}
$$

for $x \in \mathbb{R}$, where $\mu \in \mathbb{R}$ and $\sigma > 0$. Find a sufficient statistic for $\boldsymbol{\theta} = (\mu, \sigma^2)$.
A joint pdf for $\mathbf{X}$ is given by

$$
\begin{aligned}
f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) &= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\left(\frac{(x_i-\mu)^2}{\sigma^2}\right)\right\}\right] \\
&= \prod_{i=1}^{n} \left[(2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{(x_i-\mu)^2}{\sigma^2}\right)\right\}\right] \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{\sum_{i=1}^{n} -\frac{1}{2}\left(\frac{(x_i-\mu)^2}{\sigma^2}\right)\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{x_i^2}{\sigma^2} - \frac{2\mu x_i}{\sigma^2} + \frac{\mu^2}{\sigma^2}\right)\right\} \\
&= (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}\left(\frac{\sum_{i=1}^{n} x_i^2}{\sigma^2} - \frac{2\mu\sum_{i=1}^{n} x_i}{\sigma^2} + \frac{n\mu^2}{\sigma^2}\right)\right\} \\
&= \underbrace{(2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{n\mu^2}{2\sigma^2}\right\} \exp\left\{-\frac{\sum_{i=1}^{n} x_i^2}{2\sigma^2} + \frac{\mu\sum_{i=1}^{n} x_i}{\sigma^2}\right\}}_{g(T(\mathbf{x})|\theta)}.
\end{aligned}
$$

We set $h(\mathbf{x}) = 1$. Then, it follows from theorem 8.11 that $T(\mathbf{X}) = \left(\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i\right)$ is a sufficient statistic for $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$.

EXAMPLE 8.15 (Uniform sufficient statistic). Let $X \sim \mathcal{U}(0, \theta)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pdf of X, which is given by

$$f_X(x|\theta) = \frac{1}{\theta}$$

where $x \in (0, \theta)$ and $f(x|\theta) = 0$ otherwise, so that a joint pdf for $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^n \left(\frac{1}{\theta} I_{\{0<x_i<\theta\}}\right) = \left(\frac{1}{\theta}\right)^n I_{\left\{x_{(1)}>0\right\}} I_{\left\{x_{(n)}<\theta\right\}}.$$

We have $h(\mathbf{x}) = I_{\left\{x_{(1)}>0\right\}}$ and $g(T(\mathbf{x})|\theta) = \theta^{-n} I_{\left\{x_{(n)}<\theta\right\}}$. It follows from theorem 8.11 that $T(\mathbf{x}) = x_{(n)} = \max_{1\le i\le n} x_i$ is a sufficient statistic for θ.

THEOREM 8.16. *Let $X_1, \ldots, X_n$ be iid observations from a pdf or pmf $f(x|\boldsymbol{\theta})$ that belongs to an exponential family given by*

$$f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^k \omega_i(\boldsymbol{\theta})\, t_i(x)\right)$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$, $d \le k$. Then

$$T(\mathbf{X}) = \left(\sum_{j=1}^n t_1(X_j), \ldots, \sum_{j=1}^n t_k(X_j)\right)$$

*is a sufficient statistic for $\boldsymbol{\theta}$. $T(\mathbf{X})$ is called the **natural sufficient statistic.** (This is Theorem 6.2.10 from Casella & Berger.)*

PROOF. The joint pdf of $\mathbf{X}$ is given by

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{j=1}^n \left[h(x_j)\, c(\theta) \exp\left(\sum_{i=1}^k \omega_i(\theta)\, t_i(x_j)\right)\right] \\ &= \underbrace{c(\theta)^n \exp\left\{\sum_{i=1}^k \left(\omega_i(\theta) \sum_{j=1}^n t_i(x_j)\right)\right\}}_{g(T(\mathbf{x})|\theta)} \underbrace{\prod_{j=1}^n h(x_j)}_{h(\mathbf{x})}. \end{aligned}$$

It follows from theorem 8.11 that $T(\mathbf{X}) = \left(\sum_{j=1}^n t_1(X_j), \ldots, \sum_{j=1}^n t_k(X_j)\right)$ is a sufficient statistic for θ. □

8.1.4. Minimal sufficient statistics. In the preceding section, we found one sufficient statistic for each model considered. In fact, there are many sufficient statistics. It is always true that the data $\mathbf{X}$ is a sufficient statistic; we can factor the pdf of $\mathbf{X}$ as $f(\mathbf{x}|\theta) = f(T(\mathbf{X})|\theta)\, h(\mathbf{x})$, where $T(\mathbf{x}) = \mathbf{x}$ and $h(\mathbf{x}) = 1$ for all $\mathbf{x}$. Any one-to-one function of a sufficient statistic is sufficient. We might ask whether one sufficient statistic is any better than another.

DEFINITION 8.17. A sufficient statistic $T(\mathbf{X})$ is called a *minimal sufficient statistic* if, for any other sufficient statistic $T'(\mathbf{X})$, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$, i.e., $T(\mathbf{x}) = g(T'(\mathbf{x}))$.

Of all sufficient statistics, a minimal sufficient statistic provides the greatest reduction of the data. In terms of the partition sets described above, if $\{B_{t'} : t' \in \mathcal{T}'\}$ are the partition sets of $T'(\mathbf{x})$ and $\{A_t : t \in \mathcal{T}\}$ are the partition sets for $T(\mathbf{x})$, then every $B_{t'}$ is a subset of A_t. The partition associated with the minimal sufficient statistic is the coarest possible partition (among those induced by sufficient statistics).

THEOREM 8.18. *Let $f(\mathbf{x}|\theta)$ be the pmf or pdf of a sample $\mathbf{X}$. Suppose there exists a function $T(\mathbf{x})$ such that, for every two sample points $\mathbf{x}$ and $\mathbf{y}$, the ratio $f(\mathbf{x}|\theta)/f(\mathbf{y}|\theta)$ is constant as a function of θ if and only if $T(\mathbf{x}) = T(\mathbf{y})$. Then $T(\mathbf{X})$ is a minimal sufficient statistic for θ. (This is Theorem 6.2.13 from Casella & Berger; the following proof is given there.)*

PROOF. To simplify the proof, we assume $f(x|\theta) > 0$ for all $x \in \mathcal{X}$ and θ.

First, we show that $T(\mathbf{X})$ is a sufficient statistic. Let $\mathcal{T} = \{t : t = T(\mathbf{x}), \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $T(\mathbf{x})$. Define the partition sets induced by $T(\mathbf{x})$ as $A_t = \{\mathbf{x} : T(\mathbf{x}) = t\}$. For each A_t, choose and fix one element $\mathbf{x}_t \in A_t$. For any $\mathbf{x} \in \mathcal{X}$, $\mathbf{x}_{T(\mathbf{x})}$ is the fixed element that is in the same set, A_t, as $\mathbf{x}$. Since $\mathbf{x}$ and $\mathbf{x}_{T(\mathbf{x})}$ are in the same set A_t, $T(\mathbf{x}) = T\left(\mathbf{x}_{T(\mathbf{x})}\right)$ and, hence, $f(x|\theta)/f\left(\mathbf{x}_{T(\mathbf{x})}|\theta\right)$ is constant as a function of θ. Thus, we can define a function on $\mathcal{X}$ by $h(\mathbf{x}) = f(\mathbf{x}|\theta)/f\left(\mathbf{x}_{T(\mathbf{x})}|\theta\right)$ and h does not depend on θ. Define a function on $\mathcal{T}$ by $g(t|\theta) = f(\mathbf{x}_t|\theta)$. Then it can be seen that

$$f(\mathbf{x}|\theta) = \frac{f\left(\mathbf{x}_{T(\mathbf{x})}|\theta\right) f(\mathbf{x}|\theta)}{f\left(\mathbf{x}_{T(\mathbf{x})}|\theta\right)} = g(T(\mathbf{x})|\theta)\, h(\mathbf{x})$$

and, by theorem 8.11, $T(\mathbf{X})$ is a sufficient statistic for θ.

Now to show that $T(\mathbf{X})$ is minimal, let $T'(\mathbf{X})$ be any other sufficient statistic. By theorem 8.11, there exist functions g' and h' such that $f(\mathbf{x}|\theta) = g'(T'(\mathbf{x})|\theta)\, h'(\mathbf{x})$. Let $\mathbf{x}$ and $\mathbf{y}$ be any two sample points with $T'(\mathbf{x}) = T'(\mathbf{y})$. Then

$$\frac{f(\mathbf{x}|\theta)}{f(\mathbf{y}|\theta)} = \frac{g'(T'(\mathbf{x})|\theta)\, h'(\mathbf{x})}{g'(T'(\mathbf{y})|\theta)\, h'(\mathbf{y})} = \frac{h'(\mathbf{x})}{h'(\mathbf{y})}.$$

Since this ratio does not depend on θ, the assumptions of the theorem imply that $T(\mathbf{x}) = T(\mathbf{y})$. Thus, $T(\mathbf{x})$ is a function of $T'(\mathbf{x})$ and $T(\mathbf{x})$ is minimal. □

COROLLARY 8.19. *If the partition of the sample space induced by $f(x|\theta)/f(y|\theta)$ is equivalent to that induced by T, then T is minimal sufficient.*

EXAMPLE 8.20 (Bernoulli minimal sufficient statistic). Let $X_1, X_2 \sim \text{Bernoulli}(p)$. Let $V = X_1$, $T = \sum_i X_i$, and $U = (T, X_1)$. Determine whether V, T, or U is a minimal sufficient statistic for p.

The set of outcomes and the statistics are shown below.

X_1	X_2	V	T	U
0	0	0	0	$(0,0)$
0	1	0	1	$(1,0)$
1	0	1	1	$(1,1)$
1	1	1	2	$(2,1)$

Let $\mathcal{V} = \{v : v = x_1, \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$, the sample space of $\mathbf{X}$, under V. From the table above, we have $\mathcal{V} = \{0, 1\}$. Let $W_t = \{\mathbf{x} : X_1 = v\}$ be the partition sets induced by V, so that we have

$$W_0 = \{(0,0), (0,1)\}$$
$$W_1 = \{(1,0), (1,1)\}.$$

The joint pmf of $\mathbf{X} = X_1, X_2$ conditioned on V is given by

$$\begin{aligned} p_{(\mathbf{X}|V)}(\mathbf{x}|v) &= P(\{\mathbf{X} = \mathbf{x}\} \mid \{V = v\}) \\ &= \frac{p_{\mathbf{X}}(\mathbf{x}|p)}{p_V(v|p)} \\ &= \frac{\left(p^{x_1}(1-p)^{1-x_1}\right)\left(p^{x_2}(1-p)^{1-x_2}\right)}{p^{x_1}(1-p)^{1-x_1}} \\ &= p^{x_2}(1-p)^{1-x_2}. \end{aligned}$$

Clearly, this expression depends on p, so V is not a sufficient statistic for p.

Let $\mathcal{T} = \{t : t = \sum_i x_i, \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under T. From the table above, we have $\mathcal{T} = \{0, 1, 2\}$. Let $A_t = \{\mathbf{x} : \sum_i x_i = t\}$ be the partition sets induced by T, so that we have

$$A_0 = \{(0,0)\}$$
$$A_1 = \{(0,1), (1,0)\}$$
$$A_2 = \{(1,1)\}.$$

We showed in example 8.13 that T is a sufficient statistic for p. We will now show that T is minimal. From example 8.13, the pmf of $\mathbf{X}$ can be written as $p_{\mathbf{X}}(\mathbf{x}|p) = g(T(\mathbf{x})|p)\,h(\mathbf{x})$, where $h(\mathbf{x}) = 1$ and

$$g(T(\mathbf{x})|p) = (1-p)^n \left(\frac{p}{1-p}\right)^{\sum_{i=1}^n x_i} = (1-p)^2 \left(\frac{p}{1-p}\right)^{x_1+x_2}.$$

Then, we have

$$\begin{aligned}\frac{p_{\mathbf{X}}(\mathbf{x}|p)}{p_{\mathbf{X}}(\mathbf{y}|p)} &= \frac{g(T(\mathbf{x})|p)\,h(\mathbf{x})}{g(T(\mathbf{y})|p)\,h(\mathbf{y})} \\ &= \frac{\left[(1-p)^2\left(\frac{p}{1-p}\right)^{x_1+x_2}\right]\cdot 1}{\left[(1-p)^2\left(\frac{p}{1-p}\right)^{y_1+y_2}\right]\cdot 1} \\ &= \frac{\left(\frac{p}{1-p}\right)^{x_1+x_2}}{\left(\frac{p}{1-p}\right)^{y_1+y_2}}.\end{aligned}$$

We have $p \in (0,1)$, so that $p/(1-p) > 0$, so that the ratio shown above will be defined. This ratio will be constant as a function of p if and only if $x_1 + x_2 = y_1 + y_2$, i.e., if $T(\mathbf{x}) = T(\mathbf{y})$. It follows from theorem 8.18 that T is a minimal sufficient statistic for p.

Let $\mathcal{U} = \{\mathbf{u} : \mathbf{u} = (T(\mathbf{x}), x_1), \mathbf{x} \in \mathcal{X}\}$ be the image of $\mathcal{X}$ under $\mathbf{U}$. From the table above, we have $\mathcal{U} = \{(0,0),(1,0),(1,1),(2,1)\}$. Let $B_t = \{\mathbf{x} : (T(\mathbf{x}), x_1) = \mathbf{u}\}$ be the partition sets induced by $\mathbf{U}$, so that we have

$$\begin{aligned}B_{(0,0)} &= \{(0,0)\} \\ B_{(1,0)} &= \{(1,0)\} \\ B_{(1,1)} &= \{(0,1)\} \\ B_{(2,1)} &= \{(1,1)\}.\end{aligned}$$

The pmf of $\mathbf{U}$ is given by

$$\begin{aligned}p_{\mathbf{U}}(\mathbf{u}|p) &= P(\{T(\mathbf{X}) = t\} \cap \{X_1 = x_1\}) \\ &= P(\{T(\mathbf{X}) = t\} \,|\, \{X_1 = x_1\}) \cdot P(\{X_1 = x_1\}) \\ &= \left(\frac{p_{T(\mathbf{x})}(t|p)}{p_{X_1}(x_1|p)}\right) \cdot p_{X_1}(x_1|p) \\ &= p_{T(\mathbf{x})}(t|p).\end{aligned}$$

To show that $\mathbf{U}$ is a sufficient statistic for p, we examine the ratio

$$\begin{aligned}\frac{p(\mathbf{x}|\theta)}{q(T(\mathbf{x}|\theta))} &= \frac{p_{\mathbf{X}}(\mathbf{x}|p)}{p_{\mathbf{U}}(\mathbf{u}|p)} \\ &= \frac{\left(p^{x_1}(1-p)^{1-x_1}\right)\left(p^{x_2}(1-p)^{1-x_2}\right)}{(1-p)^2\left(\frac{p}{1-p}\right)^{x_1+x_2}} \\ &= \frac{p^{x_1+x_2}(1-p)^{2-x_1-x_2}}{(1-p)^2 p^{x_1+x_2}(1-p)^{-x_1-x_2}} \\ &= \frac{p^{x_1+x_x}(1-p)^{2-x_1-x_2}}{p^{x_1+x_2}(1-p)^{2-x_1-x_x}} \\ &= 1.\end{aligned}$$

Because the ratio $p_{\mathbf{X}}(\mathbf{x}|p)/p_{\mathbf{U}}(\mathbf{u}|p)$ is constant as a function of p, it follows from theorem 8.6 that $\mathbf{U}$ is a sufficient statistic for p. We will now check whether $\mathbf{U}$ is minimal. We showed above that

$$\frac{p_{\mathbf{X}}(\mathbf{x}|p)}{p_{\mathbf{X}}(\mathbf{y}|p)} = \frac{\left(\frac{p}{1-p}\right)^{x_1+x_2}}{\left(\frac{p}{1-p}\right)^{y_1+y_2}}.$$

By theorem 8.18, if this ratio is constant if and only if $\mathbf{U}(\mathbf{x}) = \mathbf{U}(\mathbf{y})$, then $\mathbf{U}$ is minimal. We have

$$\mathbf{U}(\mathbf{x}) = (T(\mathbf{x}), x_1) = (x_1 + x_2, x_1)$$

and

$$\mathbf{U}(\mathbf{y}) = (T(\mathbf{y}), y_1) = (y_1 + y_2, y_1).$$

As shown above, the ratio $p_{\mathbf{X}}(\mathbf{x}|p)/p_{\mathbf{X}}(\mathbf{y}|p)$ will be constant if and only if $x_1 + y_1 = x_2 + y_2$, which does not imply $x_1 = y_1$. Thus, it is not the case that the ratio will be constant if and only if $\mathbf{U}(\mathbf{x}) = \mathbf{U}(\mathbf{y})$, so it follows that $\mathbf{U}$ is not a minimal sufficient statistic for p. Another way to see that $\mathbf{U}$ is not minimal is to observe that

$$\begin{aligned} B_{(0,0)} &= \{(0,0)\} = A_0 \\ B_{(1,0)} &= \{(1,0)\} \subset A_1 \\ B_{(1,1)} &= \{(0,1)\} \subset A_1 \\ B_{(2,1)} &= \{(1,1)\} = A_2. \end{aligned}$$

Thus, $B_t \subseteq A_t$, i.e., the partition of $\mathcal{X}$ induced by $\mathbf{U}$ is not equivalent to that induced by T, which is minimal sufficient, so by corollary 8.19, $\mathbf{U}$ is not a minimal sufficient statistic for p. Conversely, note that the statistic $W = 17T$ generates the same partition as T, so W is also minimal sufficient.

EXAMPLE 8.21 (Exponential minimal sufficient statistic). Let $X \sim \text{Exp}(\lambda)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pdf of X, which is given by

$$f(x|\lambda) = \lambda \mathrm{e}^{-\lambda x}$$

where $x > 0$ and $f(x|\lambda) = 0$ otherwise, so that a joint pdf of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\lambda) = \prod_{i=1}^{n} \lambda \mathrm{e}^{-\lambda x} I_{\{x_i > 0\}} = \lambda^n \mathrm{e}^{-\sum_{i=1}^n \lambda x_i} I_{\{x_{(1)} > 0\}}.$$

Find a minimal sufficient statistic for λ.

We have

$$\begin{aligned} \frac{f_{\mathbf{X}}(\mathbf{x}|\lambda)}{f_{\mathbf{X}}(\mathbf{y}|\lambda)} &= \frac{\lambda^n \mathrm{e}^{-\lambda \sum_{i=1}^n x_i} I_{\{x_{(1)} > 0\}}}{\lambda^n \mathrm{e}^{-\lambda \sum_{i=1}^n y_i} I_{\{y_{(1)} > 0\}}} \\ &= \mathrm{e}^{-\lambda \sum_{i=1}^n x_i} \mathrm{e}^{\lambda \sum_{i=1}^n y_i} \left(\frac{I_{\{x_{(1)} > 0\}}}{I_{\{y_{(1)} > 0\}}} \right) \\ &= \mathrm{e}^{-\lambda \left(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i \right)} \left(\frac{I_{\{x_{(1)} > 0\}}}{I_{\{y_{(1)} > 0\}}} \right). \end{aligned}$$

This ratio will be constant as a function of λ if and only if $\sum_{i=1}^n x_i - \sum_{i=1}^n y_i = 0$, i.e., if $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. It follows from theorem 8.18 that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is a minimal sufficient statistic for λ.

EXAMPLE 8.22 (Minimal sufficient statistics for an exponential family). The set of sufficient statistics $T(X) = \{T_1(X), \ldots, T_n(X)\}$ from an exponential family are minimal sufficient. Recall that the pmf or pdf of an exponential family can be written as

$$f(x|\theta) = h^*(x) c^*(\theta) \exp\left\{ \sum_{j=1}^{k} \omega_j(\theta) T_j(x) \right\}.$$

We have

$$\begin{aligned}\frac{f(x|\theta)}{f(y|\theta)} &= \frac{h^*(x)\,c^*(\theta)\exp\left\{\sum_{j=1}^{k}\omega_j(\theta)\,T_j(x)\right\}}{h^*(y)\,c^*(\theta)\exp\left\{\sum_{j=1}^{k}\omega_j(\theta)\,T_j(y)\right\}}\\ &= \frac{h^*(x)}{h^*(y)}\exp\left\{\sum_{j=1}^{k}\omega_j(\theta)\,T_j(x)\right\}\exp\left\{-\sum_{j=1}^{k}\omega_j(\theta)\,T_j(y)\right\}\\ &= \frac{h^*(x)}{h^*(y)}\exp\left\{\left[\sum_{j=1}^{k}\omega_j(\theta)\,T_j(x)\right]-\left[\sum_{j=1}^{k}\omega_j(\theta)\,T_j(y)\right]\right\}\\ &= \frac{h^*(x)}{h^*(y)}\exp\left\{[\omega_1(\theta)\,T_1(x)+\cdots+\omega_k T_k(x)]-[\omega_1(\theta)\,T_1(y)+\cdots+\omega_k(\theta)\,T_k(y)]\right\}\\ &= \frac{h^*(x)}{h^*(y)}\exp\left\{\omega_1(\theta)\,[T_1(x)-T_1(y)]+\cdots+\omega_k\,[T_k(x)-T_k(y)]\right\}\\ &= \frac{h^*(x)}{h^*(y)}\exp\left\{\sum_{j=1}^{k}\omega_j(\theta)\,[T_j(x)-T_j(y)]\right\}.\end{aligned}$$

This ratio will be constant as a function of θ if and only if $T(X)=T(Y)$. It follows from theorem 8.18 that $T(X)=\{T_1(X),\ldots,T_n(X)\}$ is a minimal sufficient statistic for θ.

8.1.5. Ancillary statistics.

DEFINITION 8.23. A statistic $S(\mathbf{X})$ whose distribution does not depend on the parameter θ is called an *ancillary statistic*.

Alone, an ancillary statistic contains no information about θ. When used in conjunction with other statistics, it sometimes contains valuable information for inference about θ.

8.1.5.1. *Bivariate transformations.* Let (X,Y) be a random vector with a known probability distribution. Now consider a new bivariate random vector (U,V) defined by $U=g_1(X,Y)$ and $V=g_2(X,Y)$, where $g_1(x,y)$ and $g_2(x,y)$ are some specified functions. If B is any subset of $\mathbb{R}^2$, then $(U,V)\in B$ if and only if $(X,Y)\in A$, where $A=\{(x,y):(g_1(x,y),g_2(x,y))\in B\}$. Thus $P(\{(U,V)\in B\})=P(\{(X,Y)\in A\})$, and the probability distribution of (U,V) is completely determined by the probability distribution of (X,Y).

If (X,Y) is a continuous random vector with joint pdf $f_{X,Y}(x,y)$, then the joint pdf of (U,V) can be expressed in terms of $f_{X,Y}(x,y)$. Let

$$\mathcal{A}=\{(x,y):f_{X,Y}(x,y)>0\}$$

and

$$\mathcal{B}=\{(u,v):u=g_1(x,y),v=g_2(x,y),(x,y)\in\mathcal{A}\}.$$

The joint pdf $f_{U,V}(u,v)$ will be positive on the set $\mathcal{B}$. For the simplest version of this result we assume that the transformation $u=g_1(x,y)$ and $v=g_2(x,y)$ defines a one-to-one transformation of $\mathcal{A}$ onto $\mathcal{B}$. We are assuming that for each $(u,v)\in\mathcal{B}$, there is only one $(x,y)\in\mathcal{A}$ such that $(u,v)=(g_1(x,y),g_2(x,y))$. For such a one-to-one, onto transformation, we can solve the equations $u=g_1(x,y)$ and $v=g_2(x,y)$ for x and y in terms of u and v. We will denote this inverse transformation by $x=h_1(u,v)$ and $y=h_2(u,v)$. The role played by the derivative in the univariate case is now played by quantity called the Jacobian of the transformation. This function of (u,v), denoted by J, is the determinant of a matrix of partial derivatives, and is defined by

$$J=\begin{vmatrix}\dfrac{\partial x}{\partial u} & \dfrac{\partial x}{\partial v}\\ \dfrac{\partial y}{\partial u} & \dfrac{\partial y}{\partial v}\end{vmatrix}=\frac{\partial x}{\partial u}\frac{\partial y}{\partial v}-\frac{\partial y}{\partial u}\frac{\partial x}{\partial v},$$

where

$$\frac{\partial x}{\partial u} = \frac{\partial h_1(u,v)}{\partial u}, \quad \frac{\partial x}{\partial v} = \frac{\partial h_1(u,v)}{\partial v}, \quad \frac{\partial y}{\partial u} = \frac{\partial h_2(u,v)}{\partial u}, \quad \text{and} \quad \frac{\partial y}{\partial v} = \frac{\partial h_2(u,v)}{\partial v}.$$

Then, the joint pdf of (U,V) is 0 outside the set $\mathcal{B}$ and on the set $\mathcal{B}$ is given by

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v))|J|.$$

REMARK. This technique, which is easily generalized to more than two variables, can be used to find the pdf of some function of interest by first finding the joint pdf of that function and another, conveniently chosen function, then integrating the resulting joint pdf with respect to the second function, which gives the (marginal) pdf of the function of interest. This technique is demonstrated in example 8.25.

THEOREM 8.24 (Distribution of the sum of Poisson variables). *If $X \sim Poisson(\theta)$ and $Y \sim Poisson(\lambda)$ and X and Y are independent, then $X + Y \sim Poisson(\theta + \lambda)$. (This is Theorem 4.3.2 from Casella & Berger.)*

PROOF. [proof goes here] □

EXAMPLE 8.25 (Uniform ancillary statistic). Let $X_1, \ldots, X_n$ be iid uniform observations on the interval $(\theta, \theta+1)$, $-\infty < \theta < \infty$. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics from the sample. Show that the range statistic $R = X_{(n)} - X_{(1)}$ is an ancillary statistic. (This is example 6.2.17 from Casella & Berger.)
The pdf of each X_i is given by

$$\begin{aligned} f_{X_i}(x|\theta) &= \frac{1}{(\theta+1)-\theta} \\ &= 1, \end{aligned}$$

where $\theta < x < \theta + 1$ and $f_{X_i}(x|\theta) = 0$ otherwise, so that the cdf of each X_i is given by

$$F_{X_i}(x|\theta) = \begin{cases} 0, & x \le \theta \\ \int_\theta^x 1 \mathrm{d}x = x - \theta, & \theta < x < \theta + 1 \\ 1, & x \ge \theta + 1 \end{cases}.$$

Let $u = x_{(1)}$ and let $v = x_{(n)}$. By theorem 5.7, the joint pdf of $X_{(1)}$ and $X_{(n)}$ is given by

$$\begin{aligned} f_{X_{(1)},X_{(n)}}(u,v) &= \frac{n!}{(1-1)!\,(n-1-1)!\,(n-n)!} f_X(u) f_X(v) [F_X(u)]^{1-1} \\ &\quad \cdot [F_X(v) - F_X(u)]^{n-1-1} [1 - F_X(v)]^{n-n} \\ &= \frac{n!}{(n-2)!} f_X(u) f_X(v) [F_X(v) - F_X(u)]^{n-2} \\ &= \frac{n!}{(n-2)!} \cdot 1 \cdot 1 \cdot [(v-\theta) - (u-\theta)]^{n-2} \\ &= \frac{n!}{(n-2)!} (v-u)^{n-2} \\ &= \frac{n(n-1)(n-2)!}{(n-2)!} (v-u)^{n-2} \\ &= n(n-1)(v-u)^{n-2} \end{aligned}$$

where $\theta < u < v < \theta + 1$ and $f_{X_{(1)},X_{(n)}}(u,v) = 0$ otherwise. Let

$$\mathcal{A} = \left\{(u,v) : f_{X_{(1)},X_{(n)}}(u,v) > 0\right\},$$

and make the transformation $R = X_{(n)} - X_{(1)}$ and $M = \left(X_{(1)} + X_{(n)}\right)/2$, so that we have

$$\mathcal{B} = \left\{(r,m) : r = v - u, m = (u+v)/2, (u,v) \in \mathcal{A}\right\}.$$

Then, $f_{R,M}(r,m)$ will be positive on $\mathcal{B}$. We have the inverse transformation

$$\begin{aligned} X_{(1)} &= 2M - X_{(n)} \\ &= 2M - \left(R + X_{(1)}\right) \\ \Leftrightarrow 2X_{(1)} &= 2M - R \end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow X_{(1)} = (2M - R)/2 \\
&\Leftrightarrow u = (2m - r)/2
\end{aligned}$$

and

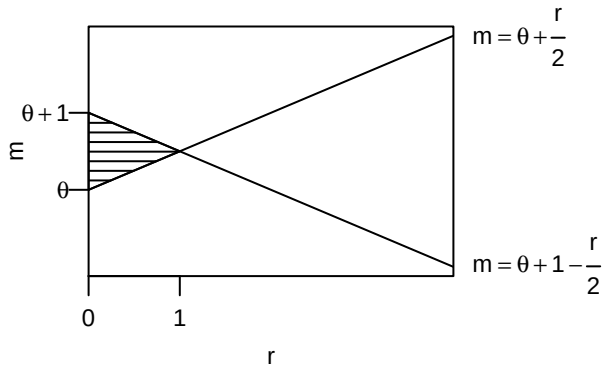
$$\begin{aligned}
X_{(n)} &= 2M - X_{(1)} \\
&= 2M - \left(X_{(n)} - R\right) \\
\Leftrightarrow 2X_{(n)} &= 2M + R \\
\Leftrightarrow X_{(n)} &= (2M + R)/2 \\
&\Leftrightarrow v = (2m + r)/2.
\end{aligned}$$

The Jacobian of the transformation is given by

$$\begin{aligned}
J &= \begin{vmatrix} \frac{\partial u}{\partial r} & \frac{\partial u}{\partial m} \\ \frac{\partial v}{\partial r} & \frac{\partial v}{\partial m} \end{vmatrix} \\
&= \begin{vmatrix} -\frac{1}{2} & 1 \\ \frac{1}{2} & 1 \end{vmatrix} \\
&= \left(-\frac{1}{2}\right)(1) - \left(\frac{1}{2}\right)(1) \\
&= -1.
\end{aligned}$$

We will now find the range for $f_{R,M}$. We have

$$\begin{aligned}
u > \theta &\Rightarrow m - \frac{r}{2} > \theta \\
&\Rightarrow m > \theta + \frac{r}{2} \\
v > u &\Rightarrow m + \frac{r}{2} > m - \frac{r}{2} \\
&\Rightarrow \frac{r}{2} > -\frac{r}{2} \\
&\Rightarrow r > 0 \\
v < \theta + 1 &\Rightarrow \frac{2m + r}{2} < \theta + 1 \\
&\Rightarrow m < \theta + 1 - \frac{r}{2}
\end{aligned}$$



From the figure above, we see that the region where $f_{R,M}$ is positive is bounded by $r > 0$, $m > \theta + (r/2)$, and $m < \theta + 1 - (r/2)$. The upper limit of integration for r is given by the intersection of $m = \theta + (r/2)$ and $m = \theta + 1 - (r/2)$. Setting them equal, we have

$$\theta + \frac{r}{2} = \theta + 1 - \frac{r}{2} \implies r = 1.$$

Then, the joint pdf of R and M is given by

$$\begin{aligned} f_{R,M}(r,m) &= f_{X_{(1)},X_{(n)}}\left(\frac{2m-r}{2},\frac{2m+r}{2}\right)|J| \\ &= n(n-1)\left(\frac{2m+r}{2}-\frac{2m-r}{2}\right)^{n-2}|-1| \\ &= n(n-1)\left(\frac{2m+r-2m+r}{2}\right)^{n-2} \\ &= n(n-1)\left(\frac{2r}{2}\right)^{n-2} \\ &= n(n-1)r^{n-2} \end{aligned}$$

where $0<r<1$ and $\theta+(r/2)<m<\theta+1-(r/2)$ and $f_{R,M}(r,m)=0$ otherwise. We will find the (marginal) pdf of R by integrating the joint pdf with respect to m.

$$\begin{aligned} f_R(r|\theta) &= \int_{\theta+(r/2)}^{\theta+1-(r/2)} n(n-1)r^{n-2}\mathrm{d}m \\ &= n(n-1)r^{n-2}\left[m\Big|_{\theta+(r/2)}^{\theta+1-(r/2)}\right] \\ &= n(n-1)r^{n-2}\left[\theta+1-\frac{r}{2}-\left(\theta+\frac{r}{2}\right)\right] \\ &= n(n-1)r^{n-2}\left[1-\frac{r}{2}-\frac{r}{2}\right] \\ &= n(n-1)r^{n-2}(1-r) \end{aligned}$$

where $0<r<1$. This expression is independent of θ, so it follows that R is ancillary.

EXAMPLE 8.26 (Ancillary statistic for location family). Let $X_1,\ldots,X_n$ be iid observations from a location parameter family with cdf $F(x-\theta)$, $-\infty<\theta<\infty$. We will show that the range, $R=X_{(n)}-X_{(1)}$, is an ancillary statistic. (This is example 6.2.18 from Casella & Berger.)

We will use theorem 7.19 and work with $Z_1,\ldots,Z_n$ iid observations from $F(x)$ (corresponding to $\theta=0$) with $X_1=Z_1+\theta,\ldots,X_n=Z_n+\theta$. Thus, the cdf of the range statistic, R, is

$$\begin{aligned} F_R(r|\theta) &= P_\theta(\{R\le r\}) \\ &= P_\theta\left(\left\{\max_i X_i-\min_i X_i\le r\right\}\right) \\ &= P_\theta(\{\max(X_1,\ldots,X_n)-\min(X_1,\ldots,X_n)\le r\}) \\ &= P_\theta(\{\max(Z_1+\theta,\ldots,Z_n+\theta)-\min(Z_1+\theta,\ldots,Z_n+\theta)\le r\}) \\ &= P_\theta\left(\left\{Z_{(n)}+\theta-\left(Z_{(1)}+\theta\right)\right\}\le r\right) \\ &= P_\theta\left(\left\{Z_{(n)}-Z_{(1)}\le r\right\}\right). \end{aligned}$$

This expression does not depend on θ, so it follows that the cdf of R does not depend on θ and therefore R is an ancillary statistic.

EXAMPLE 8.27 (Ancillary statistic for scale family). Let $X_1,\ldots,X_n$ be iid observations from a scale parameter family with cdf $F(x/\sigma)$, $\sigma>0$. Then, any statistic that depends on the sample only through the $n-1$ values $X_1/X_n,\ldots,X_{n-1}/X_n$ is an ancillary statistic. (This is example 6.2.19 from Casella & Berger.)

Let $Z_1,\ldots,Z_n$ be iid observations from $F(x)$ (corresponding to $\sigma=1$) with $X_i=\sigma Z_i$. The joint cdf of $X_1/X_n,\ldots,X_{n-1}/X_n$ is

$$\begin{aligned} F(y_1,\ldots,y_{n-1}|\sigma) &= P_\sigma\left(\left\{\frac{X_1}{X_n}\le y_1\right\}\cap\cdots\cap\left\{\frac{X_{n-1}}{X_n}\le y_{n-1}\right\}\right) \\ &= P_\sigma\left(\left\{\frac{\sigma Z_1}{\sigma Z_n}\le y_1\right\}\cap\cdots\cap\left\{\frac{\sigma Z_{n-1}}{\sigma Z_n}\le y_{n-1}\right\}\right) \end{aligned}$$

$$= P_\sigma\left(\left\{\frac{Z_1}{Z_n} \le y_1\right\} \cap \cdots \cap \left\{\frac{Z_{n-1}}{Z_n} \le y_{n-1}\right\}\right).$$

The last probability does not depend on σ because the distribution of $Z_1, \ldots, Z_n$ does not depend on σ. So the distribution of $X_1/X_n, \ldots, X_{n-1}/X_n$ is independent of σ.

EXAMPLE 8.28 (Mixture of normal distributions). Sometimes, an ancillary statistic is viewed in conjunction with another statistic, and together they form a minimal sufficient statistic, i.e., $S = (T, C)$ where C is ancillary for θ and T is minimal sufficient conditional on C.

Suppose that Y is a mixture of normal distributions $\mathcal{N}\left(\mu, \sigma_0^2\right)$ and $\mathcal{N}\left(\mu, \sigma_1^2\right)$ with σ_0^2 and σ_1^2 known (for example, any two of the three distributions in 7.2.2). Let C be defined as

$$C = \begin{cases} 0, & \text{if } Y \sim \mathcal{N}\left(\mu, \sigma_0^2\right) \\ 1, & \text{if } Y \sim \mathcal{N}\left(\mu, \sigma_1^2\right) \end{cases}.$$

Y is equally likely to be $\mathcal{N}\left(\mu, \sigma_0^2\right)$ or $\mathcal{N}\left(\mu, \sigma_1^2\right)$, so $P\left(\left\{Y \sim \mathcal{N}\left(\mu, \sigma_0^2\right)\right\}\right) = P\left(\left\{Y \sim \mathcal{N}\left(\mu, \sigma_1^2\right)\right\}\right) = 1/2$. Then, the joint pdf of Y and C is given by

$$\begin{aligned} f_{C,Y}(c, y) &= P(\{C = c\} \cap \{Y = y\}) \\ &= P(\{Y = y\} \mid \{C = c\}) \cdot P(\{C = c\}) \\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma_c^2}} e^{-(y-\mu)^2/\left(2\sigma_c^2\right)} \\ &= \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left\{-\frac{y^2}{2\sigma_c^2} - \frac{\mu^2}{2\sigma_c^2} + \frac{\mu y}{\sigma_c^2}\right\} \\ &= \underbrace{\frac{1}{2}\frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left\{-\frac{y^2}{2\sigma_c^2}\right\}}_{h(y)} \exp\left\{\underbrace{-\frac{\mu^2}{2}}_{\omega_1(\mu)}\underbrace{\frac{1}{\sigma_c^2}}_{t_1(y)} + \underbrace{\mu}_{\omega_2(\mu)}\underbrace{\frac{y}{\sigma_c^2}}_{t_2(y)}\right\}. \end{aligned}$$

This last expression has the form of an exponential family, with $c(\mu) = 1$. In example 8.22, we showed that a minimal sufficient statistic for an exponential family is given by $T(Y) = (T_1(Y), \ldots, T_n(Y))$. It follows that

$$T(Y) = \left(\frac{1}{\sigma_c^2}, \frac{Y}{\sigma_c^2}\right)$$

is a minimal sufficient statistic for μ conditioned on C.

8.1.5.2. *Expanded definition of ancillarity.* Suppose that $\theta = (\psi, \lambda)$, where λ is not of direct interest (λ is a nuisance parameter). Suppose that $S = (T, C)$ is minimal sufficient for θ, where the pdf of C depends on λ but not on ψ, and the conditional pdf of T given C depends on ψ but not λ. Then C is called ancillary in the extended sense.

8.1.6. Complete statistics.

DEFINITION 8.29. Let $f(t|\theta)$ be a family of pdfs or pmfs for a statistic $T(\mathbf{X})$. The family of probability distributions is called *complete* if $\mathrm{E}_\theta[g(T)] = 0$ for all θ implies $P_\theta(\{g(T) = 0\}) = 1$ for all θ. Equivalently, $T(\mathbf{X})$ is called a *complete statistic*. I.e., T is complete if

$$\mathrm{E}_\theta[g(T)] = 0 \quad \forall\theta \qquad \Longrightarrow \qquad P_\theta(\{g(T) = 0\}) = 1 \quad \forall g.$$

THEOREM 8.30. *If $T(X)$ is sufficient and complete for θ, then T is minimal sufficient.*

PROOF. [proof goes here] □

THEOREM 8.31 (Complete statistics in the exponential family). *Let $X_1, \ldots, X_n$ be iid observations from an exponential family with pdf or pmf of the form*

$$f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{k} \omega_j(\boldsymbol{\theta})\, t_j(x)\right),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k)$. *Then the statistic*

$$T(\mathbf{X}) = \left(\sum_{i=1}^{n} t_1(X_i), \sum_{i=1}^{n} t_2(X_i), \ldots, \sum_{i=1}^{n} t_k(X_i) \right)$$

is complete if $\{(\omega_1(\boldsymbol{\theta}), \ldots, \omega_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \Theta\}$ *contains an open set in* $\mathbb{R}^k$. *(This is Theorem 6.2.25 from Casella & Berger.)*

PROOF. [proof goes here] □

THEOREM 8.32. *If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic. (This is Theorem 6.2.28 from Casella & Berger.)*

PROOF. [proof goes here] □

REMARK 8.33. Minimal sufficiency does not imply completeness. In example 7.7, we expressed the family of distributions with densities

$$f(x|\theta) = \frac{2}{\Gamma(1/4)} \exp\left[-(x-\theta)^4 \right]$$

for $x \in \mathbb{R}$ as

$$f(x|\theta) = \underbrace{\frac{2}{\Gamma(1/4)} \exp\left\{-x^4\right\}}_{h(x)} \underbrace{\exp\left\{-\theta^4\right\}}_{c(\theta)} \exp\left\{ \underbrace{4x^3}_{t_1(x)} \underbrace{\theta}_{\omega_1(\theta)} \underbrace{-6x^2}_{t_2(x)} \underbrace{\theta^2}_{\omega_2(\theta)} + \underbrace{4x}_{t_3(x)} \underbrace{\theta^3}_{\omega_3(\theta)} \right\},$$

which is in exponential form, so that $T(\mathbf{X}) = \left(4x^3, -6x^2, 4x\right)$. It follows from example 8.22 that $T(\mathbf{X})$ is a minimal sufficient statistic for θ. The parameter space for this distribution is given by $\{\omega_1(\theta), \omega_2(\theta), \omega_3(\theta)\} = \left\{\theta, \theta^2, \theta^3\right\}$, i.e., $k = 3$. We have $d = 1$, so by definition this familiy of densities is a curved exponential family. Its graph in $\mathbb{R}^3$ is the curve shown in figure 8.1.1. This graph does not have positive length (volume in $\mathbb{R}^3$), i.e., $\left\{\theta, \theta^2, \theta^3\right\}$ does not contain an open set in $\mathbb{R}^3$, so it follows from theorem 8.31 that $T(\mathbf{X})$ is not complete.

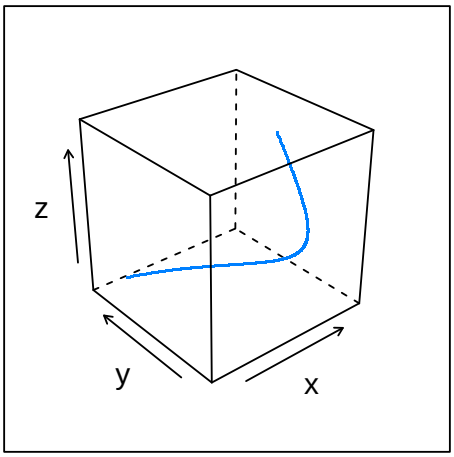


FIGURE 8.1.1. graph of $x = \theta, y = \theta^2, z = \theta^3$

REMARK 8.34. Although the condition in theorem 8.31 that the set $\{(\omega_1(\boldsymbol{\theta}), \ldots, \omega_k(\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta)\}$ contain an open set is sufficient to guarantee completeness, it is not necessary. That is, the failure of this condition does not show that a statistic is not complete, as shown in the following example.

EXAMPLE 8.35. Suppose $X \sim \text{Bernoulli}(p)$ and the parameter space consists of the two points $p = 1/4$ and $p = 3/4$. The pmf for X can be written as

$$\begin{aligned} p_X(x) &= p^x (1-p)^{1-x} \\ &= p^x (1-p)^{-x} (1-p) \\ &= (1-p) \left(\frac{p}{1-p} \right)^x \\ &= (1-p) \exp \left\{ \log \left(\left(\frac{p}{1-p} \right)^x \right) \right\} \\ &= \underbrace{(1-p)}_{c(p)} \exp \left\{ \underbrace{x}_{t_1(x)} \underbrace{\log \left(\frac{p}{1-p} \right)}_{\omega_1(p)} \right\} \end{aligned}$$

where $x \in \{0,1\}$ and $p_X(x) = 0$ otherwise, i.e., X has an exponential family distribution. Then, the parameter space is given by

$$\begin{aligned} \left\{ \omega_1(p) : p = \left\{ \frac{1}{4}, \frac{3}{4} \right\} \right\} &= \left\{ \log \left(\frac{1/4}{1-(1/4)} \right), \log \left(\frac{3/4}{1-(3-4)} \right) \right\} \\ &= \left\{ \log \left(\frac{1/4}{3/4} \right), \log \left(\frac{3/4}{1/4} \right) \right\} \\ &= \left\{ \log \frac{1}{3}, \log 3 \right\} \end{aligned}$$

The condition of theorem 8.31 is not satisfied because the parameter space has only two points in it, and hence the range of $\log(p/(1-p))$ as p varies over the parameter space does not contain an open set (an interval) in $\mathbb{R}^1$.

Yet, the distribution of X is complete. To see this, suppose g is a function defined on the sample space such that $\mathrm{E}[g(X)] = 0$. To show that the distribution of X is complete, we need to show that the only function satisfying this equality for all p in the parameter space is the function which is identically zero. The sample space consists of the two points $X = 0$ and $X = 1$, so we need to show that this implies $g(0) = g(1) = 0$.

The expected value of $g(X)$ is given by

$$\begin{aligned} \mathrm{E}[g(X)] &= \sum_{x=0}^{1} g(x) \cdot p_X(x) \\ &= \sum_{x=0}^{1} g(x) \cdot p^x (1-p)^{1-x} \\ &= g(0) \cdot p^0 (1-p)^{1-0} + g(1) \cdot p^1 (1-p)^{1-1} \\ &= (1-p) g(0) + p g(1) \end{aligned}$$

so that if $p = 1/4$, we have

$$0 = \mathrm{E}_{p=1/4}[g(X)] = \frac{3}{4} g(0) + \frac{1}{4} g(1)$$

and if $p = 3/4$, we have

$$0 = \mathrm{E}_{p=3/4}[g(X)] = \frac{1}{4} g(0) + \frac{3}{4} g(1).$$

So, we have $g(1) = -3g(0)$, which gives

$$0 = \frac{1}{4} g(0) + \frac{3}{4} g(1) = \frac{1}{4} g(0) + \frac{3}{4} (-3g(0)) = -2g(0) \implies 0 = g(0)$$

and thus $g(1) = -3(0) = 0$. Thus, the only solution is $g(1) = g(0) = 0$, and with just two points in the parameter space we have that the family of distributions is complete.

EXAMPLE 8.36. Let $X \sim \text{Bernoulli}(p)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pmf of X, which is given by

$$p_X(x) = p^x (1-p)^{1-x}$$

where $x \in \{0, 1\}$ and $p_X(x) = 0$ otherwise. Show that $T_1 = X_2 - X_1$ is not a complete statistic.

By definition, T_1 will be a complete statistic if $\mathrm{E}[g(T_1)] = 0$ for all p implies $P(\{g(T_1) = 0\}) = 1$ for all p. Suppose that $g(T_1) = T_1$. Then, we have

$$\begin{aligned}
\mathrm{E}[g(T_1)] &= \mathrm{E}[T_1] \\
&= \mathrm{E}[X_2 - X_1] \\
&= \mathrm{E}[X_2] - \mathrm{E}[X_1] \\
&= \sum_{x_2=0}^{1} \left[x_2 \cdot p^{x_2} (1-p)^{1-x_2} \right] - \sum_{x_1=0}^{1} \left[x_1 \cdot p^{x_1} (1-p)^{1-x_1} \right] \\
&= \left[0 \cdot p^0 (1-p)^{1-0} + 1 \cdot p^1 (1-p)^{1-1} \right] - \left[0 \cdot p^0 (1-p)^{1-0} + 1 \cdot p^1 (1-p)^{1-1} \right] \\
&= [0 + p] - [0 + p] \\
&= 0.
\end{aligned}$$

So, this choice of g satisfies $\mathrm{E}[g(T_1)] = 0$ for all p. Then, we have

$$\begin{aligned}
P(\{g(T_1) = 0\}) &= P(\{T_1 = 0\}) \\
&= P(\{X_2 - X_1 = 0\}).
\end{aligned}$$

This probability will be equal to 1 if and only if $X_1 = X_2$, but this will not always be the case, e.g., suppose that $X_1 = 0$ and $X_2 = 1$. Even though our choice of g satisfies $\mathrm{E}[g(T_1)] = 0$ for all p, it does not imply $P(\{g(T_1) = 0\}) = 1$ for all p. The condition of completeness applies to all functions (all choices of g), so it follows that T_1 is not complete.

EXAMPLE 8.37 (Bernoulli complete statistic). Let $X_1, \ldots, X_n$ be as in example 8.36. Show that $T_2 = \sum_{i=1}^{n} X_i$ is a complete statistic. (This is example 6.2.22 from Casella & Berger.)

T_2 is the sum of n independent Bernoulli random variables each having the same success probability p. It follows that $T_2 \sim \text{Binomial}(n, p)$, so that the pmf of T_2 is given by

$$p_{T_2}(t) = \binom{n}{t} p^t (1-p)^{n-t}$$

where $t \in \{0, 1, 2, \ldots, n\}$ and $p_{T_2}(t) = 0$ otherwise. Suppose $g(T_2)$ is a function satisfying $\mathrm{E}[g(T_2)] = 0$, so that we have

$$\begin{aligned}
0 &= \mathrm{E}[g(T_2)] \\
&= \sum_{k=0}^{n} [g(k) \cdot p_{T_2}(k)] \\
&= \sum_{k=0}^{n} \left[g(k) \cdot \binom{n}{k} p^k (1-p)^{n-k} \right] \\
&= (1-p)^n \sum_{k=0}^{n} \left[g(k) \cdot \binom{n}{k} p^k (1-p)^{-k} \right] \\
&= (1-p)^n \sum_{k=0}^{n} \left[g(k) \cdot \binom{n}{k} \left(\frac{p}{1-p} \right)^k \right].
\end{aligned}$$

We have $p \in (0, 1)$, so that $(1-p)^n > 0$ for all p. It follows that this expression will be equal to zero if and only if

$$\sum_{k=0}^{n} \left[g(k) \cdot \binom{n}{k} \left(\frac{p}{1-p} \right)^k \right] = 0.$$

Let $r = (p/(1-p))$. Then, this sum is a polynomial function of r, i.e.,

$$\sum_{k=0}^{n}\left[g(k)\cdot\binom{n}{k}r^k\right] = g(0)\cdot\binom{n}{0}r^0+\cdots+g(n)\cdot\binom{n}{n}r^n$$
$$= g(0)+g(1)\cdot nr^1+g(2)\binom{n}{2}r^2+\cdots+g(n-1)\cdot nr^{n-1}+g(n)\,r^n.$$

We have $n>0$ and $r=p/(1-p)>0$, so that for the kth term, we will have $\binom{n}{k}>0$ and $r^k>0$. It follows that this sum is equal to zero if and only if $g(k)=0$ for $k=\{0,1,\ldots,n\}$. T_2 takes on the values $0,1,\ldots,n$ with probability 1 (recall that T_2 represents the probability of k successes in n trials), so that $P(\{g(T_2)=0\})=1$ for all p. It follows that T_2 is a complete statistic.

EXAMPLE 8.38 (Poisson complete statistic). Let $X\sim\text{Poisson}(\lambda)$, and let $\mathbf{X}=X_1,\ldots,X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pmf of X, which is given by

$$p_X(x)=\frac{\mathrm{e}^{-\lambda}\lambda^x}{x!}$$

where $x\in\{0,1,2,\ldots\}$ and $p_X(x)=0$ otherwise. Show that $T(\mathbf{X})=\sum_{i=1}^n X_i$ is a complete statistic.
It follows from theorem 8.24 that $X_1+X_2\sim\text{Poisson}(\lambda+\lambda)=\text{Poisson}(2\lambda)$, that $(X_1+X_2)+X_3\sim\text{Poisson}(2\lambda+\lambda)=\text{Poisson}(3\lambda)$, and therefore that

$$T(\mathbf{X})=\sum_{i=1}^{n}X_i\sim\text{Poisson}(n\lambda),$$

so that the pmf of $T(\mathbf{X})$ is given by

$$p_{T(\mathbf{X})}(t)=\frac{\mathrm{e}^{-n\lambda}(n\lambda)^t}{t!}$$

where $t\in\{0,1,2,\ldots\}$ and $p_{T(\mathbf{X})}(t)=0$ otherwise. Suppose that $g(T)$ is a function satisfying $\mathrm{E}[g(T)]=0$, so that we have

$$0=\mathrm{E}[g(T)]=\sum_{t=0}^{\infty}\left[g(t)\cdot\frac{\mathrm{e}^{-n\lambda}(n\lambda)^t}{t!}\right]=\mathrm{e}^{-n\lambda}\sum_{t=0}^{\infty}g(t)\cdot\frac{(n\lambda)^t}{t!}.$$

We will have $\mathrm{e}^{-n\lambda}>0$ for all n and all $\lambda\geq 0$. It follows that this expression will be equal to zero if and only if

$$\sum_{t=0}^{\infty}g(t)\cdot\frac{(n\lambda)^t}{t!}=0.$$

This sum is a polynomial function of $(n\lambda)$. For the sum to be equal to zero, the coefficient $g(t)/t!$ must be equal to zero for all t. Because $t!\geq 1$, it follows that $g(t)=0$ for all t. Thus, we have

$$\mathrm{E}[g(T)]=0\quad\forall\lambda\implies P(\{g(T)=0\})=1\quad\forall\lambda,$$

so by definition $T(\mathbf{X})$ is a complete statistic.

EXAMPLE 8.39 (Uniform complete statistic). Let $X\sim\mathcal{U}(0,\theta)$, and let $\mathbf{X}=X_1,\ldots,X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pdf of X, which is given by

$$f_X(x)=\frac{1}{\theta}I_{\{0<x<\theta\}}.$$

Show that $T(\mathbf{X})=X_{(n)}$ is a complete statistic. (This is example 6.2.23 from Casella & Berger.)
From example 8.8, the cdf of X is given by $F_X(x)=x/\theta$. From theorem 5.6, the pdf of $X_{(n)}$ is given by

$$f_{X_{(n)}}(x)=\frac{n!}{(n-1)!\,(n-n)!}f_X(x)\left[F_X(x)\right]^{n-1}\left[1-F_X(x)\right]^{n-n}$$
$$=\frac{n!}{(n-1)!0!}\left(\frac{1}{\theta}\right)\left(\frac{x}{\theta}\right)^{n-1}\left(1-\frac{x}{\theta}\right)^0 I_{\{0<x<\theta\}}$$
$$=n\left(\frac{(n-1)!}{(n-1)!}\right)\left(\frac{1}{\theta}\right)\left(\frac{x^{n-1}}{\theta^{n-1}}\right)I_{\{0<x<\theta\}}$$

$$= \frac{nx^{n-1}}{\theta^n} I_{\{0<x<\theta\}}.$$

Suppose $g(T)$ is a function satisfying $\mathrm{E}[g(T)] = 0$ for all θ, so that we have

$$0 = \mathrm{E}[g(T)] = \int_0^\theta g(t) \cdot f_{X_{(n)}}(t)\,\mathrm{d}\,t = \int_0^\theta g(t) \cdot nt^{n-1}\theta^{-n}\,\mathrm{d}\,t.$$

By assumption, $\mathrm{E}[g(T)] = 0$, i.e., it is constant as a function of θ, so that its derivative with respect to θ is zero. Then, we have

$$\begin{aligned}
0 &= \frac{\mathrm{d}}{\mathrm{d}\,\theta}\left[\int_0^\theta g(t) \cdot nt^{n-1}\theta^{-n}\,\mathrm{d}\,t\right] \\
&= \frac{\mathrm{d}}{\mathrm{d}\,\theta}\left[n\theta^{-n}\int_0^\theta g(t) \cdot t^{n-1}\,\mathrm{d}\,t\right] \\
\text{(product rule)} \quad &= n\theta^{-n}\frac{\mathrm{d}}{\mathrm{d}\,\theta}\left[\int_0^\theta g(t) \cdot t^{n-1}\,\mathrm{d}\,t\right] + \left(\frac{\mathrm{d}}{\mathrm{d}\,\theta}n\theta^{-n}\right)\int_0^\theta g(t) \cdot t^{n-1}\,\mathrm{d}\,t \\
\text{(Fundamental Theorem of Calculus)} \quad &= n\theta^{-n}\left[g(\theta) \cdot \theta^{n-1} - g(0) \cdot 0^{n-1}\right] + n\left(-n\theta^{-n-1}\right)\int_0^\theta g(t) \cdot t^{n-1}\,\mathrm{d}\,t \\
&= n\theta^{-1}g(\theta) + \left[-n\theta^{-1}\int_0^\theta g(t) \cdot nt^{n-1}\theta^{-n}\,\mathrm{d}\,t\right] \\
\text{(by assumption, } \mathrm{E}[g(T)] = 0) \quad &= n\theta^{-1}g(\theta) + \left(-n\theta^{-1} \cdot 0\right) \\
&= n\theta^{-1}g(\theta).
\end{aligned}$$

We have $n > 0$ and $\theta > 0$, so that $\theta^{-1} > 0$. It follows that $n\theta^{-1}g(\theta)$ is equal to zero if and only if $g(\theta) = 0$, which is true for for all $\theta > 0$. Noting that we will always have $T(\mathbf{X}) = X_{(n)} > 0$, it follows that $P(\{g(T) = 0\}) = 1$, and thus that $T(\mathbf{X})$ is a complete statistic.

REMARK 8.40. The Fundamental Theorem of Calculus applies only to functions that are Riemann-integrable. Thus, the equation

$$\frac{\mathrm{d}}{\mathrm{d}\,\theta}\int_0^\theta g(t)\,\mathrm{d}\,t = g(\theta)$$

is valid only at points of continuity of Riemann-integrable g. The condition of completeness applies to all functions, not just Riemann-integrable ones, so the argument above does not, strictly speaking, show that T is a complete statistic. From a more practical view, though, this distinction is not of concern since the condition of Riemann-integrability is so general that it includes virtually any function we could think of.

THEOREM 8.41 (Basu's Theorem). *If $T(\mathbf{X})$ is a complete and minimal sufficient statistic, then $T(\mathbf{X})$ is independent of every ancillary statistic. (This is Theorem 6.2.24 from Casella & Berger.)*

PROOF. [proof goes here] □

That is, if T depends on some ancillary statistic and T is minimal sufficient, then T cannot be complete. Basu's Theorem allows us to deduce the independence of two statistics without finding their joint distribution. To use Basu's Theorem, we need to show that a statistic is complete.

EXAMPLE 8.42. Let $X \sim \mathrm{Exp}(\theta)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pdf of X, which is given by

$$f_X(x|\theta) = \theta \mathrm{e}^{-\theta x}$$

where $x \geq 0$, so that the joint pdf of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^n f_{X_i}(x_i|\theta) = \prod_{i=1}^n \theta \mathrm{e}^{-\theta x_i} = \theta^n \mathrm{e}^{\sum_{i=1}^n -\theta x_i} = \theta^n \mathrm{e}^{-\theta \sum_{i=1}^n x_i}.$$

Show that $T(\mathbf{X}) = \sum_{i=1}^n X_i$ is independent of $Y = X_n / \sum_{i=1}^n X_i$.

We showed in example 8.21 that $T(\mathbf{X})$ is a minimal sufficient statistic for θ. We can express $f_{\mathbf{X}}(\mathbf{x}|\theta)$ in exponential family form as

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \underbrace{\theta^n}_{c(\theta)} \exp\left\{ \underbrace{\underbrace{-\theta}_{\omega_1(\theta)} \underbrace{\sum_{i=1}^{n} x_i}_{t_1(\mathbf{x})}}_{} \right\}$$

with $h(\mathbf{x}) = 1$. The parameter space for this distribution is given by $\{\omega_1(\theta) = -\theta : \theta > 0\}$, i.e., $k = 1$. The representation of this parameter space in $\mathbb{R}^1$ is the interval $(-\infty, 0)$, which has positive length, i.e., it is an open set, so it follows from theorem 8.31 that $T(\mathbf{X})$ is complete.

The cdf of Y is given by

$$P(\{Y \leq y\}) = P\left(\left\{\frac{X_n}{\sum_{i=1}^{n} X_i} \leq y\right\}\right).$$

Let $Z_1, \ldots, Z_n$ be iid observations from $F_Y(y)$ (corresponding to $\theta = 1$) with $X_i = \theta Z_i$, so that we have

$$P(\{Y \leq y\}) = P\left(\left\{\frac{\theta Z_n}{\sum_{i=1}^{n} \theta Z_i} \leq y\right\}\right) = P\left(\left\{\frac{Z_n}{\sum_{i=1}^{n} Z_i} \leq y\right\}\right).$$

This expression does not depend on θ, so it follows that Y is an ancillary statistic. Because $T(\mathbf{X})$ is a complete and minimal sufficient statistic, it follows from theorem 8.41 that $T(\mathbf{X})$ and Y are independent.

EXAMPLE 8.43. Let $X \sim \mathcal{U}(\theta, \theta + 1)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid. Is $T(\mathbf{X}) = \left(X_{(1)}, X_{(n)}\right)$ a complete statistic?

Any one-to-one function of a minimal sufficient statistic is also minimal sufficient, so it follows that $\left(X_{(1)}, X_{(n)} - X_{(1)}\right)$ is also minimal sufficient. We showed in example 8.25 that the range statistic $R = X_{(n)} - X_{(1)}$ is an ancillary statistic for a uniform random variable. (More generally, we have $X \sim \mathcal{U}(\theta, \theta + 1)$. Suppose that $Z = \mathcal{U}(0, 1)$, so that $X = Z + \theta$. It follows that $X \sim \mathcal{U}(\theta, \theta + 1)$ is a location family, and range is ancillary for a location family.) Thus, $T(\mathbf{X})$ depends on an ancillary statistic, so it follows from theorem 8.41 that $T(\mathbf{X})$ is not complete.

EXAMPLE 8.44. We showed in example 7.4 that the $\mathcal{N}\left(\mu, \sigma^2\right)$ family is an exponential family, with $\omega_1(\boldsymbol{\theta}) = 1/\sigma^2$ and $\omega_2(\boldsymbol{\theta}) = \mu/\sigma^2$. The parameter space for this family is given by

$$\Theta = \left\{\left(\mu, \sigma^2\right) : -\infty < \mu < \infty, \sigma > 0\right\},$$

i.e., $k = 2$. The representation of this parameter space in $\mathbb{R}^2$ is the set

$$\left\{\left(1/\sigma^2, \mu/\sigma^2\right) : (\mu, \sigma) \in \Theta\right\} = \{(x, y) : x > 0, -\infty < y < \infty\},$$

which clearly contains an open set, e.g., the rectangle $\{(x, y) : 1 < x < 2, 1 < y < 2\}$. In example 8.14, we found that $T(\mathbf{X}) = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$ is a sufficient statistic for $\boldsymbol{\theta} = \left(\mu, \sigma^2\right)$. It follows from theorem 8.31 that $T(\mathbf{X})$ is a complete statistic.

CHAPTER 9

Point estimation

Many inferential problems fall into one of three types: point estimation, confidence estimation, or hypothesis testing. Point estimation refers to providing a single "best guess" of some quantity of interest, e.g., a model parameter, a cdf F, a pdf f, a regression function, a prediction for a future value Y of a random variable.

DEFINITION 9.1. A *point estimator* is any function $W(X_1, \ldots, X_n)$ of a sample; that is, any statistic $T(\mathbf{X})$ is a point estimator. An *estimate* $T(\mathbf{x})$ is the observed value of an estimator $T(\mathbf{X})$ (that is, a number) that is obtained when a sample is actually taken.

Often, we are interested in estimating some function $T(\theta)$, e.g., if $X \sim N(\mu, \sigma^2)$, then the parameter is $\boldsymbol{\theta} = (\mu, \sigma^2)$. If our goal is to estimate μ, then $\mu = T(\boldsymbol{\theta})$ is called the *parameter of interest* and σ^2 is called a *nuisance parameter*.

9.1. Methods of finding estimators

9.1.1. Method of moments estimators. Let $X \sim f(x|\theta)$. The rth moment of X is denoted $\mu_r = \mathrm{E}[X^r]$. Note that μ_r is a function of θ, e.g., $\mathrm{E}[X] = \mu_1$, $\mathrm{E}[X^2] = \mu_2$, $\mathrm{Var}(X) = \mu_2 - \mu_1^2$. A method of moments estimator for $\boldsymbol{\theta}$ is obtained by solving

$$\hat{\mu}_r = \sum_{i=1}^{n} \frac{X_i^r}{n}.$$

Moment estimators are not invariant to transformation of the data (a function of x) or to reparameterization of θ (a function of θ).

EXAMPLE 9.2 (Exponential MOM estimator). Let $X \sim \mathrm{Exp}(\theta)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid, i.e., each has the pdf of X, which is given by

$$f_X(x|\theta) = \theta \mathrm{e}^{-\theta x}$$

where $x \geq 0$. Find a moment estimator for θ.
We will find $\hat{\mu}_1 = \mathrm{E}[X]$.

$$\hat{\mu}_1 = \mathrm{E}[X] = \int_0^\infty x \cdot \theta \mathrm{e}^{-\theta x} \mathrm{d}x = \lim_{t \to \infty} \int_0^t x \cdot \theta \mathrm{e}^{-\theta x} \mathrm{d}x$$

Let $u = x$ and $v' = \theta \mathrm{e}^{-\theta x}$, so that $u' = 1$ and $v = -\mathrm{e}^{-\theta x}$. Then, we have

$$\begin{aligned}
\int_0^t uv' \,\mathrm{d}\, x &= [uv|_0^t - \int_0^t vu' \,\mathrm{d}\, x \\
&= -x\mathrm{e}^{-\theta x}\Big|_0^t - \int_0^t -\mathrm{e}^{-\theta x} \,\mathrm{d}\, x \\
&= \left(-t\mathrm{e}^{-\theta t} - 0\right) - \left(\frac{1}{\theta}\mathrm{e}^{-\theta x}\Big|_0^t \right. \\
&= -t\mathrm{e}^{-\theta t} - \left(\frac{1}{\theta}\mathrm{e}^{-\theta t} - \frac{1}{\theta}\mathrm{e}^{0}\right) \\
&= -t\mathrm{e}^{-\theta t} - \frac{1}{\theta}\left(\mathrm{e}^{-\theta t} - 1\right)
\end{aligned}$$

so that

$$\begin{aligned}\hat{\mu}_1 &= \lim_{t\to\infty}\left[-te^{-\theta t} - \frac{1}{\theta}\left(e^{-\theta t}-1\right)\right]\\ &= 0 - \lim_{t\to\infty}\frac{1}{\theta}\left(e^{-\theta t}-1\right)\\ &= -\frac{1}{\theta}\lim_{t\to\infty}\left(e^{-\theta t}-1\right)\\ &= -\frac{1}{\theta}(0-1)\\ &= \frac{1}{\theta}.\end{aligned}$$

Then, we have

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n X_i}{n} \implies \frac{1}{\theta} = \bar{x} \implies \hat{\theta} = \frac{1}{\bar{x}}.$$

REMARK 9.3. Had we used the parameterization $f_X(x|\theta) = (1/\theta)\,e^{-x/\theta}$, we would instead have found $\hat{\theta} = \bar{x}$.

EXAMPLE 9.4. Suppose we have a population with θ members labeled $1,\ldots,\theta$ from which we sample n observations with replacement and record their labels $X_1,\ldots,X_n$. Find a moment estimator for θ.

We are (randomly) sampling with replacement, so an equal-likelihood model is appropriate. Then, it follows that the pmf of X_i is given by

$$p_{X_i}(x_i|\theta) = P(\{X_i = x_i\}) = \frac{1}{\theta}I_{\{x_i\in\{1,\ldots,\theta\}\}},$$

so that the joint pmf of $\mathbf{X} = X_1,\ldots,X_n$ is given by

$$p_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^n\left[\frac{1}{\theta}I_{\{x_i\in\{1,\ldots,\theta\}\}}\right] = \frac{1}{\theta^n}I_{\{x_{(1)}\geq 1\}}I_{\{x_{(n)}\leq\theta\}}.$$

Then, we have

$$\begin{aligned}\hat{\mu}_1 &= \mathrm{E}[X]\\ &= \sum_{x=1}^{\theta} x\cdot p_X(x|\theta)\\ &= \sum_{x=1}^{\theta} x\cdot\frac{1}{\theta}\\ &= \frac{1}{\theta}\sum_{x=1}^{\theta} x\\ &= \frac{1}{\theta}\left[1+2+\cdots+(\theta-1)+\theta\right]\\ &= \frac{1}{\theta}\left(\frac{\theta(\theta+1)}{2}\right)\\ &= \frac{\theta+1}{2}.\end{aligned}$$

We must solve

$$\hat{\mu}_1 = \frac{\sum_{i=1}^n X_i}{n} \implies \frac{\hat{\theta}+1}{2} = \bar{x} \implies \hat{\theta}+1 = 2\bar{x} \implies \hat{\theta} = 2\bar{x}-1.$$

REMARK 9.5. The technique used above to find the sum of the first n numbers involves taking the sums of the respective extremes, e.g., $1+n$, $2+(n-1) = 1+n$, $3+(n-2) = 1+n$, and so on. There are $n/2$ such pairs, so it follows that the sum of the first n numbers is given by

$$\sum_{x=1}^{n} x = \frac{n(n+1)}{2}. \tag{9.1.1}$$

Example 9.6. Suppose

$$X_1, \ldots, X_n \sim F(x|\alpha, \beta) = \begin{cases} 0, & x < 0 \\ (x/\beta)^\alpha, & 0 \le x \le \beta \\ 1 & x > \beta \end{cases}.$$

Find moment estimators for α and β.

We must find estimators for two parameters, so we will need two equations, so we will take the first and second moments. The pdf of X is given by

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}\,x} F_X = \frac{\mathrm{d}}{\mathrm{d}\,x}\left(\frac{x}{\beta}\right)^\alpha = \alpha\left(\frac{x}{\beta}\right)^{\alpha-1}\left(\frac{1}{\beta}\right) = \frac{\alpha}{\beta}\left(\frac{x^{\alpha-1}}{\beta^{\alpha-1}}\right) = \frac{\alpha x^{\alpha-1}}{\beta^\alpha},$$

where $x \in [0, \beta]$ and $f_X(x) = 0$ otherwise. Then, the first moment of X is given by

$$\begin{aligned} \mathrm{E}[X] &= \int_0^\beta x \cdot f_X(x)\,\mathrm{d}\,x \\ &= \int_0^\beta x \cdot \frac{\alpha x^{\alpha-1}}{\beta^\alpha}\,\mathrm{d}\,x \\ &= \frac{\alpha}{\beta^\alpha}\int_0^\beta x^\alpha\,\mathrm{d}\,x \\ &= \frac{\alpha}{\beta^\alpha}\left[\frac{1}{\alpha+1}x^{\alpha+1}\right|_0^\beta \\ &= \frac{\alpha}{\beta^\alpha}\left(\frac{\beta^{\alpha+1}}{\alpha+1} - \frac{0}{\alpha+1}\right) \\ &= \frac{\alpha\beta}{\alpha+1} \end{aligned}$$

and the second moment is given by

$$\begin{aligned} \mathrm{E}\left[X^2\right] &= \int_0^\beta x^2 \cdot \frac{\alpha x^{\alpha-1}}{\beta^\alpha}\,\mathrm{d}\,x \\ &= \frac{\alpha}{\beta^\alpha}\int_0^\beta x^{\alpha+1}\,\mathrm{d}\,x \\ &= \frac{\alpha}{\beta^\alpha}\left[\frac{1}{\alpha+2}x^{\alpha+2}\right|_0^\beta \\ &= \frac{\alpha}{\beta^\alpha}\left(\frac{\beta^{\alpha+2}}{\alpha+2} - \frac{0}{\alpha+2}\right) \\ &= \frac{\alpha\beta^2}{\alpha+2}. \end{aligned}$$

We will solve the first equation for $\hat{\beta}$.

$$\frac{\hat{\alpha}\hat{\beta}}{\hat{\alpha}+1} = \frac{1}{n}\sum_{i=1}^n x_i = \bar{x} \implies \hat{\alpha}\hat{\beta} = \bar{x}(\hat{a}+1) \implies \hat{\beta} = \frac{\bar{x}(\hat{\alpha}+1)}{\hat{\alpha}}$$

We will solve the section equation for $\hat{\alpha}$.

$$\begin{aligned} \frac{\hat{\alpha}\hat{\beta}^2}{\hat{\alpha}+2} &= \frac{1}{n}\sum_{i=1}^n x_i^2 \\ \implies \hat{\alpha}\hat{\beta}^2 &= (\hat{\alpha}+2)\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right) \\ \implies \hat{\alpha}\left(\frac{\bar{x}(\hat{\alpha}+1)}{\hat{\alpha}}\right)^2 &= (\hat{\alpha}+2)\left(\frac{1}{n}\sum_{i=1}^n x_i^2\right) \end{aligned}$$

$$\implies \frac{\bar{x}^2\left(\hat{\alpha}+1\right)^2}{\hat{\alpha}} = \left(\hat{\alpha}+2\right)\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2\right)$$
$$\implies \frac{\left(\hat{\alpha}+1\right)^2}{\hat{\alpha}\left(\hat{\alpha}+2\right)} = \frac{1}{n\bar{x}^2}\sum_{i=1}^{n} x_i^2$$

Let $c = \left(1/n\bar{x}^2\right)\sum_{i=1}^{n} x_i^2$, so that

$$\frac{\left(\hat{\alpha}+1\right)^2}{\hat{\alpha}\left(\hat{\alpha}+2\right)} = c \implies \hat{\alpha}^2 + 2\hat{\alpha} + 1 = c\hat{\alpha}\left(\hat{\alpha}+2\right) = c\hat{\alpha}^2 + 2c\hat{\alpha} \implies \hat{\alpha}^2\left(1-c\right) + \hat{\alpha}\left(2-2c\right) + 1 = 0.$$

Then, the quadratic formula gives

$$\begin{aligned}
x &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \\
\hat{\alpha} &= \frac{-\left(2-2c\right) \pm \sqrt{\left(2-2c\right)^2 - 4\left(1-c\right)\cdot 1}}{2\left(1-c\right)} \\
&= \frac{2\left(c-1\right) \pm \sqrt{\left(2\left(1-c\right)\right)^2 - 4\left(1-c\right)}}{2\left(1-c\right)} \\
&= \frac{2\left(c-1\right) \pm \sqrt{4\left(1-c\right)^2 - 4\left(1-c\right)}}{2\left(1-c\right)} \\
&= \frac{2\left(c-1\right) \pm \sqrt{4\left(\left(1-c\right)^2 - \left(1-c\right)\right)}}{2\left(1-c\right)} \\
&= \frac{2\left(c-1\right) \pm 2\sqrt{1-2c+c^2-1+c}}{2\left(1-c\right)} \\
&= \frac{2\left(c-1\right) \pm 2\sqrt{c^2-c}}{2\left(1-c\right)} \\
&= \frac{c-1 \pm \sqrt{c^2-c}}{1-c},
\end{aligned}$$

so that

$$\hat{\beta} = \frac{\bar{x}\left(\hat{\alpha}+1\right)}{\hat{\alpha}} = \frac{\bar{x}\hat{\alpha}+\bar{x}}{\hat{\alpha}} = \bar{x} + \frac{\bar{x}}{\frac{c-1\pm\sqrt{c^2-c}}{1-c}} = \bar{x}\left(1 + \frac{1-c}{c-1\pm\sqrt{c^2-c}}\right).$$

Method of moments estimators may not be uniquely defined, as can be seen in the following example.

EXAMPLE 9.7 (Poisson MOM estimator). Suppose $X_1, \ldots, X_n \sim \text{Poisson}\left(\lambda\right)$. Recall that $\text{E}\left[X\right] = \text{Var}\left(X\right) = \lambda$. Then, we have

$$\text{E}\left[X\right] = \hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \implies \hat{\lambda} = \bar{x}$$

and

$$\begin{aligned}
\text{Var}\left(X\right) &= \text{E}\left[X^2\right] - \text{E}\left[X\right]^2 \\
\Leftrightarrow \hat{\lambda} &= \hat{\mu}_2 - \hat{\mu}_1^2 \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}^2
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{n\bar{X}^2}{n} \\
&= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2,
\end{aligned}
$$

where the final equality follows from theorem 5.2.

Standard method of moments may not work.

EXAMPLE 9.8. Suppose $X_1, \ldots, X_n \sim f(x|\theta) = \theta x^{-2}$, $0 < \theta \leq x < \infty$. Then,

$$
\begin{aligned}
\mathrm{E}_\theta[X] &= \int_\theta^\infty x\frac{\theta}{x^2}\,\mathrm{d}\,x \\
&= \theta\int_\theta^\infty \frac{1}{x}\,\mathrm{d}\,x \\
&= \theta\lim_{c\to\infty}\int_\theta^c \frac{1}{x}\,\mathrm{d}\,x \\
&= \theta\lim_{c\to\infty}\left(\log x\Big|_\theta^c\right) \\
&= \theta\lim_{c\to\infty}\left(\log c - \log\theta\right) \\
&= \infty,
\end{aligned}
$$

so that $\mu_1(\theta) = \hat{\mu}_1$ has no solution. If we consider

$$
\begin{aligned}
\mathrm{E}_\theta\left[\frac{1}{X}\right] &= \int_\theta^\infty \frac{\theta}{x^3}\,\mathrm{d}\,x \\
&= \lim_{c\to\infty}\int_\theta^c \frac{\theta}{x^3}\,\mathrm{d}\,x \\
&= \lim_{c\to\infty}\left(-\theta\frac{1}{2x^2}\Big|_\theta^c\right) \\
&= \lim_{c\to\infty}\left(-\theta\frac{1}{2c^2} - \left(-\theta\frac{1}{2\theta^2}\right)\right) \\
&= 0 + \frac{1}{2\theta} \\
&= \frac{1}{2\theta},
\end{aligned}
$$

then setting $u_{-1}(\theta) = \hat{\mu}_{-1} = (1/n)\sum_{i=1}^n 1/X_i$ gives

$$
\frac{1}{2\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{X_i} \implies \hat{\theta} = \frac{n}{2\sum_{i=1}^n \frac{1}{X_i}}.
$$

If we consider instead

$$
\begin{aligned}
\mathrm{E}_\theta\left[X^{1/2}\right] &= \int_\theta^\infty x^{1/2}\frac{\theta}{x^2}\,\mathrm{d}\,x \\
&= \lim_{c\to\infty}\left[\theta\int_\theta^c x^{-3/2}\,\mathrm{d}\,x\right] \\
&= \lim_{c\to\infty}\left[\theta\left(-2x^{-1/2}\Big|_\theta^c\right)\right] \\
&= \lim_{c\to\infty}\left[\theta\left(-2c^{-1/2} - \left(-2\theta^{-1/2}\right)\right)\right] \\
&= \theta\left(0 + 2\theta^{-1/2}\right)
\end{aligned}
$$

$$= 2\sqrt{\theta},$$

then setting $\mu_{1/2}(\theta) = \hat{\mu}_{1/2}$ gives

$$2\sqrt{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^{1/2} \implies \hat{\theta} = \frac{\left(\sum_{i=1}^{n} X_i^{1/2}\right)^2}{4n^2}.$$

9.1.2. Maximum likelihood estimators. If $X_1, \ldots, X_n$ are independent random variables from $f_i(x|\theta_1, \ldots, \theta_k)$, then the likelihood function is given by

$$\mathcal{L}(\theta|\mathbf{x}) = \mathcal{L}(\theta_1, \ldots, \theta_k|x_1, \ldots, x_n) = \prod_{i=1}^{n} f_i(x_i|\theta_1, \ldots, \theta_k) = f(\mathbf{x}|\theta),$$

i.e., the likelihood function is just the joint density of the data, except that we treat it as a function of the parameter θ. The likelihood function is not a density function, i.e., it does not integrate to 1. If $\mathbf{X}$ is a discrete random vector, then $\mathcal{L}(\theta|\mathbf{x}) = P_\theta(\{\mathbf{X} = \mathbf{x}\})$. If we compare the likelihood function at two parameter points and find that

$$P_{\theta_1}(\{\mathbf{X} = \mathbf{x}\}) = \mathcal{L}(\theta_1|\mathbf{x}) > \mathcal{L}(\theta_2|\mathbf{x}) = P_{\theta_2}(\{\mathbf{X} = \mathbf{x}\}),$$

then the sample we actually observed is more likely to have occurred if $\theta = \theta_1$ than if $\theta = \theta_2$, which can be interpreted as saying that θ_1 is a more plausible value for the true value of θ than is θ_2.

DEFINITION 9.9. For each sample point $\mathbf{x}$, let $\hat{\theta}(\mathbf{x})$ be a parameter value at which $\mathcal{L}(\theta|\mathbf{x})$ attains its maximum as a function of θ, with $\mathbf{x}$ held fixed. A *maximum likelihood estimator* (MLE) of the parameter θ based on a sample $\mathbf{X}$ is $\hat{\theta}(\mathbf{X})$.

The MLE is the parameter point for which the observed sample is most likely. Let $\mathcal{L}\left(\hat{\theta}|X\right)$ be the maximum of all likelihood functions evaluated at θ, i.e.,

$$\mathcal{L}\left(\hat{\theta}|X\right) = \max\{\mathcal{L}(\theta|X) : \theta \in \Theta\} \Leftrightarrow \hat{\theta} \text{ is an MLE.}$$

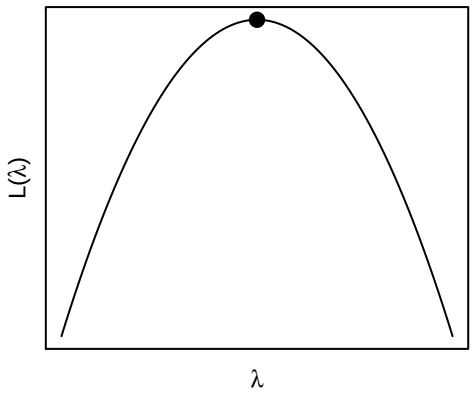


FIGURE 9.1.1. maximizing the likelihood function

If the likelihood is differentiable in θ, possible candidates for $\hat{\theta}$ are those that solve

$$\frac{\partial}{\partial\theta}\log\mathcal{L}(\theta|X) = 0,$$

which gives solutions that find interior extrema, some of which may be minima. To check whether a solution is a maximum, verify that

$$\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}(\theta|X) < 0,$$

i.e., concave. It is often easier to work with $\log\mathcal{L}(\theta|X)$ than with $\mathcal{L}(\theta|X)$, and whatever maximizes the likelihood will also maximize log-likelihood.

EXAMPLE 9.10 (Exponential MLE). Let $X \sim \text{Exp}(\lambda)$, and let $\mathbf{X} = X_1, \ldots, X_n$ be a random sample from the distribution of X, so that the X_i's are iid. Find the MLE of λ and compare it to the MOM estimator.

The pdf of X is given by

$$f_X(x|\lambda) = \frac{1}{\lambda} \mathrm{e}^{-x/\lambda}$$

where $x \geq 0$, so that the pdf of $\mathbf{X}$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\lambda) = \prod_{i=1}^{n} f_X(x_i|\lambda) = \prod_{i=1}^{n} \frac{1}{\lambda} \mathrm{e}^{-x_i/\lambda} = \frac{1}{\lambda^n} \mathrm{e}^{-\sum_{i=1}^{n} x_i/\lambda}.$$

Then, we have $\mathcal{L}(\lambda|\mathbf{x}) = f_{\mathbf{X}}(\mathbf{x}|\lambda)$, so that

$$\begin{aligned} \log \mathcal{L}(\lambda|\mathbf{x}) &= \log\left(\frac{1}{\lambda^n} \mathrm{e}^{-\sum_{i=1}^{n} x_i/\lambda}\right) \\ &= \log\left(\frac{1}{\lambda^n}\right) + \log\left(\mathrm{e}^{-\sum_{i=1}^{n} x_i/\lambda}\right) \\ &= -n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} x_i. \end{aligned}$$

We will take the derivative of $\log \mathcal{L}(\lambda|\mathbf{x})$ with respect to λ.

$$\frac{\partial}{\partial \lambda} \log \mathcal{L}(\lambda|\mathbf{x}) = \frac{\partial}{\partial \lambda}\left[-n \log \lambda - \frac{1}{\lambda} \sum_{i=1}^{n} x_i\right] = -\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} x_i.$$

Setting this equal to zero, we have

$$\frac{n}{\lambda} = \frac{1}{\lambda^2} \sum_{i=1}^{n} x_i \implies n\lambda^2 = \lambda \sum_{i=1}^{n} x_i \implies \hat{\lambda} = \frac{1}{n} \sum_{i=1}^{n} x_i = \bar{x}.$$

We will evaluate the second derivative of $\log \mathcal{L}(\lambda|\mathbf{x})$ at $\lambda = \hat{\lambda}$ to verify that this is a maximum.

$$\begin{aligned} \frac{\partial^2}{\partial \lambda^2} \log \mathcal{L}(\lambda|\mathbf{x})\Big|_{\lambda=\hat{\lambda}} &= \frac{\partial^2}{\partial \lambda^2}\left[-\frac{n}{\lambda} + \frac{1}{\lambda^2} \sum_{i=1}^{n} x_i\right|_{\lambda=\hat{\lambda}} \\ &= \frac{n}{\lambda^2} - \frac{2}{\lambda^3} \sum_{i=1}^{n} x_i\Big|_{\lambda=\hat{\lambda}} \\ &= \frac{n}{\bar{x}^2} - \frac{2}{\bar{x}^3} \cdot n\bar{x} \\ &= \frac{n}{\bar{x}^2} - \frac{2n}{\bar{x}^2} \\ &= -\frac{n}{\bar{x}^2} \end{aligned}$$

We have $n > 0$ and $\bar{x}^2 > 0$, so it follows that $-n/\bar{x}^2 < 0$, therefore $\hat{\lambda} = \bar{X}$ is the MLE. In example 9.2, we found that $\hat{\lambda}_{MOM} = \bar{x}$, so the two estimators agree.

EXAMPLE 9.11. Let's reconsider example 9.4, where we found that the pmf of $\mathbf{X}$ was given by

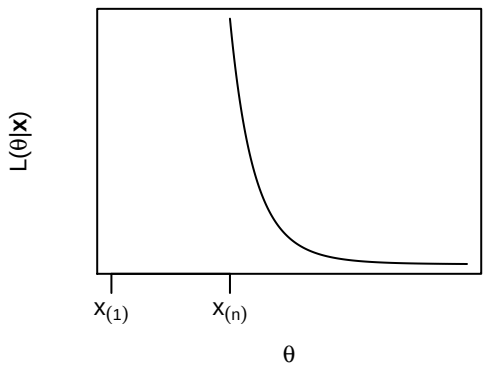
$$p_{\mathbf{X}}(\mathbf{x}|\theta) = \frac{1}{\theta^n} I_{\{x_{(1)} \geq 1\}} I_{\{x_{(n)} \leq \theta\}}.$$

Find the MLE of θ.

We have

$$\mathcal{L}(\theta|\mathbf{x}) = \begin{cases} \frac{1}{\theta^n} I_{\{x_{(1)} \geq 1\}}, & \theta \geq x_{(n)} \\ 0, & \theta < x_{(n)} \end{cases},$$

whose graph is shown below. Clearly, the maximum value of $\mathcal{L}(\theta|\mathbf{x})$ occurs at $x_{(n)}$, so it follows that $\hat{\theta} = X_{(n)}$.



REMARK. When the support of the likelihood function depends on the parameter, it may not be necessary to take the derivative of $\mathcal{L}(\theta|\mathbf{x})$, e.g., the MLE may be found graphically.

EXAMPLE 9.12 (Multinomial MLE). Suppose $X_1, \ldots, X_n$ are sampled from a multinomial distribution with 3 categories and probabilities, e.g., the genotypes AA, Aa, and aa, so that X_i represents the category of the ith observation. Suppose that the pmf of X_i is given by

$$\begin{aligned} P(\{X_i = 1|\theta\}) &= \theta^2 \\ P(\{X_i = 2|\theta\}) &= 2\theta(1-\theta) \\ P(\{X_i = 3|\theta\}) &= (1-\theta)^2 \end{aligned}$$

where $\theta \in (0,1)$. We observe $n_k = \sum_{i=1}^n I_{\{X_i=k\}}$ individuals of type $k \in \{1,2,3\}$. Find the MLE of θ. Note that the pmf specified for X_i is legitimate, i.e., is it non-negative and

$$\theta^2 + 2\theta(1-\theta) + (1-\theta)^2 = \theta^2 + 2\theta - 2\theta^2 + 1 - 2\theta + \theta^2 = 1.$$

We can write the pmf as

$$p_{X_i}(x_i|\theta) = \left(\theta^2\right)^{I_{\{x_i=1\}}} \left[2\theta(1-\theta)\right]^{I_{\{x_i=2\}}} \left[(1-\theta)^2\right]^{I_{\{x_i=3\}}}$$

where each $I_{\{x_i=k\}}$ is equal to 0 or 1. Then, the joint pmf of $\mathbf{X} = X_i, \ldots, X_n$ is given by

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{i=1}^n p_{X_i}(x_i|\theta) \\ &= \prod_{i=1}^n \left(\theta^2\right)^{I_{\{x_i=1\}}} \left[2\theta(1-\theta)\right]^{I_{\{x_i=2\}}} \left[(1-\theta)^2\right]^{I_{\{x_i=3\}}} \\ &= \left(\theta^2\right)^{\sum_{i=1}^n I_{\{x_i=1\}}} \left[2\theta(1-\theta)\right]^{\sum_{i=1}^n I_{\{x_i=2\}}} \left[(1-\theta)^2\right]^{\sum_{i=1}^n I_{\{x_i=3\}}} \end{aligned}$$

so that the likelihood function is given by

$$\begin{aligned} \mathcal{L}(\theta|\mathbf{x}) &= p_{\mathbf{X}}(\mathbf{x}|\theta) \\ &= \left(\theta^2\right)^{\sum_{i=1}^n I_{\{x_i=1\}}} \left[2\theta(1-\theta)\right]^{\sum_{i=1}^n I_{\{x_i=2\}}} \left[(1-\theta)^2\right]^{\sum_{i=1}^n I_{\{x_i=3\}}} \\ &= \left(\theta^2\right)^{n_1} \left[2\theta(1-\theta)\right]^{n_2} \left[(1-\theta)^2\right]^{n_3} \\ &= \theta^{2n_1} \left[2\theta(1-\theta)\right]^{n_2} (1-\theta)^{2n_3} \\ &= 2^{n_2} \theta^{2n_1+n_2} (1-\theta)^{n_2+2n_3} \\ \Leftrightarrow \log \mathcal{L}(\theta|\mathbf{x}) &= \log\left[2^{n_2} \theta^{2n_1+n_2} (1-\theta)^{n_2+2n_3}\right] \\ &= \log 2^{n_2} + \log \theta^{2n_1+n_2} + \log(1-\theta)^{n_2+2n_3} \\ &= n_2 \log 2 + (2n_1+n_2)\log\theta + (n_2+2n_3)\log(1-\theta). \end{aligned}$$

We will take the derivative of the log-likelihood with respect to θ.

$$\begin{aligned}\frac{\partial}{\partial\theta}\log\mathcal{L}\left(\theta|\mathbf{x}\right) &= \frac{\partial}{\partial\theta}\left[n_2\log 2 + \left(2n_1+n_2\right)\log\theta + \left(n_2+2n_3\right)\log\left(1-\theta\right)\right] \\ &= 0 + \frac{2n_1+n_2}{\theta} + \frac{n_2+2n_3}{1-\theta}\cdot -1 \\ &= \frac{2n_1+n_2}{\theta} - \frac{n_2+2n_3}{1-\theta}\end{aligned}$$

Setting this equal to zero, we have

$$\begin{aligned}\frac{2n_1+n_2}{\theta} - \frac{n_2+2n_3}{1-\theta} &= 0 \\ \Leftrightarrow \frac{n_2+2n_3}{1-\theta} &= \frac{2n_1+n_2}{\theta} \\ \Leftrightarrow \theta\left(n_2+2n_3\right) &= \left(1-\theta\right)\left(2n_1+n_2\right) \\ \Leftrightarrow \frac{1-\theta}{\theta} &= \frac{n_2+2n_3}{2n_1+n_2} \\ \Leftrightarrow \frac{1}{\theta} - 1 &= \frac{n_2+2n_3}{2n_1+n_2} \\ \Leftrightarrow \frac{1}{\theta} &= \frac{n_2+2n_3}{2n_1+n_2} + 1 \\ &= \frac{n_2+2n_3+2n_1+n_2}{2n_1+n_2} \\ &= \frac{2\left(n_1+n_2+n_3\right)}{2n_1+n_2} \\ &= \frac{2n}{2n_1+n_2} \\ \Leftrightarrow \hat{\theta} &= \frac{2n_1+n_2}{2n}.\end{aligned}$$

We will evaluate the second derivative of the log-likelihood at $\theta=\hat{\theta}$ to verify that $\hat{\theta}$ is a maximum.

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\log\mathcal{L}\left(\theta|\mathbf{x}\right)\Big|_{\theta=\hat{\theta}} &= \frac{\partial}{\partial\theta}\left[\frac{2n_1+n_2}{\theta} - \frac{n_2-2n_3}{1-\theta}\right|_{\theta=\hat{\theta}} \\ &= -\frac{2n_1+n_2}{\theta^2} - \left(\frac{n_2-2n_3}{\left(1-\theta\right)^2}\cdot -1\cdot -1\right) \\ &= -\frac{2n_1+n_2}{\theta^2} - \frac{n_2-2n_3}{\left(1-\theta\right)^2} \\ &= -\left[\frac{2n_1+n_2}{\theta^2} + \frac{n_2-2n_3}{\left(1-\theta\right)^2}\right]\end{aligned}$$

We have $n_k \geq 0$ and $n>0$, i.e., at least one of n_k is positive, and we have $\theta^2>0$ and $\left(1-\theta\right)^2>0$, so that the sum above is positive, so that the expression above is negative. It follows that $\hat{\theta}=\left(2n_1+n_2\right)/2n$ is the MLE.

EXAMPLE 9.13 (Poisson MLE). Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} \text{Poisson}\left(\lambda\right)$, where $\lambda>0$. Find the MLE of λ.
In example 8.12, we found that the joint pmf of $\mathbf{X}=X_1,\ldots,X_n$ was given by

$$p_{\mathbf{X}}\left(\mathbf{x}|\lambda\right) = \left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)\mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i},$$

so that the log-likelihood is given by

$$\log\mathcal{L}\left(\lambda|\mathbf{x}\right) = \log\left[\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)\mathrm{e}^{-n\lambda}\lambda^{\sum_{i=1}^{n}x_i}\right]$$

$$
\begin{aligned}
&= \log\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right) + \log \mathrm{e}^{-n\lambda} + \log \lambda^{\sum_{i=1}^{n} x_i} \\
&= \left(\log\frac{1}{x_1!} + \cdots + \log\frac{1}{x_n!}\right) - n\lambda + \left(\sum_{i=1}^{n} x_i\right)\log\lambda \\
&= \left[(\log 1 - \log x_1!) + \cdots + (\log 1 - \log x_n!)\right] - n\lambda + \left(\sum_{i=1}^{n} x_i\right)\log\lambda \\
&= \left[(0 - \log x_1!) + \cdots + (0 - \log x_n!)\right] - n\lambda + \log\lambda\left(\sum_{i=1}^{n} x_i\right) \\
&= \sum_{i=1}^{n} -\log x_i! - n\lambda + \log\lambda\left(\sum_{i=1}^{n} x_i\right) \\
&= -n\lambda + \log\lambda\left(\sum_{i=1}^{n} x_i\right) - \sum_{i=1}^{n}\log x_i!.
\end{aligned}
$$

We will take the derivative of the log-likelihood with respect to λ.

$$
\frac{\partial}{\partial\lambda}\log\mathcal{L}\left(\lambda|\mathbf{x}\right) = \frac{\partial}{\partial\lambda}\left[-n\lambda + \log\lambda\left(\sum_{i=1}^{n} x_i\right) - \sum_{i=1}^{n}\log x_i!\right] = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i
$$

Setting this equal to zero, we have

$$
-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \implies \frac{1}{\lambda}\sum_{i=1}^{n} x_i = n \implies \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.
$$

We will evaluate the second derivative of the log-likelihood at $\lambda = \hat{\lambda}$ to verify that $\hat{\lambda}$ is a maximum.

$$
\begin{aligned}
\frac{\partial^2}{\partial\lambda^2}\log\mathcal{L}\left(\lambda|\mathbf{x}\right)\Big|_{\lambda=\hat{\lambda}} &= \frac{\partial}{\partial\lambda}\left[-n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i\right|_{\lambda=\hat{\lambda}} \\
&= -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i\Big|_{\lambda=\hat{\lambda}} \\
&= -\frac{1}{\bar{x}^2}\sum_{i=1}^{n} x_i \\
&= -\frac{1}{\bar{x}^2} n\bar{x} \\
&= -\frac{n}{\bar{x}}
\end{aligned}
$$

We have $n > 0$ and $\bar{x} \geq 0$ (because the x_i's are each non-negative), so that $-n/\bar{x} < 0$ (assuming that $\bar{x} > 0$). It follows that $\hat{\lambda} = \bar{x}$ is the MLE.

Example 9.14 (Uniform MLE, single-parameter case). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. Find the MLE of θ.

The pdf of X_i is given by

$$
\begin{aligned}
f_{X_i}\left(x_i\right) &= \frac{1}{\left(\theta + \frac{1}{2}\right) - \left(\theta - \frac{1}{2}\right)} I_{\left\{\theta - \frac{1}{2} < x_i < \theta + \frac{1}{2}\right\}} \\
&= \frac{1}{1} I_{\left\{\theta - \frac{1}{2} < x_i < \theta + \frac{1}{2}\right\}} \\
&= I_{\left\{\theta - \frac{1}{2} < x_i < \theta + \frac{1}{2}\right\}},
\end{aligned}
$$

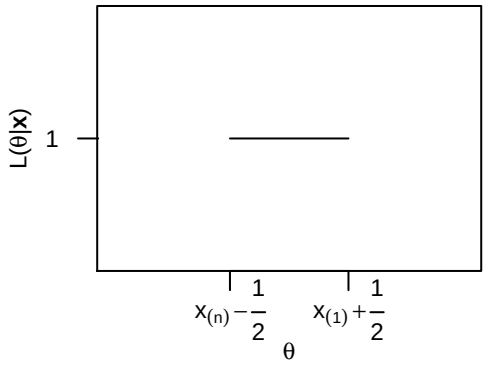
so that the joint pdf of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{i=1}^{n} I_{\left\{\theta-\frac{1}{2}<x_i<\theta+\frac{1}{2}\right\}} \\ &= I_{\left\{\theta-\frac{1}{2}<x_1<\theta+\frac{1}{2}\right\}} \cdot \ldots \cdot I_{\left\{\theta-\frac{1}{2}<x_n<\theta+\frac{1}{2}\right\}} \\ &= I_{\left\{x_{(1)}>\theta-\frac{1}{2}\right\}} I_{\left\{x_{(n)}<\theta+\frac{1}{2}\right\}}. \end{aligned}$$

Then, the likelihood function is given by

$$\mathcal{L}(\theta|\mathbf{x}) = \begin{cases} 1, & x_{(n)} - \frac{1}{2} < \theta < x_{(1)} + \frac{1}{2} \\ 0, & \text{otherwise} \end{cases},$$

and its graph is shown below. Clearly, $\mathcal{L}(\theta|\mathbf{x})$ is constant between $x_{(n)} - \frac{1}{2}$ and $x_{(1)} + \frac{1}{2}$ and 0 elsewhere, so that the MLE $\hat{\theta}$ is given by any point in the interval $\left(x_{(n)} - \frac{1}{2}, x_{(1)} + \frac{1}{2}\right)$.



EXAMPLE 9.15. Let $X_1, \ldots X_n \overset{\text{iid}}{\sim} f(x|\theta) = \theta/x^2$, $0 < \theta \leq x < \infty$. Find the MLE of θ.
The support of X depends on the parameter θ, so we will rewrite the pdf of X as

$$f_X(x|\theta) = \frac{\theta}{x^2} I_{\{x \geq \theta\}}.$$

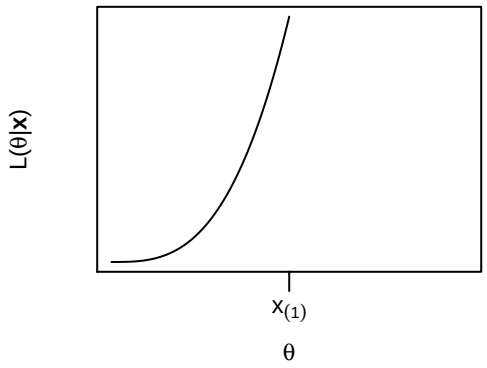
Then, the joint pdf of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_{X_i}(x_i|\theta) = \prod_{i=1}^{n} \frac{\theta}{x_i^2} I_{\{x_i \geq \theta\}} = \frac{\theta^n}{\prod_{i=1}^{n} x_i^2} I_{\left\{x_{(1)} \geq \theta\right\}},$$

so that the likelihood function is given by

$$\mathcal{L}(\theta|\mathbf{x}) = \begin{cases} \theta^n / \prod_{i=1}^{n} x_i^2, & \theta \leq x_{(1)} \\ 0, & \theta > x_{(1)} \end{cases}.$$

The graph of the likelihood function is shown below. Clearly, the maximum value of $\mathcal{L}(\theta|\mathbf{x})$ occurs at $x_{(1)}$, so it follows that $\hat{\theta} = X_{(1)}$.



EXAMPLE 9.16 (Uniform MLE, two-parameter case). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{U}\left(\mu - \sqrt{3}\sigma, \mu + \sqrt{3}\sigma\right)$. Find the MLEs of μ and σ.

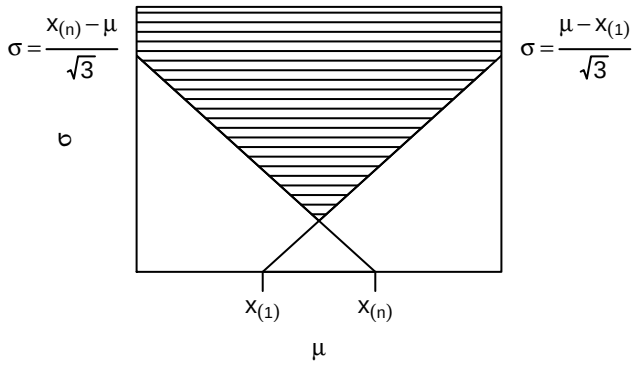
The pdf of X_i is given by

$$\begin{aligned} f_{X_i}\left(x_i|\mu,\sigma\right) &= \frac{1}{\left(\mu+\sqrt{3}\sigma\right)-\left(\mu-\sqrt{3}\sigma\right)} I_{\left\{\mu-\sqrt{3}\sigma<x_i<\mu+\sqrt{3}\sigma\right\}} \\ &= \frac{1}{2\sqrt{3}\sigma} I_{\left\{\mu-\sqrt{3}\sigma<x_i<\mu+\sqrt{3}\sigma\right\}}, \end{aligned}$$

so that the joint pdf of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$\begin{aligned} f_{\mathbf{X}}\left(\mathbf{x}|\mu,\sigma\right) &= \prod_{i=1}^{n} f_{X_i}\left(x_i|\mu,\sigma\right) \\ &= \prod_{i=1}^{n} \frac{1}{2\sqrt{3}\sigma} I_{\left\{\mu-\sqrt{3}\sigma<x_i<\mu+\sqrt{3}\sigma\right\}} \\ &= \left(2\sqrt{3}\sigma\right)^{-n} I_{\left\{x_{(1)}>\mu-\sqrt{3}\sigma\right\}} I_{\left\{x_{(n)}<\mu+\sqrt{3}\sigma\right\}}. \end{aligned}$$

To maximize $\mathcal{L}$, we note that $\left(2\sqrt{3}\sigma\right)^{-n}$ increases as σ decreases. Thus, we must find the minimum value of σ such that $\mathcal{L}$ is positive, which is true when $\sigma > \left(\mu - x_{(1)}\right)/\sqrt{3}$ and $\sigma > \left(x_{(n)} - \mu\right)/\sqrt{3}$. In the graph below, the region of positivity of $\mathcal{L}$ is shaded.



The minimum value of σ such that $\mathcal{L}$ is positive occurs precisely when the two lines intersect, i.e., when

$$\frac{\mu - x_{(1)}}{\sqrt{3}} = \frac{x_{(n)} - \mu}{\sqrt{3}} \implies \mu - x_{(1)} = x_{(n)} - \mu \implies 2\mu = x_{(n)} + x_{(1)} \implies \hat{\mu} = \frac{x_{(n)} + x_{(1)}}{2}.$$

We substitute into one of the line equations to find $\hat{\sigma}$.

$$\hat{\sigma} = \frac{\hat{\mu} - x_{(1)}}{\sqrt{3}} = \frac{\frac{1}{2}\left(x_{(n)} + x_{(1)}\right) - x_{(1)}}{\sqrt{3}} = \frac{x_{(n)} + x_{(1)} - 2x_{(1)}}{2\sqrt{3}} = \frac{x_{(n)} - x_{(1)}}{2\sqrt{3}}$$

Thus, the MLEs for μ and σ are given by $\hat{\mu} = \left(x_{(n)} + x_{(1)}\right)/2$ and $\hat{\sigma} = \left(x_{(n)} - x_{(1)}\right)/2\sqrt{3}$, respectively.

EXAMPLE 9.17. Suppose that X_i and Y_i are independent for $i = 1, \ldots, n$, where $X_i \sim f\left(x|\lambda\right) = \left(1/\lambda\right) e^{-x/\lambda}$ and $Y_i \sim f\left(y|\mu\right) = \left(1/\mu\right) e^{-y/\mu}$. We observe

$$Z_i = \min\left(X_i, Y_i\right) \quad \text{and} \quad W_i = \begin{cases} 1, & \text{if } Z_i = X_i \\ 0, & \text{if } Z_i = Y_i \end{cases}.$$

Find the MLEs of μ and λ.

In the case where $W_i = 1$, we have $Z_i = \min\left(X_i, Y_i\right) = X_i$, so that our information about Y_i is limited to $Y_i > Z_i$, and we have

$$\begin{aligned} P\left(\{W_i = 1\} \cap \{Z_i = z_i\}\right) &= P\left(\{X_i = z_i\} \cap \{Y_i > z_i\}\right) \\ &= P\left(\{X_i = z_i\}\right) \cdot P\left(\{Y_i > z_i\}\right) \\ &= f_{X_i}\left(z_i|\lambda\right) \cdot \left(1 - F_{Y_i}\left(z_i|\mu\right)\right). \end{aligned}$$

In the case where $W_i = 0$, we have $Z_i = \min(X_i, Y_i) = Y_i$, so that our information about X_i is limited to $X_i > Z_i$, and we have

$$\begin{aligned} P(\{W_i = 0\} \cap \{Z_i = z_i\}) &= P(\{Y_i = z_i\} \cap \{X_i > z_i\}) \\ &= P(\{Y_i = z_i\}) \cdot P(\{X_i > z_i\}) \\ &= f_{Y_i}(z_i|\mu) \cdot (1 - F_{X_i}(z_i|\lambda)). \end{aligned}$$

The cdf of Y_i is given by

$$F_{Y_i}(y|\mu) = \int_0^y f_{Y_i}(t|\mu)\,\mathrm{d}t = \int_0^y \frac{1}{\mu} e^{-t/\mu} \mathrm{d}t = \left[-e^{-t/\mu}\right|_0^y = -e^{-y/\mu} - \left(-e^0\right) = 1 - e^{-y/\mu}.$$

By symmetry, the cdf of X_i is given by $F_{X_i}(x|\lambda) = 1 - e^{-x/\lambda}$. So, we have

$$P(\{W_i = 1\} \cap \{Z_i = z_i\}) = \frac{1}{\lambda} e^{-z_i/\lambda} \cdot e^{-z_i/\mu}$$

and

$$P(\{W_i = 0\} \cap \{Z_i = z_i\}) = \frac{1}{\mu} e^{-z_i/\mu} \cdot e^{-z_i/\lambda}.$$

Then, the joint pdf of W_i and Z_i is given by

$$\begin{aligned} f_{W_i,Z_i}(w_i, z_i|\lambda, \mu) &= P(\{Z_i = z_i\} \cap \{W_i = w_i\}) \\ &= \left[\frac{1}{\lambda} e^{-z_i/\lambda} e^{-z_i/\mu}\right]^{w_i} \left[\frac{1}{\mu} e^{-z_i/\mu} e^{-z_i/\lambda}\right]^{1-w_i} \\ &= \left(\frac{1}{\lambda}\right)^{w_i} e^{(-z_i/\lambda)w_i} e^{(-z_i/\mu)w_i} \left(\frac{1}{\mu}\right)^{1-w_i} e^{(-z_i/\mu)(1-w_i)} e^{(-z_i/\lambda)(1-w_i)} \\ &= \left(\frac{1}{\lambda}\right)^{w_i} \left(\frac{1}{\mu}\right)^{1-w_i} \exp\left\{-w_i\frac{z_i}{\lambda} - w_i\frac{z_i}{\mu} - \frac{z_i}{\mu} + w_i\frac{z_i}{\mu} - \frac{z_i}{\lambda} + w_i\frac{z_i}{\lambda}\right\} \\ &= \left(\frac{1}{\lambda}\right)^{w_i} \left(\frac{1}{\mu}\right)^{1-w_i} \exp\left\{-\frac{z_i}{\mu} - \frac{z_i}{\lambda}\right\} \\ &= \left(\frac{1}{\lambda}\right)^{w_i} \left(\frac{1}{\mu}\right)^{1-w_i} e^{-z_i/\mu} e^{-z_i/\lambda}, \end{aligned}$$

so that the joint pdf of $\mathbf{W} = W_1, \ldots, W_n$ and $\mathbf{Z} = Z_1, \ldots, Z_n$ is given by

$$\begin{aligned} f_{\mathbf{W},\mathbf{Z}}(\mathbf{w}, \mathbf{z}|\lambda, \mu) &= \prod_{i=1}^n f_{W_i,Z_i}(w_i, z_i|\lambda, \mu) \\ &= \prod_{i=1}^n \left[\left(\frac{1}{\lambda}\right)^{w_i} \left(\frac{1}{\mu}\right)^{1-w_i} e^{-z_i/\mu} e^{-z_i/\lambda}\right] \\ &= \prod_{i=1}^n \left[\left(\frac{1}{\lambda}\right)^{w_i} \left(\frac{1}{\mu}\right)^{1-w_i} \exp\left\{-z_i\left(\frac{1}{\mu} + \frac{1}{\lambda}\right)\right\}\right] \\ &= \lambda^{-\sum_{i=1}^n w_i} \mu^{-\sum_{i=1}^n (1-w_i)} \exp\left\{-\sum_{i=1}^n z_i\left(\frac{1}{\mu} + \frac{1}{\lambda}\right)\right\}. \end{aligned}$$

Then, the log-likelihood is given by

$$\begin{aligned} \log L(\lambda, \mu|\mathbf{w}, \mathbf{z}) &= \log\left[\lambda^{-\sum_{i=1}^n w_i} \mu^{-\sum_{i=1}^n (1-w_i)} \exp\left\{-\sum_{i=1}^n z_i\left(\frac{1}{\mu} + \frac{1}{\lambda}\right)\right\}\right] \\ &= \log \lambda^{-\sum_{i=1}^n w_i} + \log \mu^{-\sum_{i=1}^n (1-w_i)} - \sum_{i=1}^n z_i\left(\frac{1}{\mu} + \frac{1}{\lambda}\right) \\ &= -\log\lambda \sum_{i=1}^n w_i - \log\mu \sum_{i=1}^n (1 - w_i) - \left(\frac{1}{\mu} + \frac{1}{\lambda}\right) \sum_{i=1}^n z_i. \end{aligned}$$

We will take the derivative of $\log L$ with respect to λ.

$$\begin{aligned}\frac{\partial}{\partial\lambda}\log L\left(\lambda,\mu|\mathbf{w},\mathbf{z}\right) &= \frac{\partial}{\partial\lambda}\left[-\log\lambda\sum_{i=1}^{n}w_i-\log\mu\sum_{i=1}^{n}\left(1-w_i\right)-\left(\frac{1}{\mu}+\frac{1}{\lambda}\right)\sum_{i=1}^{n}z_i\right]\\ &= -\frac{1}{\lambda}\sum_{i=1}^{n}w_i-0+\frac{1}{\lambda^2}\sum_{i=1}^{n}z_i\\ &= -\frac{1}{\lambda}\sum_{i=1}^{n}w_i+\frac{1}{\lambda^2}\sum_{i=1}^{n}z_i\end{aligned}$$

Setting this equal to zero, we have

$$\begin{aligned}\frac{1}{\hat{\lambda}^2}\sum_{i=1}^{n}z_i &= \frac{1}{\hat{\lambda}}\sum_{i=1}^{n}w_i\\ \Leftrightarrow \hat{\lambda}^2\sum_{i=1}^{n}w_i &= \hat{\lambda}\sum_{i=1}^{n}z_i\\ \Leftrightarrow \hat{\lambda} &= \frac{\sum_{i=1}^{n}z_i}{\sum_{i=1}^{n}w_i}.\end{aligned}$$

We will now take the derivative of $\log L$ with respect to μ.

$$\begin{aligned}\frac{\partial}{\partial\mu}\log L\left(\lambda,\mu|\mathbf{w},\mathbf{z}\right) &= \frac{\partial}{\partial\mu}\left[-\log\lambda\sum_{i=1}^{n}w_i-\log\mu\sum_{i=1}^{n}\left(1-w_i\right)-\left(\frac{1}{\mu}+\frac{1}{\lambda}\right)\sum_{i=1}^{n}z_i\right]\\ &= 0-\frac{1}{\mu}\sum_{i=1}^{n}\left(1-w_i\right)+\frac{1}{\mu^2}\sum_{i=1}^{n}z_i\\ &= -\frac{1}{\mu}\sum_{i=1}^{n}\left(1-w_i\right)+\frac{1}{\mu^2}\sum_{i=1}^{n}z_i.\end{aligned}$$

Setting this equal to zero, we have

$$\begin{aligned}\frac{1}{\hat{\mu}^2}\sum_{i=1}^{n}z_i &= \frac{1}{\hat{\mu}}\sum_{i=1}^{n}\left(1-w_i\right)\\ \Leftrightarrow \hat{\mu}^2\sum_{i=1}^{n}\left(1-w_i\right) &= \hat{\mu}\sum_{i=1}^{n}z_i\\ \Leftrightarrow \hat{\mu} &= \frac{\sum_{i=1}^{n}z_i}{\sum_{i=1}^{n}\left(1-w_i\right)}.\end{aligned}$$

REMARK. The above example is a common setting for survival (or failure) analysis. For example, X_i might be the time of cure for the ith subject, and Y_i the follow-up time.

THEOREM 9.18 (Invariance property of MLEs). *If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau\left(\hat{\theta}\right)$. (This is Theorem 7.2.10 from Casella & Berger; the following proof is given there.)*

PROOF. Let $\hat{\eta}$ denote the value the maximizes $L^*(\eta|\mathbf{x})$. We must show that $L^*(\hat{\eta}|\mathbf{x}) = L^*\left[\tau\left(\hat{\theta}\right)|\mathbf{x}\right]$. Now, the maxima of L and L^* coincide, so we have

$$L^*(\hat{\eta}|\mathbf{x}) = \sup_{\eta}\sup_{\{\theta:\tau(\theta)=\eta\}}L(\theta|\mathbf{x}) = \sup_{\theta}L(\theta|\mathbf{x}) = L\left(\hat{\theta}|\mathbf{x}\right),$$

where the second equality follows because the iterated maximization is equal to the unconditional maximization over θ, which is attained at $\hat{\theta}$. Furthermore

$$L\left(\hat{\theta}|\mathbf{x}\right) = \sup_{\left\{\theta:\tau(\theta)=\tau\left(\hat{\theta}\right)\right\}}L(\theta|\mathbf{x}) = L^*\left[\tau\left(\hat{\theta}\right)|\mathbf{x}\right].$$

Hence the string of equalities shows that $L^*\left(\hat{\eta}|\mathbf{x}\right) = L^*\left(\tau\left(\hat{\theta}\right)|\mathbf{x}\right)$ and that $\tau\left(\hat{\theta}\right)$ is the MLE of $\tau\left(\theta\right)$. □

EXAMPLE 9.19. Suppose that $X_1, \ldots, X_n \overset{\text{iid}}{\sim} \text{Exp}\left(\lambda\right)$, so that the pdf of X_i is given by $f_{X_i}\left(x|\lambda\right) = \lambda \text{e}^{-\lambda x}$, and the MLE of λ is $\hat{\lambda} = 1/\bar{X}$.

(1) Find the MLE of $P\left(\{X > a\}\right)$.

Let $\eta = P\left(\{X > a\}\right)$, so that we have

$$\begin{aligned}
\eta &= P\left(\{X > a\}\right) \\
&= \int_a^\infty f_X\left(x|\lambda\right) \text{d}x \\
&= \lim_{c\to\infty} \int_a^c \lambda \text{e}^{-\lambda x} \text{d}x \\
&= \lim_{c\to\infty} \left[-\text{e}^{-\lambda x}\right|_a^c \\
&= \lim_{c\to\infty} \left(-\text{e}^{-\lambda c} - \left(-\text{e}^{-\lambda a}\right)\right) \\
&= \lim_{c\to\infty} \left(-\text{e}^{-\lambda c} + \text{e}^{-\lambda a}\right) \\
&= 0 + \text{e}^{-\lambda a} \\
&= \text{e}^{-\lambda a}.
\end{aligned}$$

Then, by theorem 9.18, we have $\hat{\eta} = \text{e}^{-(1/\bar{x})a} = \text{e}^{-a/\bar{x}}$. Note that

$$\log \eta = \log \text{e}^{-\lambda a} \implies -\lambda a = \log \eta \implies \lambda = -\frac{\log \eta}{a},$$

so that

$$f_{X_i}\left(x|\eta\right) = \left(-\frac{\log \eta}{a}\right) \text{e}^{-(-\log \eta/a)x} = \left(-\frac{\log \eta}{a}\right) \text{e}^{(x \log \eta)/a}.$$

Then, the joint pdf of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$f_{\mathbf{X}}\left(\mathbf{x}|\eta\right) = \prod_{i=1}^n f_{X_i}\left(x|\eta\right) = \prod_{i=1}^n \left(-\frac{\log \eta}{a}\right) \text{e}^{(x_i \log \eta)/a} = \left(-\frac{\log \eta}{a}\right)^n \text{e}^{\sum_{i=1}^n (x_i \log \eta)/a},$$

so that the log-likelihood function is given by

$$\begin{aligned}
\log \mathcal{L}\left(\eta|\mathbf{x}\right) &= \log \left[\left(-\frac{\log \eta}{a}\right)^n \text{e}^{\sum_{i=1}^n (x_i \log \eta)/a}\right] \\
&= \log \left[\left(-\frac{\log \eta}{a}\right)^n\right] + \log \left[\text{e}^{\sum_{i=1}^n (x_i \log \eta)/a}\right] \\
&= n \log \left(-\frac{\log \eta}{a}\right) + \sum_{i=1}^n \left(\frac{x_i \log \eta}{a}\right) \\
&= n \log \left(-\frac{\log \eta}{a}\right) + \frac{\log \eta}{a} \sum_{i=1}^n x_i.
\end{aligned}$$

Taking the derivative with respect to η, we have

$$\begin{aligned}
\frac{\partial}{\partial \eta} \log \mathcal{L}\left(\eta|\mathbf{x}\right) &= \frac{\partial}{\partial \eta} \left[n \log \left(-\frac{\log \eta}{a}\right) + \log \eta \left(a \sum_{i=1}^n x_i\right)\right] \\
&= \frac{n}{-\frac{\log \eta}{a}} \left(-\frac{1}{\eta a}\right) + \frac{1}{\eta a} \sum_{i=1}^n x_i \\
&= \frac{n}{\eta \log \eta} + \frac{1}{\eta a} \sum_{i=1}^n x_i.
\end{aligned}$$

Setting this equal to zero, we have

$$\begin{aligned}
&\frac{n}{\hat{\eta}\log\hat{\eta}}+\frac{1}{\hat{\eta}a}\sum_{i=1}^{n}x_i=0\\
&\qquad\Longrightarrow\quad\frac{n}{\hat{\eta}\log\hat{\eta}}=-\frac{1}{\hat{\eta}a}\sum_{i=1}^{n}x_i\\
&\Longrightarrow\quad\hat{\eta}\log\hat{\eta}\sum_{i=1}^{n}x_i=-n\hat{\eta}a\\
&\qquad\Longrightarrow\quad\log\hat{\eta}=-\frac{na}{\sum_{i=1}^{n}x_i}\\
&\qquad\qquad=-\frac{a}{\bar{x}}\\
&\qquad\qquad\Longrightarrow\quad\hat{\eta}=\mathrm{e}^{-a/\bar{x}},
\end{aligned}$$

which agrees with the result obtained from the theorem.

(2) Find the MLE of median $(X_1,\ldots,X_n)$.

The median of a distribution is a value m such that $P(\{Y\leq m\})\geq\frac{1}{2}$ and $P(\{Y\geq m\})\geq\frac{1}{2}$. If Y is continuous, m satisfies

$$\int_{-\infty}^{m}f(y)\,\mathrm{d}\,y=\int_{m}^{\infty}f(y)\,\mathrm{d}\,y=\frac{1}{2}.$$

X is continuous, so it follows that

$$\frac{1}{2}=P(\{X\leq m\})=F_X(m).$$

The cdf of X is given by

$$F_X(x)=\int_0^x f_X(x)\,\mathrm{d}\,x=\int_0^x\lambda\mathrm{e}^{-\lambda x}\,\mathrm{d}\,x=-\mathrm{e}^{-\lambda x}\Big|_0^x=-\mathrm{e}^{-\lambda x}-\left(-\mathrm{e}^{0}\right)=1-\mathrm{e}^{-\lambda x},$$

so it follows that

$$F_X(m)=1-\mathrm{e}^{-\lambda m}=\frac{1}{2}\implies\mathrm{e}^{-\lambda m}=\frac{1}{2}\implies-\lambda m=\log\left(2^{-1}\right)=-\log 2\implies m=\frac{\log 2}{\lambda}.$$

Then, by theorem 9.18, we have

$$\hat{m}=\frac{\log 2}{\hat{\lambda}}=\frac{\log 2}{1/\bar{x}}=\bar{x}\log 2.$$

THEOREM 9.20. *If an MLE is unique, then it is a function of the sufficient statistics.*

PROOF. Suppose that $X_1,\ldots,X_n$ is a random sample from the distribution of X, whose pdf is given by $f(x|\theta)$. Suppose also that $\hat{\theta}$ is the MLE of θ, and that the associated likelihood function is given by $\mathcal{L}(\theta|\mathbf{x})$. Then, we have

$$\mathcal{L}(\theta|\mathbf{x})=f(\mathbf{x}|\theta)=h(\mathbf{x})\,g(T(\mathbf{x})|\theta),$$

where the final equality follows from theorem 8.11. Then, maximizing $\mathcal{L}(\theta|\mathbf{x})$ with respect to θ is equivalent to maximizing $g(T(\mathbf{x})|\theta)$, which is a function of $T(\mathbf{x})$, which is a sufficient statistic. □

Often, it is not possible to find the MLE analytically and we need to use numerical methods. Two commonly used methods are the Newton-Raphson algorithm and the Expectation-Maximization (EM) algorithm. Both are iterative methods that produce a sequence of values $\theta^{(0)},\theta^{(1)},\ldots$ that, under ideal conditions, converge to the MLE $\hat{\theta}$. It is helpful to use a good starting value $\theta^{(0)}$. Often, the method of moments estimator is a good starting value.

9.1.2.1. *Newton-Raphson.* Let $\ell(\theta|\mathbf{x}) = \log \mathcal{L}(\theta|\mathbf{x})$. When we maximize the log-likelihood by setting its derivative equal to zero, we are solving the equation

$$\frac{\partial}{\partial\theta}\ell(\theta|\mathbf{x}) = 0.$$

To motivate Newton-Raphson, we will expand the derivative of $\ell(\theta|\mathbf{x})$ around $\theta^{(j)}$. Recall from definition 5.10 that the Taylor series expansion of f around a is given by

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{(n)}(a)}{n!}(x-a)^n = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots,$$

so that we have

$$\ell'\left(\hat{\theta}\right) \approx \ell'\left(\theta^{(j)}\right) + \ell''\left(\theta^{(j)}\right)\left(\hat{\theta} - \theta^{(j)}\right).$$

Setting ℓ' equal to zero and solving for $\hat{\theta}$, we have

$$\begin{aligned}
0 &= \ell'\left(\hat{\theta}\right) \\
&\approx \ell'\left(\theta^{(j)}\right) + \ell''\left(\theta^{(j)}\right)\left(\hat{\theta} - \theta^{(j)}\right) \\
&\approx \ell'\left(\theta^{(j)}\right) + \hat{\theta}\ell''\left(\theta^{(j)}\right) - \theta^{(j)}\ell''\left(\theta^{(j)}\right) \\
\implies \hat{\theta}\ell''\left(\theta^{(j)}\right) &\approx \theta^{(j)}\ell''\left(\theta^{(j)}\right) - \ell'\left(\theta^{(j)}\right) \\
\implies \hat{\theta} &\approx \frac{\theta^{(j)}\ell''\left(\theta^{(j)}\right) - \ell'\left(\theta^{(j)}\right)}{\ell''\left(\theta^{(j)}\right)} \\
&\approx \theta^{(j)} - \frac{\ell'\left(\theta^{(j)}\right)}{\ell''\left(\theta^{(j)}\right)}.
\end{aligned}$$

This suggests the following iterative scheme:

$$\hat{\theta}^{(j+1)} = \theta^{(j)} - \frac{\ell'\left(\theta^{(j)}\right)}{\ell''\left(\theta^{(j)}\right)}. \tag{9.1.2}$$

We will continue to improve the estimator in this way until convergence.

In the multiparameter case, the MLE $\hat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \ldots, \hat{\theta}_k\right)$ is a vector and the method becomes

$$\hat{\boldsymbol{\theta}}^{(j+1)} = \boldsymbol{\theta}^{(j)} - \mathbf{H}_\ell^{-1}\ell'\left(\boldsymbol{\theta}^{(j)}\right),$$

where $\ell'\left(\boldsymbol{\theta}^{(j)}\right)$ is a vector of first derivatives, i.e.,

$$\ell'\left(\boldsymbol{\theta}^{(j)}\right) = \left(\frac{\partial}{\partial\theta_1}\ell(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_k}\ell(\boldsymbol{\theta})\right),$$

and $\mathbf{H}_\ell(\boldsymbol{\theta})$ is the (Hessian) matrix of second partial derivatives of the log-likelihood, i.e.,

$$\mathbf{H}_\ell(\boldsymbol{\theta}) = l''(\boldsymbol{\theta}) = \begin{vmatrix} \frac{\partial^2}{\partial\theta_1^2}\ell(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_1\theta_k}\ell(\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\theta_k\theta_1}\ell(\boldsymbol{\theta}) & \cdots & \frac{\partial^2}{\partial\theta_k^2}\ell(\boldsymbol{\theta}) \end{vmatrix}.$$

Suppose ε is sufficiently close to zero. Then, the convergence criterion may be $\left|\theta^{(j+1)} - \theta^{(j)}\right| < \varepsilon$, $\ell'\left(\theta^{(j+1)}\right) < \varepsilon$, or $\mathcal{L}\left(\theta^{(j+1)}\right) - \mathcal{L}\left(\theta^{(j)}\right) < \varepsilon$.

EXAMPLE 9.21. Let $Y_1, \ldots, Y_n$ with

$$Y_i = \begin{cases} 0, & \text{if subject } i \text{ did not experience the event} \\ 1, & \text{if subject } i \text{ experienced the event} \end{cases}.$$

Let $X_1, \ldots, X_n$ be the fixed covariate for each subject and

$$P\left(\{Y_i = 1\}\right) = \frac{\mathrm{e}^{\theta x_i}}{1 + \mathrm{e}^{\theta x_i}}.$$

Suppose $n = 5$ and (x_i, y_i) are as shown below. Solve for the MLE of θ using the Newton-Raphson algorithm.

i	1	2	3	4	5
x_i	4.1	2.2	3.9	7.1	6.2
y_i	0	1	0	1	1

We have $Y_i \sim \text{Bernoulli}\left(p_i\right)$, so that the pmf of Y_i is given by

$$p_{Y_i}\left(y_i | p_i\right) = p_i^{y_i}\left(1 - p_i\right)^{1 - y_i}.$$

We are given that $P\left(\{Y_i = 1\}\right) = \mathrm{e}^{\theta x_i} / \left(1 + \mathrm{e}^{\theta x_i}\right)$, so we can write

$$P\left(\{Y_i = 1\}\right) = p_i^1\left(1 - p_i\right)^{1-1} = p_i = \frac{\mathrm{e}^{\theta x_i}}{1 + \mathrm{e}^{\theta x_i}}.$$

Then, we can reparameterize p_{Y_i} as

$$\begin{aligned}
p_{Y_i}\left(y_i | \theta\right) &= \left(\frac{e^{\theta x_i}}{1 + e^{\theta x_i}}\right)^{y_i}\left(1 - \frac{e^{\theta x_i}}{1 + e^{\theta x_i}}\right)^{1 - y_i} \\
&= \left(\frac{e^{\theta x_i}}{1 + e^{\theta x_i}}\right)^{y_i}\left(\frac{1 + e^{\theta x_i} - e^{\theta x_i}}{1 + e^{\theta x_i}}\right)^{1 - y_i} \\
&= \left(\frac{e^{\theta x_i}}{1 + e^{\theta x_i}}\right)^{y_i}\left(\frac{1}{1 + e^{\theta x_i}}\right)^{1 - y_i} \\
&= \left(\frac{e^{\theta x_i}}{1 + e^{\theta x_i}}\right)^{y_i}\left(\frac{1}{1 + e^{\theta x_i}}\right)\left(\frac{1}{1 + e^{\theta x_i}}\right)^{-y_i} \\
&= \frac{\left(e^{\theta x_i}\right)^{y_i}}{\left(1 + e^{\theta x_i}\right)^{y_i}}\left(1 + e^{\theta x_i}\right)^{-1}\left(1 + e^{\theta x_i}\right)^{y_i} \\
&= e^{\theta x_i y_i}\left(1 + e^{\theta x_i}\right)^{-1},
\end{aligned}$$

so that the joint pmf of $\mathbf{Y} = Y_1, \ldots, Y_n$ is given by

$$p_{\mathbf{Y}}\left(\mathbf{y} | \theta\right) = \prod_{i=1}^{n} p_{Y_i}\left(y_i | \theta\right) = \prod_{i=1}^{n} \mathrm{e}^{\theta x_i y_i}\left(1 + \mathrm{e}^{\theta x_i}\right)^{-1} = \mathrm{e}^{\sum_{i=1}^{n} \theta x_i y_i} \prod_{i=1}^{n}\left(1 + \mathrm{e}^{\theta x_i}\right)^{-1}.$$

Then, the log-likelihood is given by

$$\begin{aligned}
\ell\left(\theta\right) &= \log \mathcal{L}\left(\theta | \mathbf{y}\right) \\
&= \log\left[\mathrm{e}^{\sum_{i=1}^{n} \theta x_i y_i} \prod_{i=1}^{n}\left(1 + \mathrm{e}^{\theta x_i}\right)^{-1}\right] \\
&= \log \mathrm{e}^{\theta \sum_{i=1}^{n} x_i y_i} + \log \prod_{i=1}^{n}\left(1 + \mathrm{e}^{\theta x_i}\right)^{-1} \\
&= \theta \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{\theta x_i}\right)^{-1} \\
&= \theta \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \log\left(1 + \mathrm{e}^{\theta x_i}\right).
\end{aligned}$$

We will take the derivative with respect to θ.

$$\begin{aligned}\frac{\partial}{\partial\theta}\ell\left(\theta\right) &= \frac{\partial}{\partial\theta}\left[\theta\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}\log\left(1+\mathrm{e}^{\theta x_i}\right)\right] \\ &= \sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}\frac{1}{1+\mathrm{e}^{\theta x_i}}\cdot x_i\mathrm{e}^{\theta x_i}\end{aligned}$$

To implement Newton-Raphson algorithm, we will also need the second derivative with respect to θ.

$$\begin{aligned}\frac{\partial}{\partial\theta^2}\ell\left(\theta\right) &= \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}\frac{1}{1+e^{\theta x_i}}\cdot x_i\mathrm{e}^{\theta x_i}\right] \\ &= 0 - \sum_{i=1}^{n}\frac{\left(x_i^2\mathrm{e}^{\theta x_i}\right)\left(1+\mathrm{e}^{\theta x_i}\right) - \left(x_i\mathrm{e}^{\theta x_i}\right)\left(x_i\mathrm{e}^{\theta x_i}\right)}{\left(1+\mathrm{e}^{\theta x_i}\right)^2} \\ &= -\sum_{i=1}^{n}\frac{x_i^2\mathrm{e}^{\theta x_i} + x_i^2\mathrm{e}^{2\theta x_i} - x_i^2\mathrm{e}^{2\theta x_i}}{\left(1+\mathrm{e}^{\theta x_i}\right)^2} \\ &= -\sum_{i=1}^{n}\frac{x_i^2\mathrm{e}^{\theta x_i}}{\left(1+\mathrm{e}^{\theta x_i}\right)^2}\end{aligned}$$

Thus, we must solve

$$\begin{aligned}\hat{\theta} &= \theta^{(j)} - \frac{\ell'\left(\theta^{(j)}\right)}{\ell''\left(\theta^{(j)}\right)} \\ &= \theta^{(j)} - \frac{\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}\frac{1}{1+\mathrm{e}^{\theta x_i}}\cdot x_i\mathrm{e}^{\theta x_i}}{-\sum_{i=1}^{n}\frac{x_i^2\mathrm{e}^{\theta x_i}}{\left(1+\mathrm{e}^{\theta x_i}\right)^2}} \\ &= \theta^{(j)} + \frac{\sum_{i=1}^{n}x_iy_i - \sum_{i=1}^{n}\frac{1}{1+\mathrm{e}^{\theta x_i}}\cdot x_i\mathrm{e}^{\theta x_i}}{\sum_{i=1}^{n}\frac{x_i^2\mathrm{e}^{\theta x_i}}{\left(1+\mathrm{e}^{\theta x_i}\right)^2}}.\end{aligned}$$

We can implement the algorithm using the following R code.

```
newtraph <- function(x, y, theta, eps = 1e-6) {
        diff <- 1
        theta.hat <- theta
        while (diff > eps) {
                eta <- theta * x
                U <- sum(x * y) - sum((x * exp(eta)) / (1 + exp(eta)))
                I <- sum((x^2 * exp(eta)) / (1 + exp(eta)^2))
                theta.new <- theta + (U / I)
                diff <- abs(theta.new - theta)
                theta <- theta.new
                theta.hat <- c(theta.hat, theta)
        }
        return(theta.hat)
}

x <- c(4.1,2.2,3.9,7.1,6.2)
y <- c(0,1,0,1,1)
newtraph(x, y, 0.3)
```

```
##  [1] 0.3000000 0.1310898 0.1275242 0.1260055 0.1253472 0.1250597 0.1249338
##  [8] 0.1248786 0.1248544 0.1248437 0.1248391 0.1248370 0.1248361
```

So, we see that our estimates for $\hat{\theta}$ converge and $\hat{\theta} \approx 0.125$.

9.1.2.2. *EM algorithm.* The idea behind the EM algorithm is to iterate between taking an expectation and maximizing. Consider data $\mathbf{Y}$ whose density $f(\mathbf{y}|\theta)$ leads to a log-likelihood function that is hard to maximize. Suppose we can find another random variable Z such that $f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{z}|\theta)\,\mathrm{d}\,z$ and such that the likelihood based on $f(\mathbf{y}, \mathbf{z}|\theta)$ is easy to maximize. $\mathbf{Y}$ is the observed data, and Z is called the augmented (or latent, or missing) data, and $\mathcal{L}(\theta|\mathbf{y}, \mathbf{z}) = f(\mathbf{y}, \mathbf{z}|\theta)$ is the complete-data likelihood. Conceptually, the EM algorithm works by filling in the missing data, maximizing the expectation of the complete log-likelihood, and iterating until convergence.

The EM algorithm proceeds as follows:

(0) Pick a starting value $\theta^{(0)}$. For $j = 1, 2, \ldots$, repeat steps 1 and 2 below.
(1) **E-step:** Calculate

$$\begin{aligned} J\left(\theta|\theta^{(j)}\right) &= \mathrm{E}\left[\log \mathcal{L}(\theta|\mathbf{y}, \mathbf{z})\,|\theta^{(j)}, \mathbf{y}\right] \\ &= \mathrm{E}\left[\log f(y_1, \ldots, y_n, z_1, \ldots, z_n|\theta)\,|\theta^{(j)}, Y_1 = y_1, \ldots, Y_n = y_n\right] \end{aligned}$$

The expectation is over the missing data $Z_1, \ldots, Z_n$ treating $\theta^{(j)}$ and the observed data $\mathbf{Y}$ as fixed.
(2) **M-step:** Find $\theta^{(j+1)}$ to maximize $J\left(\theta|\theta^{(j)}\right)$, i.e., take the derivative, set it equal to zero, and solve for $\theta \to \theta^{(j+1)}$.

Example 9.22 (Mixture of normals). Often the data distribution may arise as a mixture of normal densities. Consider, for instance, heights of people arising as a mixture of men's and women's heights. The density of a mixture of two Normals is

$$f(y|\theta) = (1-p)\,\phi(y|\mu_0, \sigma_0) + p\phi(y|\mu_1, \sigma_1)$$

where $\phi(y|\mu, \sigma)$ denotes a normal density with mean μ and standard deviation σ. The idea is that an observation is drawn from the first normal with probability p and the second with probability $1-p$. However, we don't know from which Normal it was drawn. The parameters are $\boldsymbol{\theta} = (\mu_0, \sigma_0, \mu_1, \sigma_1, p)$.

The incomplete-data likelihood is given by

$$\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} \left[(1-p)\,\phi(y_i|\mu_0, \sigma_0) + p\phi(y_i|\mu_1, \sigma_1)\right],$$

so that the log-likelihood is given by

$$\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \log\left[(1-p)\,\phi(y_i|\mu_0, \sigma_0) + p\phi(y_i|\mu_1, \sigma_1)\right].$$

The sum inside the logarithm makes this difficult to maximize. Instead, we will define an indicator variable Z_i as

$$Z_i = \begin{cases} 0, & \text{if } Y_i \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right) \\ 1, & \text{if } Y_i \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right) \end{cases}.$$

Then, we can write the joint density of $\mathbf{Y}$ conditioned on $\mathbf{Z}$ as

$$f_{\mathbf{Y}}(\mathbf{y}|\theta, z_i = j) = \phi(y_i|\mu_j, \sigma_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(y_i-\mu_j)^2/\left(2\sigma_j^2\right)}.$$

Recall that the probability of the intersection of two events A and B is equal to the probability of A given B multiplied by the probability of B, i.e.,

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

Then, abusing notation slightly, we have

$$P(\{Y_i = y_i\} \cap \{Z_i = z_i\}) = P(\{Y_i = y_i|Z_i = z_i\})\,P(\{Z_i = z_i\}),$$

i.e., the joint density of $\mathbf{Y}$ and $\mathbf{Z}$ is given by

$$f_{\mathbf{Y},\mathbf{Z}}(y_i, z_i = j) = f_{\mathbf{Y}}(y_i|z_i = j, \theta)\,P(\{Z_i = j\}),$$

where $P(\{Z_i = 1\}) = p$ and $P(\{Z_i = 0\}) = 1 - p$. Then, the complete-data likelihood is given by

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) &= \prod_{i=1}^{n} f_{\mathbf{Y},\mathbf{Z}}(y_i, z_i|\boldsymbol{\theta}) \\
&= \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-(y_i-\mu_j)^2/\left(2\sigma_j^2\right)} \cdot P(\{Z_i = j\}) \right] \\
&= \prod_{i=1}^{n} \left[\left[\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(y_i-\mu_0)^2/\left(2\sigma_0^2\right)} \cdot (1-p) \right]^{1-z_i} \cdot \left[\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(y_i-\mu_1)^2/\left(2\sigma_1^2\right)} \cdot p \right]^{z_i} \right],
\end{aligned}$$

so that the log-likelihood is given by

$$\begin{aligned}
\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) &= \log \prod_{i=1}^{n} \left[\left[\frac{1}{\sqrt{2\pi\sigma_0^2}} e^{-(y_i-\mu_0)^2/\left(2\sigma_0^2\right)} \cdot (1-p) \right]^{1-z_i} \cdot \left[\frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-(y_i-\mu_1)^2/\left(2\sigma_1^2\right)} \cdot p \right]^{z_i} \right] \\
&= \sum_{i=1}^{n} \left\{ (1-z_i) \log \left[\frac{1-p}{\sqrt{2\pi\sigma_0^2}} e^{-(y_i-\mu_0)^2/\left(2\sigma_0^2\right)} \right] + z_i \log \left[\frac{p}{\sqrt{2\pi\sigma_1^2}} e^{-(y_i-\mu_1)^2/\left(2\sigma_1^2\right)} \right] \right\}.
\end{aligned}$$

We will now find the expected value of the complete-data log likelihood (the E-step).

$$\begin{aligned}
J\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}\right) &= \mathrm{E}\left[\log \mathcal{L}(\boldsymbol{\theta}|\mathbf{y}, \mathbf{z}) \,|\mathbf{y}, \boldsymbol{\theta}^{(j)} \right] \\
&= \sum_{i=1}^{n} \left\{ \mathrm{E}\left[1 - z_i|\mathbf{y}, \boldsymbol{\theta}^{(j)} \right] \log \left[\frac{1-p}{\sqrt{2\pi\sigma_0^2}} e^{-(y_i-\mu_0)^2/\left(2\sigma_0^2\right)} \right] + \mathrm{E}\left[z_i|\mathbf{y}, \boldsymbol{\theta}^{(j)} \right] \log \left[\frac{p}{\sqrt{2\pi\sigma_1^2}} e^{-(y_i-\mu_1)^2/\left(2\sigma_1^2\right)} \right] \right\}
\end{aligned}$$

We will evaluate $\mathrm{E}\left[z_i|\mathbf{y}, \boldsymbol{\theta}^{(j)} \right] = P\left(\left\{ z_i = 1|y_i, \boldsymbol{\theta}^{(j)} \right\} \right)$ using Bayes' Rule.

$$\begin{aligned}
\mathrm{E}\left[z_i|\mathbf{y}, \boldsymbol{\theta}^{(j)} \right] &= P\left(\left\{ z_i = 1|y_i, \boldsymbol{\theta}^{(j)} \right\} \right) \\
&= \frac{f\left(y_i|z_i = 1, \boldsymbol{\theta}^{(j)} \right) P(\{Z_i = 1\})}{\sum_{k=0}^{1} f\left(y_i|Z_i = k, \boldsymbol{\theta}^{(j)} \right) P(\{Z_i = k\})} \\
&= \frac{\phi\left(y_i|\mu_1^{(j)}, \sigma_1^{(j)} \right) \cdot p^{(j)}}{p^{(j)} \phi\left(y_i|\mu_1^{(j)}, \sigma_1^{(j)} \right) + \left(1 - p^{(j)}\right) \phi\left(y_i|\mu_0^{(j)}, \sigma_0^{(j)} \right)} \\
&= \tau_i^{(j)}
\end{aligned}$$

Then, we have

$$\begin{aligned}
J\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}\right) &= \sum_{i=1}^{n} \left\{ \left(1 - \tau_i^{(j)}\right) \left[\log \left(\frac{1-p}{\sqrt{2\pi\sigma_0^2}} \right) + \log \left(e^{-(y_i-\mu_0)^2/\left(2\sigma_0^2\right)} \right) \right] \right. \\
&\qquad \left. + \tau_i^{(j)} \left[\log \frac{p}{\sqrt{2\pi\sigma_1^2}} + \log \left(e^{-(y_i-\mu_1)^2/\left(2\sigma_1^2\right)} \right) \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ \left(1 - \tau_i^{(j)}\right) \left[\log(1-p) - \log\left(2\pi\sigma_0^2\right)^{1/2} - \frac{(y_i-\mu_0)^2}{2\sigma_0^2} \right] + \tau_i^{(j)} \left[\log p - \log\left(2\pi\sigma_1^2\right)^{1/2} - \frac{(y_i-\mu_1)^2}{2\sigma_1^2} \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ \left(1 - \tau_i^{(j)}\right) \left[\log(1-p) - \frac{1}{2}\log 2\pi\sigma_0^2 - \frac{(y_i-\mu_0)^2}{2\sigma_0^2} \right] + \tau_i^{(j)} \left[\log p - \frac{1}{2}\log 2\pi\sigma_1^2 - \frac{(y_i-\mu_1)^2}{2\sigma_1^2} \right] \right\}.
\end{aligned}$$

We will take the derivative with respect to p (the M-step).

$$\frac{\partial}{\partial p} J\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{(j)}\right) = \sum_{i=1}^{n} \left\{ -\left(1 - \tau_i^{(j)}\right) \frac{1}{1-p} + \tau_i^{(j)} \frac{1}{p} \right\}$$

(Although p is incorporated in $\tau_i^{(j)}$, it is considered to have been evaluated there, so that $\tau_i^{(j)}$ is treated as constant when taking the deriative of J with respect to p.) Setting this equal to zero, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \frac{\left(1-\tau_i^{(j)}\right)}{1-p} = \sum_{i=1}^{n} \frac{\tau_i^{(j)}}{p} \\
\Leftrightarrow\ &\frac{1}{1-p}\sum_{i=1}^{n}\left(1-\tau_i^{(j)}\right) = \frac{1}{p}\sum_{i=1}^{n}\tau_i^{(j)} \\
\Leftrightarrow\ &(1-p)\sum_{i=1}^{n}\tau_i^{(j)} = p\sum_{i=1}^{n}\left(1-\tau_i^{(j)}\right) \\
\Leftrightarrow\ &\frac{1-p}{p} = \frac{\sum_{i=1}^{n}\left(1-\tau_i^{(j)}\right)}{\sum_{i=1}^{n}\tau_i^{(j)}} \\
\Leftrightarrow\ &\frac{1}{p}-1 = \frac{n-\sum_{i=1}^{n}\tau_i^{(j)}}{\sum_{i=1}^{n}\tau_i^{(j)}} \\
\Leftrightarrow\ &\frac{1}{p}-1 = \frac{n}{\sum_{i=1}^{n}\tau_i^{(j)}}-1 \\
\Leftrightarrow\ &\frac{1}{p} = \frac{n}{\sum_{i=1}^{n}\tau_i^{(j)}} \\
\Leftrightarrow\ &p^{(j+1)} = \frac{\sum_{i=1}^{n}\tau_i^{(j)}}{n}.
\end{aligned}$$

We can proceed similarly to find the other parameter estimates, i.e., by taking the derivative of J with respect to each parameter, setting the result equal to zero, and solving for the parameter.

We have introduced four methods for finding MLEs. In order of approximate difficulty, the methods are

(1) Set the derivative of the log-likelihood equal to zero and solve or $\hat{\theta}$. This is the simplest method when a closed-form solution is possible.
(2) When the support depends on the parameter θ, plot the likelihood function and determine where it attains its maximum.
(3) Use the Newton-Raphson method to estimate the MLE.
(4) Use the EM method to find the expectation of the complete log-likelihood using the observed $\mathbf{y}$ and an initial estimate for $\theta^{(j)}$.
   (a) Introduce a latent variable Z.
   (b) Find the complete log-likelihood $L(\theta|\mathbf{y},\mathbf{z})$.
   (c) Use the E-step to estimate $z^{(j)}$.
   (d) Use the M-step to maximize $J\left(\theta|\theta^{(j)}\right)$.

9.1.3. Bayes estimators. In the classical (frequentist) approach, θ is considered to be unknown, but fixed. in the Bayesian approach, θ is treated as a random variable. A prior distribution for θ, $\pi(\theta)$, is specified. This may be non-informative or may describe expert opinion (or subjective belief). Inference is based on the posterior distribution,

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\,\pi(\theta)}{\int f(\mathbf{x}|\theta)\,\pi(\theta)\,\mathrm{d}\,\theta},$$

which is obtained from Bayes' rule and updates the prior, $\pi(\theta)$, given the observed data. The expectation of $\pi(\theta|\mathbf{x})$, which is given by

$$\mathrm{E}\left[\pi(\theta|\mathbf{x})\right] = \int_{\Theta} \theta\cdot\pi(\theta|\mathbf{x})\,\mathrm{d}\,\theta,$$

where Θ is the support of θ, i.e., the parameter space, is called the posterior mean or the *Bayes estimator*. One could also use the most probable value of θ, the *posterior mode*, as a point estimator.

EXAMPLE 9.23. Let $X_1,\ldots,X_n \sim \text{Bernoulli}(\theta)$ and consider $\theta \sim \mathcal{U}(0,1)$.

(1) Find the Bayes estimator for θ.

The prior distribution is uniform on the interval $(0,1)$, so its density is given by

$$\pi(\theta) = \frac{1}{1-0} = 1.$$

From example 8.13, the joint pmf of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$p_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i},$$

so that the posterior distribution is given by

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &= \frac{p_{\mathbf{X}}(\mathbf{x}|\theta)\,\pi(\theta)}{\int p_{\mathbf{X}}(\mathbf{x}|\theta)\,\pi(\theta)\,\mathrm{d}\theta} \\
&= \frac{\left[\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}\right] \cdot 1}{\int_0^1 \left[\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}\right] \cdot 1 \mathrm{d}\theta} \\
&= \frac{\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}}{\int_0^1 \theta^{\sum_{i=1}^n x_i} (1-\theta)^{n-\sum_{i=1}^n x_i}\,\mathrm{d}\theta}.
\end{aligned}$$

Let $\alpha = \sum_{i=1}^n x_i + 1$ and let $\beta = n - \sum_{i=1}^n x_i + 1$, so that we have

$$\pi(\theta|\mathbf{x}) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta}.$$

We recognize the integrand as the kernel of a Beta (α, β) distribution, so we write

$$\pi(\theta|\mathbf{x}) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)\int_0^1 \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\,\mathrm{d}\theta},$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha+\beta)$. Then, the integrand is the pdf of a Beta (α,β) distribution, which integrates to 1, so that we have

$$\begin{aligned}
\pi(\theta|\mathbf{x}) &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)\cdot 1} \\
&= \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},
\end{aligned}$$

which is the pdf of a Beta (α,β) random variable. It follows that

$$\pi(\theta|\mathbf{x}) \sim \text{Beta}\left(\sum_{i=1}^n x_i + 1, n - \sum_{i=1}^n x_i + 1\right).$$

The expected value (mean) of a Beta (γ,ψ) random variable is given by $\gamma/(\gamma+\psi)$, so it follows that

$$\begin{aligned}
\mathrm{E}[\pi(\theta|\mathbf{x})] &= \frac{\alpha}{\alpha+\beta} \\
&= \frac{\sum_{i=1}^n x_i + 1}{(\sum_{i=1}^n x_i + 1) + (n - \sum_{i=1}^n x_i + 1)} \\
&= \frac{\sum_{i=1}^n x_i + 1}{n+2}
\end{aligned}$$

is the Bayes estimator (posterior mean) for θ.

(2) Show that the Bayes estimator is a weighted average of the MLE and the prior mean.

The log-likelihood is given by

$$\begin{aligned}
\log L(\theta|\mathbf{x}) &= \log\left[\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}\right] \\
&= \log\theta^{\sum_{i=1}^n x_i} + \log(1-\theta)^{n-\sum_{i=1}^n x_i}
\end{aligned}$$

$$= \left(\sum_{i=1}^{n} x_i\right) \log\theta + \left(n - \sum_{i=1}^{n} x_i\right) \log\left(1-\theta\right).$$

We will take the derivative with respect to θ.

$$\begin{aligned}
\frac{\partial}{\partial\theta} \log L\left(\theta|\mathbf{x}\right) &= \frac{\partial}{\partial\theta}\left[\left(\sum_{i=1}^{n} x_i\right) \log\theta + \left(n - \sum_{i=1}^{n} x_i\right) \log\left(1-\theta\right)\right] \\
&= \frac{1}{\theta}\sum_{i=1}^{n} x_i + \left(n - \sum_{i=1}^{n} x_i\right) \frac{1}{1-\theta} \cdot -1 \\
&= \frac{1}{\theta}\sum_{i=1}^{n} x_i - \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} x_i\right)
\end{aligned}$$

Setting this equal to zero, we have

$$\begin{aligned}
\frac{1}{\theta}\sum_{i=1}^{n} x_i &= \frac{1}{1-\theta}\left(n - \sum_{i=1}^{n} x_i\right) \\
\Leftrightarrow \theta\left(n - \sum_{i=1}^{n} x_i\right) &= (1-\theta)\sum_{i=1}^{n} x_i \\
\Leftrightarrow \theta\left(n - \sum_{i=1}^{n} x_i\right) &= \sum_{i=1}^{n} x_i - \theta\sum_{i=1}^{n} x_i \\
\Leftrightarrow \theta\left(n - \sum_{i=1}^{n} x_i\right) + \theta\sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} x_i \\
\Leftrightarrow \theta\left[\left(n - \sum_{i=1}^{n} x_i\right) + \sum_{i=1}^{n} x_i\right] &= \sum_{i=1}^{n} x_i \\
\Leftrightarrow \theta n &= \sum_{i=1}^{n} x_i \\
\Leftrightarrow \hat{\theta} &= \bar{x}.
\end{aligned}$$

The prior mean is given by

$$\mathrm{E}\left[\theta\right] = \frac{0+1}{2} = \frac{1}{2}.$$

Suppose that a is the weight of the MLE and $1-a$ is the weight of the prior mean, so that we have

$$\begin{aligned}
a\bar{x} + (1-a)\,\frac{1}{2} &= \frac{n\bar{x}+1}{n+2} \\
&= \frac{n}{n+2}\bar{x} + \frac{1}{n+2} \\
&= \frac{n}{n+2}\bar{x} + \left(\frac{1}{n+2}\right)\left(\frac{2}{2}\right) \\
&= \frac{n}{n+2}\bar{x} + \left(\frac{2+n-n}{n+2}\right)\frac{1}{2} \\
&= \frac{n}{n+2}\bar{x} + \left(1 - \frac{n}{n+2}\right)\frac{1}{2}
\end{aligned}$$

Then, it follows that

$$a = \frac{n}{n+2} \quad \text{and} \quad 1-a = \frac{2}{n+2}.$$

EXAMPLE 9.24 (Poisson Bayes estimator). Suppose $X_1, \ldots, X_n$ are iid Poisson (λ), i.e.,

$$f\left(\mathbf{x}|\lambda\right) = \frac{\lambda^{\sum_{i=1}^{n} x_i} \mathrm{e}^{-n\lambda}}{\prod_{i=1}^{n} x_i!}.$$

The investigator decides to use a prior distribution that captures his prior opinion using a gamma density (conjugate prior for Poisson) with mean of 15 and standard deviation of 5. Find the posterior mean (Bayes estimator) of λ.

The posterior distribution is given by

$$\begin{aligned}
\pi(\lambda|\mathbf{x}) &= \frac{f(\mathbf{x}|\lambda)\,\pi(\lambda)}{\int f(\mathbf{x}|\lambda)\,\pi(\lambda)\,\mathrm{d}\,\lambda} \\
&= \frac{\left[\frac{\lambda^{\sum_{i=1}^n x_i}\mathrm{e}^{-n\lambda}}{\prod_{i=1}^n x_i!}\right]\left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}\mathrm{e}^{-\lambda/\beta}\right]}{\int_0^\infty \left[\frac{\lambda^{\sum_{i=1}^n x_i}\mathrm{e}^{-n\lambda}}{\prod_{i=1}^n x_i!}\right]\left[\frac{1}{\Gamma(\alpha)\beta^\alpha}\lambda^{\alpha-1}\mathrm{e}^{-\lambda/\beta}\right]\mathrm{d}\,\lambda} \\
&= \frac{\lambda^{\sum_{i=1}^n x_i+\alpha-1}\mathrm{e}^{-n\lambda-\lambda/\beta}}{\int_0^\infty \lambda^{\sum_{i=1}^n x_i+\alpha-1}\mathrm{e}^{-n\lambda-\lambda/\beta}\,\mathrm{d}\,\lambda} \\
&= \frac{\lambda^{\sum_{i=1}^n x_i+\alpha-1}\mathrm{e}^{-\lambda(n+1/\beta)}}{\int_0^\infty \lambda^{\sum_{i=1}^n x_i+\alpha-1}\mathrm{e}^{-\lambda(n+1/\beta)}\,\mathrm{d}\,\lambda} \\
&= \frac{\lambda^{\sum_{i=1}^n x_i+\alpha-1}\mathrm{e}^{-\lambda/(n+1/\beta)^{-1}}}{\int_0^\infty \lambda^{\sum_{i=1}^n x_i+\alpha-1}\mathrm{e}^{-\lambda/(n+1/\beta)^{-1}}\,\mathrm{d}\,\lambda}
\end{aligned}$$

Let $\alpha' = \sum_{i=1}^n x_i + \alpha$ and let $\beta' = 1/(n+1/\beta)$, so that we have

$$\pi(\lambda|\mathbf{x}) = \frac{\lambda^{\alpha'-1}\mathrm{e}^{-\lambda/\beta'}}{\int_0^\infty \lambda^{\alpha'-1}\mathrm{e}^{-\lambda/\beta'}\,\mathrm{d}\,\lambda}.$$

We recognize the integrand as the kernel of a $\Gamma(\alpha',\beta')$ distribution, so we write

$$\pi(\lambda|\mathbf{x}) = \frac{\lambda^{\alpha'-1}\mathrm{e}^{-\lambda/\beta'}}{\Gamma(\alpha')\,\beta'^{\alpha'}\int_0^\infty \frac{1}{\Gamma(\alpha')\beta'^{\alpha'}}\lambda^{\alpha'-1}\mathrm{e}^{-\lambda/\beta'}\,\mathrm{d}\,\lambda}.$$

Then, the integrand is the pdf of a $\Gamma(\alpha',\beta')$ distribution, which integrates to 1, so that we have

$$\pi(\lambda|\mathbf{x}) = \frac{\lambda^{\alpha'-1}\mathrm{e}^{-\lambda/\beta'}}{\Gamma(\alpha')\,\beta'^{\alpha'}\cdot 1} = \frac{1}{\Gamma(\alpha')\,\beta'^{\alpha'}}\lambda^{\alpha'-1}\mathrm{e}^{-\lambda/\beta'},$$

which is the pdf of a Gamma(α',β') random variable, i.e., $\pi(\lambda|\mathbf{x}) \sim \text{Gamma}\left(\sum_{i=1}^n x_i + \alpha, 1/(n+1/\beta)\right)$. The expected value (mean) of a Gamma(γ,ψ) random variable is given by $\gamma\psi$, so it follows that

$$\mathrm{E}[\pi(\lambda|\mathbf{x})] = \left(\sum_{i=1}^n x_i + \alpha\right)\left(\frac{1}{n+1/\beta}\right) = \frac{\sum_{i=1}^n x_i + \alpha}{n+1/\beta}$$

is the Bayes estimator of λ.

DEFINITION 9.25. Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$ (indexed by θ). A class Π of prior distributions is a *conjugate family* for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π, and all $x \in \mathcal{X}$, i.e., if

$$\pi(\theta) \in \Pi \implies f(\theta|x) \in \Pi.$$

Sampling models from exponential families all have conjugate priors, including the following examples.

Sampling model	Prior & Posterior
Binomial(n,θ)	Beta
Poisson(θ)	Gamma
Exponential(θ)	Gamma
Gamma(α,θ) (α known)	Gamma
Normal(θ,σ^2) (σ^2 known)	Normal

Without a conjugate prior, it is not usually possible to obtain a closed-form solution for a posterior distribution.

Contemporary computational resources have had an enormous impact on Bayesian inference. The computationally difficult part of Bayesian inference is the calculation of the normalizing constant that makes the posterior density integrate to 1. Traditionally, such calculations were performed analytically, often using conjugate priors so that the integrations could be done explicitly. For more complex problems, Markov chain Monte Carlo methods are used to sample from the posterior distribution.

9.2. Methods of evaluating estimators

In the previous section, we discussed techniques for finding point estimators of parameters. Since we can usually apply more than one of these methods, we are often faced with the task of choosing between estimators. What qualities should a "good" estimator have? Is it possible to find a "best" $\hat{\theta}$?

DEFINITION 9.26. The *bias* of a point estimator W of a parameter θ is the difference between the expected value of W and θ; that is, $\text{Bias}_\theta W = \text{E}_\theta[W] - \theta$. An estimator whose bias is identically (in θ) equal to zero is called *unbiased* and satisfied $\text{E}_\theta[W] = \theta$ for all θ.

EXAMPLE 9.27. Let $Y_1, \ldots, Y_n$ be a random sample from the pdf

$$f(y|\theta) = \frac{2y}{\theta^2}, \quad 0 \leq y \leq \theta.$$

(1) Find the moment estimator and the MLE of θ.
The first moment of Y_i is given by

$$\begin{aligned}
\text{E}[Y_i] &= \int_0^\theta y \cdot f_{Y_i}(y|\theta)\,\text{d}y \\
&= \int_0^\theta y \cdot \frac{2y}{\theta^2}\text{d}y \\
&= \frac{2}{3\theta^2}y^3\Big|_0^\theta \\
&= \frac{2\theta^3}{3\theta^2} - 0 \\
&= \frac{2}{3}\theta,
\end{aligned}$$

so that we have

$$\begin{aligned}
\hat{\mu}_1 &= \frac{\sum_{i=1}^n y_i}{n} \\
\Leftrightarrow \frac{2}{3}\hat{\theta} &= \bar{y} \\
\Leftrightarrow \hat{\theta}_{MOM} &= \frac{3}{2}\bar{y}.
\end{aligned}$$

The support depends on the parameter θ, so we can write the pdf as

$$f_Y(y|\theta) = \frac{2y}{\theta^2} I_{\{0 \leq y \leq \theta\}}.$$

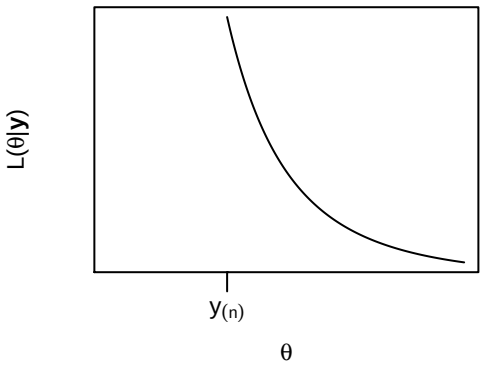
Then, the likelihood function is given by

$$\begin{aligned}
L(\theta|\mathbf{y}) &= \prod_{i=1}^n f_{Y_i}(y_i|\theta) \\
&= \prod_{i=1}^n \frac{2y_i}{\theta^2} I_{\{0 \leq y_i \leq \theta\}} \\
&= \frac{2^n \prod_{i=1}^n y_i}{\theta^{2n}} I_{\left\{0 \leq y_{(n)} \leq \theta\right\}},
\end{aligned}$$

which we can write as

$$L(\theta|\mathbf{y}) = \begin{cases} \dfrac{2^n \prod_{i=1}^n y_i}{\theta^{2n}}, & \theta \geq y_{(n)} \\ 0, & \theta < y_{(n)} \end{cases},$$

and whose graph is shown below.



Clearly, the likelihood is maximized at $Y_{(n)}$, so it follows that $\hat{\theta}_{MLE} = Y_{(n)}$.

(2) Is either estimator unbiased?

The expected value of $\hat{\theta}_{MOM}$ is given by

$$\begin{aligned}
\mathrm{E}\left[\hat{\theta}_{MOM}\right] &= \mathrm{E}\left[\frac{3}{2}\bar{Y}\right] \\
&= \frac{3}{2}\mathrm{E}\left[\bar{Y}\right] \\
&= \frac{3}{2}\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \\
&= \frac{3}{2n}\sum_{i=1}^{n}\mathrm{E}\left[Y_i\right] \\
&= \frac{3}{2n}\sum_{i=1}^{n}\frac{2}{3}\theta \\
&= \frac{3}{2n}\left(n\cdot\frac{2}{3}\theta\right) \\
&= \theta,
\end{aligned}$$

so it follows that $\hat{\theta}_{MOM}$ is unbiased for θ. The cdf of Y is given by

$$F_Y(y) = \int_0^y f_Y(y|\theta)\,\mathrm{d}\,t = \int_0^y \frac{2t}{\theta^2}\,\mathrm{d}\,t = \frac{t^2}{\theta^2}\Big|_0^y = \frac{y^2}{\theta^2} - 0 = \frac{y^2}{\theta^2},$$

so that the pdf of $\hat{\theta}_{MLE} = Y_{(n)}$ is given by

$$f_{Y_{(n)}}(y|\theta) = n\left[F_y(y)\right]^{n-1} f_Y(y) = n\left(\frac{y^2}{\theta^2}\right)^{n-1}\left(\frac{2y}{\theta^2}\right) = n\frac{y^{2(n-1)}}{\theta^{2(n-1)}}\left(\frac{2y}{\theta^2}\right) = \frac{2ny^{2n-1}}{\theta^{2n}}.$$

Then, the expected value of $\hat{\theta}_{MLE}$ is given by

$$\begin{aligned}
\mathrm{E}\left[\hat{\theta}_{MLE}\right] &= \mathrm{E}\left[Y_{(n)}\right] \\
&= \int_0^\theta y\cdot f_{Y_{(n)}}(y)\,\mathrm{d}y \\
&= \int_0^\theta y\cdot\frac{2ny^{2n-1}}{\theta^{2n}}\mathrm{d}y \\
&= \frac{2n}{\theta^{2n}}\int_0^\theta y^{2n}\mathrm{d}y \\
&= \frac{2n}{\theta^{2n}}\left[\frac{1}{2n+1}y^{2n+1}\right|_0^\theta
\end{aligned}$$

$$= \frac{2n}{\theta^{2n}}\left(\frac{\theta^{2n+1}}{2n+1} - 0\right)$$
$$= \frac{2n}{2n+1}\theta.$$

We see that $\mathrm{E}\left[\hat{\theta}_{MLE}\right]$ is not equal to θ, so it follows that $\hat{\theta}_{MLE}$ is a biased estimator for θ.

(3) Construct an unbiased estimator based on the MLE.

An unbiased estimator $\tilde{\lambda}$ for λ satisfies the condition that $\mathrm{E}\left[\tilde{\lambda}\right] = \lambda$ for all λ. So, to construct an unbiased estimator $\tilde{\theta}$ based on the MLE $\hat{\theta}_{MLE}$, we need to find a constant such that $\mathrm{E}\left[\hat{\theta}_{MLE}\right] = \theta$. We see that

$$\tilde{\theta} = \frac{2n+1}{2n}\hat{\theta}_{MLE} = \frac{2n+1}{2n}Y_{(n)}$$

satisfies the condition that

$$\mathrm{E}\left[\tilde{\theta}\right] = \mathrm{E}\left[\frac{2n+1}{2n}Y_{(n)}\right] = \frac{2n+1}{2n}\mathrm{E}\left[Y_{(n)}\right] = \frac{2n+1}{2n}\left(\frac{2n}{2n+1}\theta\right) = \theta,$$

so it follows that $\tilde{\theta}$ is an unbiased estimator for θ based on the MLE.

The precision of an estimator is another important property beside unbiasedness.

DEFINITION 9.28. Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be two unbiased estimators for a parameter θ. If $\mathrm{Var}\left(\hat{\theta}_1\right) < \mathrm{Var}\left(\hat{\theta}_2\right)$, we say that $\hat{\theta}_1$ is *more efficient* than $\hat{\theta}_2$. The relative efficiency of $\hat{\theta}_1$ with respect to $\hat{\theta}_2$ is the ratio $\mathrm{Var}\left(\hat{\theta}_2\right) / \mathrm{Var}\left(\hat{\theta}_1\right)$.

EXAMPLE 9.29. Consider the two unbiased estimators from the example above. Which estimator is more efficient?

The variance of $\hat{\theta}_{MOM}$ is given by

$$\mathrm{Var}\left(\hat{\theta}_{MOM}\right) = \mathrm{Var}\left(\frac{3}{2}\bar{Y}\right) = \mathrm{Var}\left(\frac{3}{2n}\sum_{i=1}^{n} Y_i\right) = \frac{9}{4n^2}\mathrm{Var}\left(\sum_{i=1}^{n} Y_i\right) = \frac{9}{4n^2}\sum_{i=1}^{n}\mathrm{Var}\left(Y_i\right),$$

where the final equality follows from theorem 4.7, i.e., the variance of the sum of independent random variables is equal to the sum of their variances. To find the variance of Y_i, we will need to find its second moment.

$$\mathrm{E}\left[Y^2\right] = \int_0^\theta y^2 \cdot f_Y(y)\,\mathrm{d}\,y = \int_0^\theta y^2\frac{2y}{\theta^2}\,\mathrm{d}\,y = \frac{1}{2\theta^2}y^4\Big|_0^\theta = \frac{\theta^4}{2\theta^2} - 0 = \frac{\theta^2}{2}$$

Then, we have

$$\begin{aligned}
\mathrm{Var}\left(\hat{\theta}_{MOM}\right) &= \frac{9}{4n^2}\sum_{i=1}^{n}\mathrm{Var}\left(Y_i\right) \\
&= \frac{9}{4n^2}\sum_{i=1}^{n}\left[\mathrm{E}\left[Y^2\right] - \left(\mathrm{E}\left[Y\right]\right)^2\right] \\
&= \frac{9}{4n^2}\sum_{i=1}^{n}\left[\frac{\theta^2}{2} - \left(\frac{2}{3}\theta\right)^2\right] \\
&= \frac{9}{4n^2}\sum_{i=1}^{n}\left[\frac{\theta^2}{2} - \frac{4\theta^2}{9}\right] \\
&= \frac{9}{4n^2}\sum_{i=1}^{n}\left[\frac{9\theta^2}{18} - \frac{8\theta^2}{18}\right] \\
&= \frac{9}{4n^2}\sum_{i=1}^{n}\frac{\theta^2}{18}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{9}{4n^2} \cdot n\frac{\theta^2}{18} \\
&= \frac{\theta^2}{8n}.
\end{aligned}
$$

The variance of $\tilde{\theta}$ is given by

$$
\operatorname{Var}\left(\tilde{\theta}\right) = \operatorname{Var}\left(\frac{2n+1}{2n}\hat{\theta}_{MLE}\right) = \left(\frac{2n+1}{2n}\right)^2 \operatorname{Var}\left(\hat{\theta}_{MLE}\right) = \frac{(2n+1)^2}{4n^2} \operatorname{Var}\left(\hat{\theta}_{MLE}\right).
$$

We will now find the second moment of $\hat{\theta}_{MLE} = Y_{(n)}$.

$$
\begin{aligned}
\mathrm{E}\left[Y_{(n)}^2\right] &= \int_0^\theta y^2 \cdot f_{Y_{(n)}}(y)\,\mathrm{d}\,y \\
&= \int_0^\theta y^2 \cdot \frac{2ny^{2n-1}}{\theta^{2n}}\,\mathrm{d}\,y \\
&= \frac{2n}{\theta^{2n}} \int_0^\theta y^{2n+1}\,\mathrm{d}\,y \\
&= \frac{2n}{\theta^{2n}} \left[\frac{1}{2n+2} y^{2n+2}\right|_0^\theta \\
&= \frac{2n}{\theta^{2n}} \left(\frac{\theta^{2n+2}}{2n+2} - 0\right) \\
&= \frac{2n\theta^2}{2n+2} \\
&= \frac{n\theta^2}{n+1}
\end{aligned}
$$

Then, the variance of $\hat{\theta}_{MLE}$ is given by

$$
\begin{aligned}
\operatorname{Var}\left(\hat{\theta}_{MLE}\right) &= \operatorname{Var}\left(Y_{(n)}\right) \\
&= \mathrm{E}\left[Y_{(n)}^2\right] - \left(\mathrm{E}\left[Y_{(n)}\right]\right)^2 \\
&= \frac{n\theta^2}{n+1} - \left(\frac{2n\theta}{2n+1}\right)^2 \\
&= \frac{n\theta^2}{n+1} - \frac{4n^2\theta^2}{(2n+1)^2} \\
&= \frac{\theta^2\left[n(2n+1)^2 - 4n^2(n+1)\right]}{(n+1)(2n+1)^2} \\
&= \frac{\theta^2\left[n\left(4n^2+4n+1\right) - 4n^3 - 4n^2\right]}{(n+1)(2n+1)^2} \\
&= \frac{\theta^2\left[4n^3+4n^2+n-4n^3-4n^2\right]}{(n+1)(2n+1)^2} \\
&= \frac{n\theta^2}{(n+1)(2n+1)^2},
\end{aligned}
$$

so that we have

$$
\begin{aligned}
\operatorname{Var}\left(\tilde{\theta}\right) &= \frac{(2n+1)^2}{4n^2} \operatorname{Var}\left(\hat{\theta}_{MLE}\right) \\
&= \frac{(2n+1)^2}{4n^2} \left(\frac{n\theta^2}{(n+1)(2n+1)^2}\right)
\end{aligned}
$$

$$= \frac{\theta^2}{4n\,(n+1)}.$$

If $\operatorname{Var}\left(\hat{\theta}_{MOM}\right) < \operatorname{Var}\left(\tilde{\theta}\right)$, then $\hat{\theta}_{MOM}$ is more efficient than $\tilde{\theta}$.

$$\begin{aligned}
\operatorname{Var}\left(\hat{\theta}_{MOM}\right) &< \operatorname{Var}\left(\tilde{\theta}\right) \\
\Leftrightarrow \frac{\theta^2}{8n} &< \frac{\theta^2}{4n\,(n+1)} \\
\Leftrightarrow 8n &> 4n\,(n+1) \\
\Leftrightarrow 8n - 4n\,(n+1) &> 0 \\
\Leftrightarrow 4n\,(2-(n+1)) &> 0 \\
\Leftrightarrow 4n\,(2-n-1) &> 0 \\
\Leftrightarrow 4n\,(1-n) &> 0 \\
\Leftrightarrow 4n - 4n^2 &> 0
\end{aligned}$$

We have $n > 0$, so it follows that $4n - 4n^2 < 0$, so it cannot be the case that $\hat{\theta}_{MOM}$ is more efficient than $\tilde{\theta}$. Therefore, $\tilde{\theta}$ is more efficient than $\hat{\theta}_{MOM}$. The relative efficiency of $\tilde{\theta}$ with respect to $\hat{\theta}_{MOM}$ is given by

$$\frac{\operatorname{Var}\left(\hat{\theta}_{MOM}\right)}{\operatorname{Var}\left(\tilde{\theta}\right)} = \frac{\theta^2/\,(8n)}{\theta^2/\,(4n\,(n+1))} = \frac{4n\,(n+1)}{8n} = \frac{n+1}{2}.$$

9.2.1. Mean squared error.

DEFINITION 9.30. The *mean squared error* (MSE) of an estimator W of a parameter θ is the function of θ defined by $\mathrm{E}_\theta\left[(W-\theta)^2\right]$.

Any increasing function of $|W-\theta|$ would serve to measure the goodness of an estimator, but MSE has at least two advantages over other distance measures. First, it is analytically tractable (absolute value would lead to discontinuities). Second, it has the interpretation

$$\begin{aligned}
\mathrm{E}\left[(W-\theta)^2\right] &= \mathrm{E}\left[(W-\mathrm{E}[W]+\mathrm{E}[W]-\theta)^2\right] \\
&= \mathrm{E}\left[(W-\mathrm{E}[W])^2 + 2\,(W-\mathrm{E}[W])\,(\mathrm{E}[W]-\theta) + (\mathrm{E}[W]-\theta)^2\right] \\
&= \mathrm{E}\left[(W-\mathrm{E}[W])^2\right] + 2\,\mathrm{E}\left[(W-\mathrm{E}[W])\,(\mathrm{E}[W]-\theta)\right] + \mathrm{E}\left[(\mathrm{E}[W]-\theta)^2\right] \\
&= \operatorname{Var}(W) + 2\,\mathrm{E}\left[W\,\mathrm{E}[W] - \theta W - (\mathrm{E}[W])^2 + \theta\,\mathrm{E}[W]\right] + \mathrm{E}\left[(\operatorname{Bias}(W))^2\right] \\
&= \operatorname{Var}(W) + 2\left[\mathrm{E}[W\,\mathrm{E}[W]] - \mathrm{E}[\theta W] - \mathrm{E}\left[(\mathrm{E}[W])^2\right] + \mathrm{E}[\theta\,\mathrm{E}[W]]\right] + [\operatorname{Bias}(W)]^2 \\
&= \operatorname{Var}(W) + 2\left[(\mathrm{E}[W])^2 - \theta\,\mathrm{E}[W] - (\mathrm{E}[W])^2 + \theta\,\mathrm{E}[W]\right] + [\operatorname{Bias}(W)]^2 \\
&= \operatorname{Var}(W) + 2\,(0) + [\operatorname{Bias}(W)]^2 \\
&= \operatorname{Var}(W) + [\operatorname{Bias}(W)]^2.
\end{aligned}$$

If W is unbiased, $\operatorname{MSE}[W] = \operatorname{Var}(W)$. An estimator that has good MSE properties has small combined variance and bias. Controlling bias does not guarantee that MSE is controlled. There is often a bias-variance trade-off, such that a small increase in bias can be traded for a larger decrease in variance, resulting in an improvement in MSE.

EXAMPLE 9.31. Let $X_1, \ldots, X_n$ be iid Bernoulli (θ). Consider the statistics

$$T_1 = \frac{\sum_i X_i + 1}{n+2} \quad \text{and} \quad T_2 = \frac{\sum_i X_i}{n}.$$

Find the MSE for T_1 and T_2.

Noting that the expected value of a Bernoulli (p) random variable is given by p, we will begin by finding the expected value of T_1.

$$\begin{aligned}
\mathrm{E}[T_1] &= \mathrm{E}\left[\frac{\sum_i X_i + 1}{n+2}\right] \\
&= \frac{1}{n+2}\,\mathrm{E}\left[\sum_{i=1}^{n} X_i + 1\right] \\
&= \frac{1}{n+2}\left(\sum_{i=1}^{n} \mathrm{E}[X_i] + 1\right) \\
&= \frac{1}{n+2}\left(\sum_{i=1}^{n} \theta + 1\right) \\
&= \frac{n\theta + 1}{n+2}.
\end{aligned}$$

Then, the bias of T_1 is given by

$$\mathrm{Bias}(T_1) = \mathrm{E}[T_1] - \theta = \frac{n\theta+1}{n+2} - \frac{\theta(n+2)}{n+2} = \frac{n\theta + 1 - n\theta - 2\theta}{n+2} = \frac{1-2\theta}{n+2}.$$

Noting that the variance of a Bernoulli (p) random variable is given by $p(1-p)$, it follows that the variance of T_1 is given by

$$\begin{aligned}
\mathrm{Var}(T_1) &= \mathrm{Var}\left(\frac{\sum_i X_i + 1}{n+2}\right) \\
&= \left(\frac{1}{n+2}\right)^2 \mathrm{Var}\left(\sum_{i=1}^{n} X_i + 1\right) \\
&= \frac{1}{(n+2)^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{(n+2)^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \\
&= \frac{1}{(n+2)^2}\sum_{i=1}^{n} \theta(1-\theta) \\
&= \frac{n\theta(1-\theta)}{(n+2)^2},
\end{aligned}$$

where the fourth equality follows from the independence of the X_i's. Then, the MSE for T_1 is given by

$$\begin{aligned}
\mathrm{MSE}(T_1) &= \mathrm{Var}(T_1) + [\mathrm{Bias}(T_1)]^2 \\
&= \frac{n\theta(1-\theta)}{(n+2)^2} + \left[\frac{1-2\theta}{n+2}\right]^2 \\
&= \frac{n\theta(1-\theta) + (1-2\theta)^2}{(n+2)^2} \\
&= \frac{n\theta - n\theta^2 + \left(1 - 4\theta + 4\theta^2\right)}{(n+2)^2} \\
&= \frac{\theta^2(4-n) - \theta(4-n) + 1}{(n+2)^2}.
\end{aligned}$$

The expected value of T_2 is given by

$$\mathrm{E}[T_2] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[X_i] = \frac{1}{n}\sum_{i=1}^{n} \theta = \frac{n\theta}{n} = \theta,$$

so that we have

$$\operatorname{Bias}(T_2) = \mathrm{E}[T_2] - \theta = \theta - \theta = 0.$$

The variance of T_2 is given by

$$\begin{aligned}
\operatorname{Var}(T_2) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\operatorname{Var}(X_i) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\theta(1-\theta) \\
&= \frac{n\theta(1-\theta)}{n^2} \\
&= \frac{\theta(1-\theta)}{n},
\end{aligned}$$

so that the MSE for T_2 is given by

$$\operatorname{MSE}(T_2) = \operatorname{Var}(T_2) + \left[\operatorname{Bias}(T_2)^2\right] = \frac{\theta(1-\theta)}{n} + 0 = \frac{\theta(1-\theta)}{n}.$$

REMARK 9.32. The MSEs may differ based on how large n is, i.e., one may be superior for certain ranges of n.

9.2.2. Best unbiased estimators. A common way to make the problem of finding a "best" estimator tractable is to limit the class of estimators to unbiased estimators.

DEFINITION 9.33. An estimator W^* is a *best unbiased estimator* of $\tau(\theta)$ if it satisfies $\mathrm{E}_\theta[W^*] = \tau(\theta)$ for all θ and, for any other estimator W with $\mathrm{E}_\theta[W] = \tau(\theta)$, we have $\operatorname{Var}_\theta(W^*) \le \operatorname{Var}_\theta(W)$ for all θ. W^* is also called a *uniform minimum variance unbiased estimator* (UMVUE) of $\tau(\theta)$.

THEOREM 9.34 (Cramér-Rao Inequality). *Let $X_1, \ldots, X_n$ be a sample with pdf $f(\mathbf{x}|\theta)$, and let $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ be any estimator satisfying*

$$\frac{\mathrm{d}}{\mathrm{d}\,\theta}\mathrm{E}_\theta[W(\mathbf{X})] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}[W(\mathbf{x}) f(\mathbf{x}|\theta)]\,\mathrm{d}\,\mathbf{x} \quad \textit{and} \quad \operatorname{Var}_\theta(W(\mathbf{X})) < \infty.$$

Then

$$\operatorname{Var}_\theta(W(\mathbf{X})) \ge \frac{\left[\frac{\mathrm{d}}{\mathrm{d}\,\theta}\mathrm{E}_\theta[W(\mathbf{X})]\right]^2}{\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]} = \frac{[\tau'(\theta)]^2}{\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]}$$

with Fisher information

$$I = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right] = -\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{X}|\theta)\right].$$

(This is Theorem 7.3.9 from Casella & Berger; the following proof is given there.)

PROOF. Let

$$U = W(\mathbf{X}) \quad \text{and} \quad V = \frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta).$$

Then, from the definition of correlation, we have

$$\operatorname{Corr}(U,V) = \frac{\operatorname{Cov}(U,V)}{\sqrt{\operatorname{Var}(U)}\sqrt{\operatorname{Var}(V)}}.$$

From theorem 4.8, it follows that

$$0 \leq [\operatorname{Corr}(U,V)]^2 = \frac{[\operatorname{Cov}(U,V)]^2}{\operatorname{Var}(U)\operatorname{Var}(V)} \leq 1,$$

so that we have

$$[\operatorname{Cov}(U,V)]^2 \leq \operatorname{Var}(U)\operatorname{Var}(V) \implies \operatorname{Var}(U) \geq \frac{[\operatorname{Cov}(U,V)]^2}{\operatorname{Var}(V)}.$$

Then,

$$\begin{aligned}
\operatorname{Cov}(U,V) &= \mathrm{E}\left[(U-\mu_U)(V-\mu_V)\right] \\
&= \mathrm{E}\left[UV - \mu_V U - \mu_U V + \mu_U\mu_V\right] \\
&= \mathrm{E}[UV] - \mu_V \mathrm{E}[U] - \mu_U \mathrm{E}[V] + \mu_U\mu_V \\
&= \mathrm{E}[UV] - \mathrm{E}[V]\mathrm{E}[U] - \mathrm{E}[U]\mathrm{E}[V] + \mathrm{E}[U]\mathrm{E}[V] \\
&= \mathrm{E}[UV] - \mathrm{E}[V]\mathrm{E}[U] \\
&= \mathrm{E}\left[W(\mathbf{X})\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right] - \mathrm{E}[W(\mathbf{X})]\,\mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right] \\
&= \int_{\mathcal{X}} W(\mathbf{X})\left[\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right] f(\mathbf{x}|\theta)\,\mathrm{d}\mathbf{x} - \mathrm{E}[W(\mathbf{X})]\int_{\mathcal{X}}\left[\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right] f(\mathbf{x}|\theta)\,\mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} W(\mathbf{X})\left[\frac{1}{f(\mathbf{x}|\theta)}\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right] f(\mathbf{x}|\theta)\,\mathrm{d}\mathbf{x} - \mathrm{E}[W(\mathbf{X})]\int_{\mathcal{X}}\left[\frac{1}{f(\mathbf{x}|\theta)}\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right] f(\mathbf{x}|\theta)\,\mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{X}} W(\mathbf{X})\left[\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right]\mathrm{d}\mathbf{x} - \mathrm{E}[W(\mathbf{X})]\int_{\mathcal{X}}\left[\frac{\partial}{\partial\theta}f(\mathbf{x}|\theta)\right]\mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial\theta}\int_{\mathcal{X}} W(\mathbf{X}) f(\mathbf{x}|\theta)\,\mathrm{d}\mathbf{x} - \mathrm{E}[W(\mathbf{X})]\frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(\mathbf{x}|\theta)\,\mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial\theta}\mathrm{E}[W(\mathbf{X})] - \mathrm{E}[W(\mathbf{X})]\frac{\partial}{\partial\theta}1 \\
&= \frac{\partial}{\partial\theta}\mathrm{E}[W(\mathbf{X})] - \mathrm{E}[W(\mathbf{X})]\cdot 0 \\
&= \frac{\partial}{\partial\theta}\mathrm{E}[W(\mathbf{X})] \\
&= \tau'(\theta).
\end{aligned}$$

Then, we have

$$\begin{aligned}
\operatorname{Var}(V) &= \operatorname{Var}\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right) \\
&= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right] - \left(\mathrm{E}\left[\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right]\right)^2 \\
&= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right] - 0 \\
&= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right],
\end{aligned}$$

so that

$$\operatorname{Var}(W(\mathbf{X})) \geq \frac{[\tau'(\theta)]^2}{\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]}.$$

□

The Cramér-Rao lower bound gives the smallest variance that can be attained for an unbiased estimator provided the conditions of the theorem are met. Under independence, $f(\mathbf{x}|\theta) = \prod_{i=1}^{n} f_i(x_i|\theta)$, so that we have

$$\begin{aligned}
\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right] &= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^{n} f_i(x_i|\theta)\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f_i(x_i|\theta)\right)^2\right] \\
&= \mathrm{E}\left[\left(\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f_i(x_i|\theta)\right)^2\right] \\
&= \sum_{i=1}^{n}\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f_i(x_i|\theta)\right)^2\right],
\end{aligned}$$

where the third quality follows from the linearity of differentiation and the final equality from the linearity of expected value. If the X_i's are also identically distributed, i.e., they are iid, then we have

$$\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right] = \sum_{i=1}^{n}\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f_i(x_i|\theta)\right)^2\right] = n\,\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right],$$

and theorem 9.34 reduces to

$$\mathrm{Var}_\theta(W(\mathbf{X})) \geq \frac{[\tau'(\theta)]^2}{n\,\mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right]}$$

for all estimators $W(\mathbf{X})$ that are unbiased for $\tau(\theta)$.

EXAMPLE 9.35 (Exponential CRLB). Let $X_1, \ldots, X_n \sim \mathrm{Exp}(\theta)$, i.e.,

$$f(x|\theta) = \theta \mathrm{e}^{-\theta x}, \quad x > 0, \quad \theta > 0.$$

Give the Cramér-Rao lower bound for any unbiased estimator of $\tau(\theta)$, where

(1) $\tau(\theta) = \theta$.

From example 8.21, the joint density of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$f(\mathbf{x}|\theta) = \theta^n \mathrm{e}^{-\theta\sum_{i=1}^{n} x_i},$$

so that we have

$$\log f(\mathbf{x}|\theta) = \log\theta^n \mathrm{e}^{-\theta\sum_{i=1}^{n} x_i} = n\log\theta - \theta\sum_{i=1}^{n} x_i.$$

Taking the derivative with respect to θ, we have

$$\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i,$$

so that

$$\begin{aligned}
\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right] &= \mathrm{E}\left[\frac{n^2}{\theta^2} - 2\frac{n}{\theta}\sum_{i=1}^{n} x_i + \left(\sum_{i=1}^{n} x_i\right)^2\right] \\
&= \mathrm{E}\left[\frac{n^2}{\theta^2}\right] - \mathrm{E}\left[2\frac{n}{\theta}\sum_{i=1}^{n} x_i\right] + \mathrm{E}\left[\left(\sum_{i=1}^{n} x_i\right)^2\right] \\
&= \frac{n^2}{\theta^2} - \frac{2n}{\theta}\mathrm{E}\left[\sum_{i=1}^{n} x_i\right] + \mathrm{E}\left[\left(\sum_{i=1}^{n} x_i\right)^2\right].
\end{aligned}$$

An exponential random variable with parameter λ is a gamma random variable with parameters $(1, \lambda)$. The sum of n independent gamma random variables $W_i \sim \Gamma(k_i, \lambda)$ has the distribution of a $\Gamma\left(\sum_{i=1}^n k_i, \lambda\right)$ random variable. We have $X_i \sim \text{Exp}(\theta)$, so it follows that $X_i \sim \Gamma(1, \theta)$ and that $\sum_{i=1}^n X_i \sim \Gamma(n, \theta)$. The expectation of a $\Gamma(\alpha, \beta)$ random variable is given by α/β , so it follows that

$$\text{E}\left[\sum_{i=1}^n x_i\right] = \frac{n}{\theta}.$$

The variance of a $\Gamma(\alpha, \beta)$ random variable is given by α/β^2, so it follows that

$$\begin{aligned}
\frac{n}{\theta^2} &= \text{Var}\left(\sum_{i=1}^n x_i\right) \\
&= \text{E}\left[\left(\sum_{i=1}^n x_i\right)^2\right] - \text{E}\left[\sum_{i=1}^n x_i\right]^2 \\
&= \text{E}\left[\left(\sum_{i=1}^n x_i\right)^2\right] - \left(\frac{n}{\theta}\right)^2 \\
\implies \text{E}\left[\left(\sum_{i=1}^n x_i\right)^2\right] &= \frac{n}{\theta^2} + \frac{n^2}{\theta^2} \\
&= \frac{n + n^2}{\theta^2}.
\end{aligned}$$

Then, we have

$$\text{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right] = \frac{n^2}{\theta^2} - \frac{2n}{\theta}\cdot\frac{n}{\theta} + \frac{n+n^2}{\theta^2} = \frac{n}{\theta^2}.$$

We could also have taken the second derivative of $\ln f(x|\theta)$ with respect to θ to obtain

$$\text{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right] = \text{E}\left[\frac{\partial}{\partial\theta}\left(\frac{n}{\theta} - \sum_{i=1}^n x_i\right)\right] = \text{E}\left[-\frac{n}{\theta^2}\right] = -\frac{n}{\theta^2}.$$

Then, using the fact that

$$\text{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right] = -\,\text{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(x|\theta)\right],$$

we have

$$\text{E}\left[\left(\frac{\partial}{\partial\theta}\log f(x|\theta)\right)^2\right] = -\left(-\frac{n}{\theta^2}\right) = \frac{n}{\theta^2}.$$

We have $\tau(\theta) = \theta$, so that $\tau'(\theta) = 1$. Then, the Cramér-Rao lower bound is given by

$$\frac{[\tau'(\theta)]^2}{\text{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right]} = \frac{1^2}{\frac{n}{\theta^2}} = \frac{\theta^2}{n}.$$

(2) $\tau(\theta) = 1/\theta$.

We have $\tau'(\theta) = -\theta^{-2}$, so the Cramér-Rao lower bound is given by

$$\frac{[\tau'(\theta)]^2}{\text{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta)\right)^2\right]} = \frac{\left(-\theta^{-2}\right)^2}{n\theta^{-2}} = \frac{\theta^{-4}}{n\theta^{-2}} = \frac{1}{n\theta^2}.$$

(3) In example 9.10, we found that the MLE of θ was given by $\hat{\theta} = \bar{X}$. Using the parameterization for $f(x|\theta)$ given here, we have $\hat{\theta} = 1/\bar{X}$. Is $\hat{\theta}$ a best unbiased estimator for $\tau(\theta) = 1/\theta$?

We must first show that $\hat{\theta}$ is unbiased. From theorem 9.18, we have

$$\widehat{\tau(\theta)} = \frac{1}{\hat{\theta}} = \frac{1}{\frac{1}{\bar{X}}} = \bar{X},$$

so that

$$\mathrm{E}\left[\bar{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\left(\frac{n}{\theta}\right) = \frac{1}{\theta} = \tau(\theta).$$

It follows that $\hat{\theta}$ is an unbiased estimator for $\tau(\theta)$. Then, the variance of $\widehat{\tau(\theta)}$ is given by

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

where the third equality follows from theorem 4.7. We see that the variance of $\widehat{\tau(\theta)}$ is equal to the Cramér-Rao lower bound found above, so it follows that $\hat{\theta}$ is a best unbiased estimator (UMVUE) for $\tau(\theta) = 1/\theta$.

(4) Is the MLE of θ a best unbiased estimator for $\tau(\theta) = \theta$?

We have $\widehat{\tau(\theta)} = \hat{\theta} = 1/\bar{X}$, so that the variance of $\hat{\theta}$ is given by

$$\begin{aligned}
\mathrm{Var}\left(\hat{\theta}\right) &= \mathrm{E}\left[\hat{\theta}^2\right] - \left(\mathrm{E}\left[\hat{\theta}\right]\right)^2 \\
&= \mathrm{E}\left[\left(\frac{1}{\bar{X}}\right)^2\right] - \left(\mathrm{E}\left[\frac{1}{\bar{X}}\right]\right)^2 \\
&= \mathrm{E}\left[\frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2}\right] - \left(\mathrm{E}\left[\frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i}\right]\right)^2 \\
&= \mathrm{E}\left[\frac{n^2}{\left(\sum_{i=1}^{n} X_i\right)^2}\right] - \left(\mathrm{E}\left[\frac{n}{\sum_{i=1}^{n} X_i}\right]\right)^2 \\
&= n^2\,\mathrm{E}\left[\frac{1}{\left(\sum_{i=1}^{n} X_i\right)^2}\right] - \left(n\,\mathrm{E}\left[\frac{1}{\sum_{i=1}^{n} X_i}\right]\right)^2 \\
&= n^2\,\mathrm{E}\left[\frac{1}{\left(\sum_{i=1}^{n} X_i\right)^2}\right] - n^2\left(\mathrm{E}\left[\frac{1}{\sum_{i=1}^{n} X_i}\right]\right)^2.
\end{aligned}$$

Let $U = \sum_{i=1}^{n} X_i \sim \Gamma(n, \theta)$, so that the pdf of U is given by

$$f_U(u|n,\theta) = \frac{\theta^n}{\Gamma(n)} u^{n-1} \mathrm{e}^{-\theta u} \qquad u > 0,$$

so that we have

$$\mathrm{E}\left[\frac{1}{\bar{X}}\right] = n\,\mathrm{E}\left[\frac{1}{U}\right] \quad \text{and} \quad \mathrm{E}\left[\left(\frac{1}{\bar{X}^2}\right)\right] = n^2\,\mathrm{E}\left[\frac{1}{U^2}\right].$$

Noting that $\Gamma(n) = (n-1)\,\Gamma(n-1)$, the expected value of $1/U$ is given by

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{U}\right] &= \int_0^\infty \frac{1}{u} \cdot f_U(u|n,\theta)\,\mathrm{d}\,u \\
&= \int_0^\infty \frac{1}{u} \cdot \frac{\theta^n}{\Gamma(n)} u^{n-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \int_0^\infty \frac{\theta^n}{\Gamma(n)} u^{(n-1)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \int_0^\infty \frac{\theta\cdot\theta^{n-1}}{(n-1)\,\Gamma(n-1)} u^{(n-1)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \frac{\theta}{n-1}\int_0^\infty \frac{\theta^{n-1}}{\Gamma(n-1)} u^{(n-1)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u.
\end{aligned}$$

We recognize the integrand as the pdf of a $\Gamma(n-1,\theta)$ random variable, so that we have

$$\mathrm{E}\left[\frac{1}{U}\right] = \frac{\theta}{n-1}\int_0^\infty \frac{\theta^{n-1}}{\Gamma(n-1)} u^{(n-1)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u = \frac{\theta}{n-1}\cdot 1 = \frac{\theta}{n-1}.$$

Then, the expected value of $1/U^2$ is given by

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{U^2}\right] &= \int_0^\infty \frac{1}{u^2}\cdot f_U(u|n\theta)\,\mathrm{d}\,u \\
&= \int_0^\infty \frac{1}{u^2}\cdot\frac{\theta^n}{\Gamma(n)} u^{n-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \int_0^\infty \frac{\theta^2\cdot\theta^{n-2}}{(n-1)\,\Gamma(n-1)} u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \int_0^\infty \frac{\theta^2\cdot\theta^{n-2}}{(n-1)\,(n-2)\,\Gamma(n-2)} u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \frac{\theta^2}{(n-1)\,(n-2)}\int_0^\infty \frac{\theta^{n-2}}{\Gamma(n-2)} u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u.
\end{aligned}$$

We recognize the integrand as the pdf of a $\Gamma(n-2,\theta)$ random variable, so that we have

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{U^2}\right] &= \frac{\theta^2}{(n-1)\,(n-2)}\int_0^\infty \frac{\theta^{n-2}}{\Gamma(n-2)} u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \frac{\theta^2}{(n-1)\,(n-2)}\cdot 1 \\
&= \frac{\theta^2}{(n-1)\,(n-2)}.
\end{aligned}$$

Then, the variance of $\hat{\theta}$ is given by

$$\begin{aligned}
\mathrm{Var}\left(\hat{\theta}\right) &= n^2\,\mathrm{E}\left(\frac{1}{U^2}\right) - n^2\left(\mathrm{E}\left[\frac{1}{U}\right]\right)^2 \\
&= n^2\left(\frac{\theta^2}{(n-1)\,(n-2)}\right) - n^2\left(\frac{\theta}{n-1}\right)^2 \\
&= \frac{n^2\theta^2}{(n-1)\,(n-2)} - \frac{n^2\theta^2}{(n-1)^2} \\
&= \frac{n^2\theta^2\,(n-1) - n^2\theta^2\,(n-2)}{(n-1)^2\,(n-2)} \\
&= \frac{n^2\theta^2\,[(n-1)-(n-2)]}{(n-1)^2\,(n-2)} \\
&= \frac{n^2\theta^2\,(n-1-n+2)}{(n-1)^2\,(n-2)} \\
&= \frac{n^2\theta^2}{(n-1)^2\,(n-2)}.
\end{aligned}$$

We will check whether the variance of $\hat{\theta}$ attains the Cramér-Rao Lower Bound for the variance of an estimator of θ, which we found above to be θ^2/n.

$$\mathrm{Var}\left(\hat{\theta}\right) > \frac{\theta^2}{n} \implies \frac{n^2\theta^2}{(n-1)^2\,(n-2)} - \frac{\theta^2}{n} > 0 \implies \frac{n^2}{(n-1)^2\,(n-2)} - \frac{1}{n} = \frac{n^3-(n-1)^2\,(n-2)}{n\,(n-1)^2\,(n-2)} > 0$$

We will expand the expression $(n-1)^2\,(n-2)$ as

$$\begin{aligned}
(n-1)^2\,(n-2) &= \left(n^2-2n+1\right)(n-2) \\
&= \left(n^2-(2n-1)\right)(n-2)
\end{aligned}$$

$$= n^3 - 2n^2 - n(2n-1) + 2(2n-1)$$
$$= n^3 - 2n^2 - 2n^2 + n + 4n - 2$$
$$= n^3 - 4n^2 + 5n - 2,$$

so that we have

$$\frac{n^3 - \left(n^3 - 4n^2 + 5n - 2\right)}{n\left(n^3 - 4n^2 + 5n - 2\right)} > 0$$
$$\frac{4n^2 - 5n + 2}{n(n-1)^2(n-2)} > 0.$$

This expression will be defined if and only if the denominator is non-zero. By definition, we have $n \geq 1$. In the case that $n = 1$ or $n = 2$, the denominator will be equal to zero, so the expression will be defined if and only if $n > 2 \Leftrightarrow n \geq 3$. In the case that $n = 3$, the numerator is equal to $36 - 15 + 2 = 23 > 0$, and as n increases, we see that $4n^2 - 5n + 2 > 0$, so that the numerator is positive. Thus, we conclude that the variance of $\hat{\theta}$ is greater than the Cramér-Rao Lower Bound, so $\hat{\theta}$ is not a best unbiased estimator for θ.

EXAMPLE 9.36 (Beta CRLB). Let $X_1, \ldots, X_n$ denote a random sample of size $n > 2$ from a Beta$(\theta, 1)$ distribution, i.e.,

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

(1) Find the MLE of θ.

The joint density of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \theta x_i^{\theta-1} = \theta^n \prod_{i=1}^{n} x_i^{\theta-1},$$

so that the log-likelihood is given by

$$\log L(\theta|\mathbf{x}) = \log \theta^n \prod_{i=1}^{n} x_i^{\theta-1}$$
$$= n \log \theta + \log \prod_{i=1}^{n} x_i^{\theta-1}$$
$$= n \log \theta + \sum_{i=1}^{n} \log x_i^{\theta-1}$$
$$= n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log x_i.$$

Taking the derivative with respect to θ gives

$$\frac{\partial}{\partial \theta} \log L(\theta|\mathbf{x}) = \frac{\partial}{\partial \theta} \left[n \log \theta + (\theta - 1) \sum_{i=1}^{n} \log x_i \right] = \frac{n}{\theta} + \sum_{i=1}^{n} \log x_i.$$

Setting this equal to zero, we have

$$\frac{n}{\hat{\theta}} = -\sum_{i=1}^{n} \log x_i \implies \hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \log x_i}.$$

We will evaluate the second derivative of the log-likelihood at $\theta = \hat{\theta}$ to verify that $\hat{\theta}$ is a maximum.

$$\frac{\partial^2}{\partial \theta} \log L(\theta|\mathbf{x})\Big|_{\theta=\hat{\theta}} = \frac{\partial}{\partial \theta} \left[\frac{n}{\theta} + \sum_{i=1}^{n} \log x_i \right]_{\theta=\hat{\theta}}$$
$$= \left[-\frac{n}{\theta^2} \right]_{\theta=\hat{\theta}}$$

$$
\begin{aligned}
&= -\frac{n}{\left(-\frac{n}{\sum_{i=1}^{n}\log x_i}\right)^2} \\
&= -\frac{n}{\frac{n^2}{\left(\sum_{i=1}^{n}\log x_i\right)^2}} \\
&= -\frac{\left(\sum_{i=1}^{n}\log x_i\right)^2}{n}
\end{aligned}
$$

We have $n \geq 1$, so that the denominator is positive. We have $0 < x < 1$, so that $\log x < 0$, so that $\sum_{i=1}^{n}\log x_i < 0$, so that $\left(\sum_{i=1}^{n}\log x_i\right)^2 > 0$. Both the numerator and the denominator are positive, so it follows that the expression is negative, so that $\hat{\theta}$ is the MLE.

(2) Find an unbiased estimator based on the MLE.

We have

$$
\mathrm{E}\left[\hat{\theta}\right] = \mathrm{E}\left[-\frac{n}{\sum_{i=1}^{n}\log X_i}\right] = \mathrm{E}\left[\frac{n}{\sum_{i=1}^{n}-\log X_i}\right].
$$

Let $Y = -\log X$, so that

$$
Y = -\log X \implies -Y = \log X \implies \mathrm{e}^{-Y} = \mathrm{e}^{\log X} \implies X = \mathrm{e}^{-Y}.
$$

Noting that $Y = -\log X$ is monotone, we will find a pdf for Y using theorem 3.2. We have

$$
\mathcal{Y} = \{y : y = -\log x, 0 < x < 1\} = \{y : y > 0\}
$$

and

$$
f_Y(y) = f_X\left(\mathrm{e}^{-y}\right)\left|\frac{\mathrm{d}}{\mathrm{d}\,y}\mathrm{e}^{-y}\right| = \theta\left(\mathrm{e}^{-y}\right)^{\theta-1}\left|-\mathrm{e}^{-y}\right| = \theta\mathrm{e}^{y-\theta y}\mathrm{e}^{-y} = \theta\mathrm{e}^{-\theta y},
$$

which is the pdf of an $\mathrm{Exp}(\theta)$ random variable, i.e., $Y \sim \mathrm{Exp}(\theta)$. Let

$$
U = \sum_{i=1}^{n} -\log X_i = \sum_{i=1}^{n} Y_i \sim \Gamma(n, \theta),
$$

i.e., U is a $\mathrm{Gamma}(n, \theta)$ random variable with pdf

$$
f_U(u|n, \theta) = \frac{\theta^n}{\Gamma(n)}u^{n-1}\mathrm{e}^{-\theta u} \qquad u > 0.
$$

Noting that $\Gamma(n) = (n-1)\,\Gamma(n-1)$, we have

$$
\begin{aligned}
\mathrm{E}\left[\hat{\theta}\right] &= \mathrm{E}\left[\frac{n}{\sum_{i=1}^{n}-\log X_i}\right] \\
&= \mathrm{E}\left[\frac{n}{U}\right] \\
&= n\,\mathrm{E}\left[\frac{1}{U}\right] \\
&= n\int_0^\infty \frac{1}{u}\cdot f_U(u|n,\theta)\,\mathrm{d}\,u \\
&= n\int_0^\infty \frac{1}{u}\cdot\frac{\theta^n}{\Gamma(n)}u^{n-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= n\int_0^\infty \frac{\theta\cdot\theta^{n-1}}{(n-1)\,\Gamma(n-1)}u^{(n-1)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \frac{n\theta}{n-1}\int_0^\infty \frac{\theta^{n-1}}{\Gamma(n-1)}u^{(n-1)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u.
\end{aligned}
$$

We recognize the integrand as the pdf of a $\Gamma(n-1, \theta)$ random variable, so that we have

$$
\mathrm{E}\left[\hat{\theta}\right] = \frac{n\theta}{n-1}\int_0^\infty \frac{\theta^{n-1}}{\Gamma(n-1)}u^{(n-1)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u = \frac{n\theta}{n-1}.
$$

It follows that $\hat{\theta}$ is a biased estimator for θ. We can construct an unbiased estimator for θ based on $\hat{\theta}$ by multiplying $\hat{\theta}$ by $(n-1)/n$, i.e., let

$$\tilde{\theta} = \frac{n-1}{n}\hat{\theta}, \quad \text{so that} \quad \mathrm{E}\left[\tilde{\theta}\right] = \mathrm{E}\left[\frac{n-1}{n}\hat{\theta}\right] = \frac{n-1}{n}\,\mathrm{E}\left[\hat{\theta}\right] = \frac{n-1}{n}\cdot\frac{n\theta}{n-1} = \theta.$$

(3) Does the unbiased estimator from (b) attain the Cramér-Rao Lower Bound?

The Cramér-Rao Lower Bound for the variance of an estimator $W(\mathbf{X})$ of θ is given by

$$\mathrm{Var}\left(W(\mathbf{X})\right) \geq \frac{\left[\tau'(\theta)\right]^2}{-\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{X}|\theta)\right]}.$$

We have $\tau(\theta) = \theta$, so that $\tau'(\theta) = 1$. Then, we have

$$\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{X}|\theta)\right] = \mathrm{E}\left[-\frac{n}{\theta^2}\right] = -\frac{n}{\theta^2},$$

so that we have

$$\mathrm{Var}\left(W(\mathbf{X})\right) \geq \frac{1}{-\left(-\frac{n}{\theta^2}\right)} = \frac{\theta^2}{n}.$$

The variance of $\tilde{\theta}$ is given by

$$\mathrm{Var}\left(\tilde{\theta}\right) = \mathrm{Var}\left(\frac{n-1}{n}\hat{\theta}\right) = \left(\frac{n-1}{n}\right)^2\mathrm{Var}\left(\hat{\theta}\right) = \frac{(n-1)^2}{n^2}\left(\mathrm{E}\left[\hat{\theta}^2\right] - \left(\mathrm{E}\left[\hat{\theta}\right]\right)^2\right).$$

The expected value of $\hat{\theta}^2$ is given by

$$\begin{aligned}
\mathrm{E}\left[\hat{\theta}^2\right] &= \mathrm{E}\left[\left(-\frac{n}{\sum_{i=1}^n \log X_i}\right)^2\right] \\
&= \mathrm{E}\left[\left(\frac{n}{\sum_{i=1}^n -\log X_i}\right)^2\right] \\
&= n^2\,\mathrm{E}\left[\frac{1}{U^2}\right] \\
&= n^2\int_0^\infty \frac{1}{u^2}\cdot f_U(u|n,\theta)\,\mathrm{d}\,u \\
&= n^2\int_0^\infty \frac{1}{u^2}\cdot\frac{\theta^n}{\Gamma(n)}u^{n-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= n^2\int_0^\infty \frac{\theta^2\cdot\theta^{n-2}}{(n-1)\,\Gamma(n-1)}u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= n^2\int_0^\infty \frac{\theta^2\cdot\theta^{n-2}}{(n-1)\,(n-2)\,\Gamma(n-2)}u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u \\
&= \frac{n^2\theta^2}{(n-1)\,(n-2)}\int_0^\infty \frac{\theta^{n-2}}{\Gamma(n-2)}u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u.
\end{aligned}$$

We recognize the integrand as the pdf of a Gamma$(n-2,\theta)$ random variable, so that we have

$$\mathrm{E}\left[\hat{\theta}^2\right] = \frac{n^2\theta^2}{(n-1)\,(n-2)}\int_0^\infty \frac{\theta^{n-2}}{\Gamma(n-2)}u^{(n-2)-1}\mathrm{e}^{-\theta u}\,\mathrm{d}\,u = \frac{n^2\theta^2}{(n-1)\,(n-2)}.$$

Then, we have

$$\begin{aligned}
\mathrm{Var}\left(\tilde{\theta}\right) &= \frac{(n-1)^2}{n^2}\left(\mathrm{E}\left[\hat{\theta}^2\right] - \left(\mathrm{E}\left[\hat{\theta}\right]\right)^2\right) \\
&= \frac{(n-1)^2}{n^2}\left(\frac{n^2\theta^2}{(n-1)\,(n-2)} - \left(\frac{n\theta}{n-1}\right)^2\right)
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(n-1)^2}{n^2}\left(\frac{n^2\theta^2}{(n-1)(n-2)} - \frac{n^2\theta^2}{(n-1)^2}\right) \\
&= \frac{(n-1)^2}{n^2}\left(\frac{n^2\theta^2(n-1) - n^2\theta^2(n-2)}{(n-1)^2(n-2)}\right) \\
&= \frac{1}{n^2}\left(\frac{n^2\theta^2[(n-1)-(n-2)]}{n-2}\right) \\
&= \frac{\theta^2(n-1-n+2)}{n-2} \\
&= \frac{\theta^2}{n-2}.
\end{aligned}
$$

We will check whether the variance of $\tilde{\theta}$ attains the Cramér-Rao Lower Bound. Noting that $\theta > 0$, we have

$$\operatorname{Var}\left(\tilde{\theta}\right) > \frac{\theta^2}{n} \implies \frac{\theta^2}{n-2} > \frac{\theta^2}{n} \implies \frac{1}{n-2} > \frac{1}{n}.$$

By definition, we have $n \geq 1$. In the case that $n = 1$, the inequality will be false. In the case that $n = 2$, the denominator will be equal to zero, so the expression will be defined (and true) if and only if $n > 2 \Leftrightarrow n \geq 3$. Thus, we conclude that the variance of $\tilde{\theta}$ is greater than the Cramér-Rao Lower Bound, so $\tilde{\theta}$ is not a best unbiased estimator for θ.

The corollary below gives a way to find a UMVUE by giving an estimator that attains the Cramér-Rao Lower Bound.

COROLLARY 9.37 (Attainment). *Let $X_1, \ldots, X_n$ be iid $f(x|\theta)$, where $f(x|\theta)$ satisfies the conditions of the Cramér-Rao Theorem. Let $\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta)$ denote the likelihood function. If $W(\mathbf{X}) = W(X_1, \ldots, X_n)$ is any unbiased estimator of $\tau(\theta)$, then $W(\mathbf{X})$ attains the Cramér-Rao Lower Bound if and only if*

$$a(\theta)\left[W(\mathbf{x}) - \tau(\theta)\right] = \frac{\partial}{\partial\theta}\log\mathcal{L}(\theta|\mathbf{x})$$

for some function $a(\theta)$. (This is Corollary 7.3.15 from Casella & Berger; the following proof is given there.)

PROOF. As shown in the proof of theorem 9.34, the Cramér-Rao Inequality can be written as

$$\left[\operatorname{Cov}(U, V)\right]^2 \leq \operatorname{Var}(U)\operatorname{Var}(V),$$

where

$$U = W(\mathbf{X}) \quad \text{and} \quad V = \frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta),$$

i.e.,

$$\left[\operatorname{Cov}_\theta\left(W(\mathbf{X}), \frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)\right)\right]^2 \leq \operatorname{Var}_\theta(W(\mathbf{X}))\operatorname{Var}_\theta\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)\right).$$

Recalling that $\mathrm{E}_\theta[W(\mathbf{X})] = \tau(\theta)$ and that

$$\mathrm{E}_\theta\left[\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(X_i|\theta)\right] = 0,$$

and using the results of theorem 4.8, we can have equality if and only if $W(\mathbf{x}) - \tau(\theta)$ is proportional to $\frac{\partial}{\partial\theta}\log\prod_{i=1}^n f(x_i|\theta)$. That is exactly what is expressed in the corollary above. □

EXAMPLE 9.38. Let $X_1, \ldots, X_n$ be iid with pdf

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

(1) Find the MLE of θ.

The joint density of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{\theta}{(1+x_i)^{1+\theta}} = \theta^n \prod_{i=1}^{n} \frac{1}{(1+x_i)^{1+\theta}},$$

so that the log-likelihood is given by

$$\begin{aligned}
\log L(\theta|\mathbf{x}) &= \log \theta^n \prod_{i=1}^{n} \frac{1}{(1+x_i)^{1+\theta}} \\
&= n \log \theta + \log \prod_{i=1}^{n} \frac{1}{(1+x_i)^{1+\theta}} \\
&= n \log \theta + \sum_{i=1}^{n} \log (1+x_i)^{-(1+\theta)} \\
&= n \log \theta - (1+\theta) \sum_{i=1}^{n} \log (1+x_i).
\end{aligned}$$

Taking the derivative with respect to θ gives

$$\frac{\partial}{\partial \theta} \log L(\theta|\mathbf{x}) = \frac{\partial}{\partial \theta} \left[n \log \theta - (1+\theta) \sum_{i=1}^{n} \log (1+x_i) \right] = \frac{n}{\theta} - \sum_{i=1}^{n} \log (1+x_i).$$

Setting this equal to zero, we have

$$\frac{n}{\hat{\theta}} = \sum_{i=1}^{n} \log (1+x_i) \implies \hat{\theta} = \frac{n}{\sum_{i=1}^{n} \log (1+x_i)}.$$

We will evaluate the second derivative of the log-likelihood at $\theta = \hat{\theta}$ to verify that $\hat{\theta}$ is a maximum.

$$\begin{aligned}
\frac{\partial^2}{\partial \theta^2} \log L(\theta|\mathbf{x})\Big|_{\theta=\hat{\theta}} &= \frac{\partial}{\partial \theta} \left[\frac{n}{\theta} - \sum_{i=1}^{n} \log (1+x_i) \right]_{\theta=\hat{\theta}} \\
&= \left[-\frac{n}{\theta^2} \right]_{\theta=\hat{\theta}} \\
&= -\frac{n}{\left(\frac{n}{\sum_{i=1}^{n} \log(1+x_i)} \right)^2} \\
&= -\frac{n}{\frac{n^2}{\left(\sum_{i=1}^{n} \log(1+x_i) \right)^2}} \\
&= -\frac{\left(\sum_{i=1}^{n} \log (1+x_i) \right)^2}{n}
\end{aligned}$$

We have $x > 0$, so that $1 + x > 1$, so that $\log(1+x) > 0$, so that the numerator is positive. We have $n \geq 1$, so that the denominator is positive. It follows that the expression will be negative, so that $\hat{\theta}$ is the MLE.

(2) Find the Cramér-Rao Lower Bound for θ.

The Cramér-Rao Lower Bound for the variance of an estimator $W(\mathbf{X})$ of θ is given by

$$\operatorname{Var}(W(\mathbf{X})) \geq \frac{[\tau'(\theta)]^2}{-\operatorname{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{X}|\theta) \right]}.$$

We have $\tau(\theta) = \theta$, so that $\tau'(\theta) = 1$. Then, we have

$$\operatorname{E}\left[\frac{\partial^2}{\partial \theta^2} \log f(\mathbf{X}|\theta) \right] = \operatorname{E}\left[-\frac{n}{\theta^2} \right] = -\frac{n}{\theta^2},$$

so that we have

$$\operatorname{Var}(W(\mathbf{X})) \geq \frac{1}{-\left(-\frac{n}{\theta^2}\right)} = \frac{\theta^2}{n}.$$

(3) Find the UMVUE of $1/\theta$.

We found above that

$$\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x}) = \frac{n}{\theta} - \sum_{i=1}^{n}\log(1+x_i),$$

which we can write as

$$\frac{\partial}{\partial\theta}\log f(\mathbf{x}|\theta) = -n\left(-\frac{1}{\theta} + \frac{1}{n}\sum_{i=1}^{n}\log(1+x_i)\right) = \underbrace{-n}_{a(\theta)}\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n}\log(1+x_i)}_{W(\mathbf{x})} - \underbrace{\frac{1}{\theta}}_{\tau(\theta)}\right).$$

We will now check whether $W(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n}\log(1+X_i)$ is unbiased for $1/\theta$.

$$\begin{aligned}
\mathrm{E}[W(\mathbf{X})] &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\log(1+X_i)\right] \\
&= \frac{1}{n}\mathrm{E}\left[\sum_{i=1}^{n}\log(1+X_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}[\log(1+X_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty \log(1+x_i)\cdot f(x_i|\theta)\,\mathrm{d}x_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^\infty \log(1+x_i)\cdot\frac{\theta}{(1+x_i)^{1+\theta}}\mathrm{d}x_i \\
&= \frac{1}{n}\sum_{i=1}^{n}\lim_{c\to\infty}\int_0^c \log(1+x_i)\cdot\frac{\theta}{(1+x_i)^{1+\theta}}\mathrm{d}x_i
\end{aligned}$$

Let

$$u = \log(1+x), \quad \text{so that} \quad \mathrm{d}u = \frac{1}{1+x}\mathrm{d}x$$

and let

$$\mathrm{d}v = \frac{\theta}{(1+x)^{1+\theta}}\mathrm{d}x = \theta(1+x)^{-(1+\theta)}\,\mathrm{d}x = \theta(1+x)^{-\theta-1},$$

so that

$$v = -(1+x)^{-\theta} = -\frac{1}{(1+x)^{\theta}}.$$

Then, integration by parts gives

$$\int_a^b u\mathrm{d}v = [uv]_a^b - \int_a^b v\mathrm{d}u$$

$$\begin{aligned}
\Leftrightarrow \int_0^c \log(1+x)\cdot\frac{\theta}{(1+x)^{1+\theta}}\mathrm{d}x &= \left[\log(1+x)\cdot\left(-\frac{1}{(1+x)^{\theta}}\right)\right|_0^c - \int_0^c\left(-\frac{1}{(1+x)^{\theta}}\right)\frac{1}{1+x}\mathrm{d}x \\
&= \left[-\frac{\log(1+c)}{(1+c)^{\theta}} - \left(-\frac{\log(1+0)}{(1+0)^{\theta}}\right)\right] + \int_0^c\frac{1}{(1+x)^{\theta+1}}\mathrm{d}x \\
&= \left[-\frac{\log(1+c)}{(1+c)^{\theta}} + \log 1\right] + \left[-\frac{1}{\theta(1+x)^{\theta}}\right|_0^c
\end{aligned}$$

$$= -\frac{\log(1+c)}{(1+c)^{\theta}} + 0 + \left[-\frac{1}{\theta(1+c)^{\theta}} - \left(-\frac{1}{\theta(1+0)^{\theta}} \right) \right]$$
$$= -\frac{\log(1+c)}{(1+c)^{\theta}} - \frac{1}{\theta(1+c)^{\theta}} + \frac{1}{\theta},$$

so that we have

$$\mathrm{E}[W(\mathbf{X})] = \frac{1}{n}\sum_{i=1}^{n} \lim_{c\to\infty} \left[-\frac{\log(1+c)}{(1+c)^{\theta}} - \frac{1}{\theta(1+c)^{\theta}} + \frac{1}{\theta} \right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left[-\lim_{c\to\infty} \frac{\log(1+c)}{(1+c)^{\theta}} - \lim_{c\to\infty} \frac{1}{\theta(1+c)^{\theta}} + \lim_{c\to\infty} \frac{1}{\theta} \right]$$
$$= \frac{1}{n}\sum_{i=1}^{n} \left[-\lim_{c\to\infty} \frac{\log(1+c)}{(1+c)^{\theta}} - 0 + \frac{1}{\theta} \right].$$

Using L'Hôpital's rule to evaluate this limit gives

$$\lim_{c\to\infty} \frac{\log(1+c)}{(1+c)^{\theta}} = \lim_{c\to\infty} \frac{(1+c)^{-1}}{\theta(1+c)^{\theta-1}} = \lim_{c\to\infty} \frac{1}{\theta(1+c)^{\theta}} = 0,$$

so that we have

$$\mathrm{E}[W(\mathbf{X})] = \frac{1}{n}\sum_{i=1}^{n} \left[0 + \frac{1}{\theta} \right] = \frac{1}{n}\left(n\frac{1}{\theta} \right) = \frac{1}{\theta}.$$

We have $\mathrm{E}[W(\mathbf{X})] = 1/\theta$ for all θ, so it follows that $W(\mathbf{X})$ is an unbiased estimator for θ. Then, by corollary 9.37, $W(\mathbf{X})$ attains the Cramér-Rao Lower Bound, and it follows that $W(\mathbf{X}) = \frac{1}{n}\sum_{i=1}^{n} \log(1+X_i)$ is the UMVUE of θ.

9.2.3. Sufficiency and Unbiasedness.

THEOREM 9.39 (Rao-Blackwell). *Let W be any unbiased estimator of $\tau(\theta)$, and let T be a sufficient statistic for θ. Define $\phi(T) = \mathrm{E}[W|T]$. Then $\mathrm{E}_{\theta}[\phi(T)] = \tau(\theta)$ and $\mathrm{Var}_{\theta}(\phi(T)) \leq \mathrm{Var}_{\theta}(W)$ for all θ; that is, $\phi(T)$ is a uniformly better unbiased estimator of $\tau(\theta)$. (This is Theorem 7.3.17 from Casella & Berger; the following proof is given there.)*

PROOF. [proof goes here] □

EXAMPLE 9.40 (Bernoulli Rao-Blackwellization). Let $X_1, \ldots, X_n \sim \mathrm{Bernoulli}(\theta)$ and consider $\tau(\theta) = \theta$. In example 8.13, we found that $T(\mathbf{X}) = \sum_{i=1}^{n} X_i$ was a sufficient statistic for θ. Consider the estimator $W(\mathbf{X}) = X_1$, which has the pmf $p_{X_1}(x) = \theta^x(1-\theta)^{1-x}$, where $x \in \{0,1\}$. Then,

$$\mathrm{E}[X_1] = \sum_{x=0}^{1} x \cdot p_{X_1}(x|\theta) = 0 \cdot \theta^0(1-\theta)^{1-0} + 1 \cdot \theta^1(1-\theta)^{1-1} = 0 + \theta = \theta,$$

so it follows that $W(\mathbf{X})$ is an unbiased estimator for θ. Then, it follows from theorem 9.39 that

$$\phi(T) = \mathrm{E}[W(\mathbf{X})|T(\mathbf{X})] = \mathrm{E}\left[X_1 | \sum_{i=1}^{n} X_i \right]$$

is a uniformly better unbiased estimator of $\tau(\theta)$ than $W(\mathbf{X})$. We will show that $\phi(T)$ is unbiased, beginning by finding an expression for $\phi(T)$. Noting that

$$\sum_{i=1}^{n} X_i \sim \mathrm{Binomial}(n, \theta),$$

we have

$$\phi(T) = \mathrm{E}\left[X_1 | \sum_{i=1}^{n} X_i \right]$$

$$
\begin{aligned}
&= \sum_{x_1=0}^{1} x_1 \cdot P\left(\{X_1 = x_1\} \mid \left\{\sum_{i=1}^{n} X_i = s\right\}\right) \\
&= 0 \cdot P\left(\{X_1 = 0\} \mid \left\{\sum_{i=1}^{n} X_i = s\right\}\right) + 1 \cdot P\left(\{X_1 = 1\} \mid \left\{\sum_{i=1}^{n} X_i = s\right\}\right) \\
&= P\left(\{X_1 = 1\} \mid \left\{\sum_{i=1}^{n} X_i = s\right\}\right) \\
&= \frac{P\left(\{X_1 = 1\} \cap \{\sum_{i=2}^{n} X_i = s-1\}\right)}{P\left(\{\sum_{i=1}^{n} X_i = s\}\right)} \\
&= \frac{P\left(\{X_1 = 1\}\right) \cdot P\left(\{\sum_{i=2}^{n} X_i = s-1\}\right)}{P\left(\{\sum_{i=1}^{n} X_i = s\}\right)} \\
&= \frac{\left[\theta^1 (1-\theta)^0\right]\left[\binom{n-1}{s-1}\theta^{s-1}(1-\theta)^{(n-1)-(s-1)}\right]}{\binom{n}{s}\theta^s (1-\theta)^{n-s}} \\
&= \frac{\theta^s (1-\theta)^{n-s}}{\theta^s (1-\theta)^{n-s}} \frac{\frac{(n-1)!}{(s-1)!((n-1)-(s-1))!}}{\frac{n!}{s!(n-s)!}} \\
&= \frac{s!\,(n-1)!\,(n-s)!}{n!\,(s-1)!\,(n-s)!} \\
&= \frac{s\,(s-1)!\,(n-1)!}{n\,(n-1)!\,(s-1)!} \\
&= \frac{s}{n} \\
&= \frac{1}{n}\sum_{i=1}^{n} X_i \\
&= \bar{X}.
\end{aligned}
$$

Then, the expected value of $\phi(T)$ is

$$
\mathrm{E}\left[\phi(T)\right] = \mathrm{E}\left[\bar{X}\right] = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\,\mathrm{E}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] = \frac{1}{n}\left(n\theta\right) = \theta,
$$

so it follows that $\phi(T)$ is unbiased for θ. The variance of $\phi(T)$ is given by

$$
\begin{aligned}
\mathrm{Var}\left(\phi(T)\right) &= \mathrm{Var}\left(\bar{X}\right) \\
&= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left(X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\theta\left(1-\theta\right) \\
&= \frac{n\theta\left(1-\theta\right)}{n^2} \\
&= \frac{\theta\left(1-\theta\right)}{n}.
\end{aligned}
$$

Noting that the variance of $W(\mathbf{X}) = X_1$ is given by

$$\operatorname{Var}(W(\mathbf{X})) = \operatorname{Var}(X_1) = \theta(1-\theta),$$

we will verify that $\operatorname{Var}(\phi(T)) \leq \operatorname{Var}(W)$, i.e., that $\phi(T)$ is a uniformly better estimator than $W(\mathbf{X})$. We have $n \geq 1$ and $\theta \in (0,1)$, so it follows that

$$\operatorname{Var}(\phi(T)) \leq \operatorname{Var}(W) \Leftrightarrow \frac{\theta(1-\theta)}{n} \leq \theta(1-\theta) \Leftrightarrow \theta(1-\theta) \leq n\theta(1-\theta) \Leftrightarrow 1 \leq n,$$

and we see that the variance of $\phi(T)$ is indeed less than or equal to the variance of $W(\mathbf{X})$. We will now check whether the variance of $\phi(T)$ attains the Cramér-Rao lower bound for the variance of an unbiased estimator $V(\mathbf{X})$ of θ. The CRLB is given by

$$\operatorname{Var}(V(\mathbf{X})) \geq \frac{[\tau'(\theta)]^2}{-\operatorname{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{X}|\theta)\right]}.$$

From example 8.13, the joint pmf of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i},$$

so that the denominator above is

$$\begin{aligned}
-\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{X}|\theta)\right] &= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\log\theta^{\sum_{i=1}^n X_i}(1-\theta)^{n-\sum_{i=1}^n X_i}\right] \\
&= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\left[\log(\theta)\sum_{i=1}^n X_i + \log(1-\theta)\left(n - \sum_{i=1}^n X_i\right)\right]\right] \\
&= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\left[n\bar{X}\log\theta + n\left(1-\bar{X}\right)\log(1-\theta)\right]\right] \\
&= -\operatorname{E}\left[\frac{\partial}{\partial\theta}\left[\frac{n\bar{X}}{\theta} - \frac{n\left(1-\bar{X}\right)}{1-\theta}\right]\right] \\
&= -\operatorname{E}\left[-\frac{n\bar{X}}{\theta^2} + \frac{n\left(1-\bar{X}\right)}{(1-\theta)^2}\right] \\
&= -\left[\operatorname{E}\left[-\frac{n\bar{X}}{\theta^2}\right] + \operatorname{E}\left[\frac{n\left(1-\bar{X}\right)}{(1-\theta)^2}\right]\right] \\
&= -\left[-\frac{n}{\theta^2}\operatorname{E}\left[\bar{X}\right] + \frac{n}{(1-\theta)^2}\operatorname{E}\left[1-\bar{X}\right]\right] \\
&= -\left[-\frac{n\theta}{\theta^2} + \frac{n}{(1-\theta)^2}\left(\operatorname{E}[1] - \operatorname{E}\left[\bar{X}\right]\right)\right] \\
&= -\left[-\frac{n}{\theta} + \frac{n(1-\theta)}{(1-\theta)^2}\right] \\
&= -\left[-\frac{n}{\theta} + \frac{n}{1-\theta}\right] \\
&= -\left[-\frac{n(1-\theta)+n\theta}{\theta(1-\theta)}\right] \\
&= \frac{n-n\theta+n\theta}{\theta(1-\theta)} \\
&= \frac{n}{\theta(1-\theta)}.
\end{aligned}$$

Then, we have $\tau(\theta) = \theta$, so that $\tau'(\theta) = 1$, so that the Cramér-Rao lower bound is given by

$$\operatorname{Var}(V(\mathbf{X})) \geq \frac{[\tau'(\theta)]^2}{-\operatorname{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\log f(\mathbf{X}|\theta)\right]} = \frac{1}{\frac{n}{\theta(1-\theta)}} = \frac{\theta(1-\theta)}{n},$$

and we see that $\phi(T)$ attains this lower bound.

THEOREM 9.41. *If W is a best unbiased estimator of $\tau(\theta)$, then W is unique. (This is Theorem 7.3.19 from Casella & Berger; the following proof is given there.)*

PROOF. Suppose W' is another best unbiased estimator, and consider the estimator $W^* = \frac{1}{2}(W + W')$. Note that

$$\mathrm{E}_\theta[W^*] = \mathrm{E}_\theta\left[\frac{1}{2}(W+W')\right] = \frac{1}{2}(\mathrm{E}_\theta[W] + \mathrm{E}_\theta[W']) = \frac{1}{2}(\tau(\theta)+\tau(\theta)) = \frac{1}{2}(2\tau(\theta)) = \tau(\theta)$$

and

$$\begin{aligned}\mathrm{Var}_\theta(W^*) &= \mathrm{Var}_\theta\left(\frac{1}{2}W + \frac{1}{2}W'\right)\\ &= \left(\frac{1}{2}\right)^2 \mathrm{Var}(W) + \left(\frac{1}{2}\right)^2 \mathrm{Var}(W') + 2\left(\frac{1}{2}\cdot\frac{1}{2}\right)\mathrm{Cov}(W,W')\\ &= \frac{1}{4}\mathrm{Var}(W) + \frac{1}{4}\mathrm{Var}(W') + \frac{1}{2}\mathrm{Cov}(W,W').\end{aligned}$$

From theorem 4.12, we have

$$|\mathrm{Cov}(W,W')| \le [\mathrm{Var}(W)\,\mathrm{Var}(W')]^{1/2}.$$

Because W and W' are both best unbiased estimators, it follows that $\mathrm{Var}(W) = \mathrm{Var}(W')$, so that we have

$$|\mathrm{Cov}(W,W')| \le \left[\mathrm{Var}(W)^2\right]^{1/2} = \mathrm{Var}(W)$$

and

$$\begin{aligned}\mathrm{Var}_\theta(W^*) &= \frac{1}{4}\mathrm{Var}(W) + \frac{1}{4}\mathrm{Var}(W') + \frac{1}{2}\mathrm{Cov}(W,W')\\ &\le \frac{1}{4}\mathrm{Var}(W) + \frac{1}{4}\mathrm{Var}(W) + \frac{1}{2}\mathrm{Var}(W)\\ &= \mathrm{Var}(W).\end{aligned}$$

But if the above inequality is strict, then the best unbiasedness of W is contradicted, so we must have equality for all θ. Since the inequality is an application of Cauchy-Schwarz, we can have the equality only if $W' = a(\theta)W + b(\theta)$. Now using properties of covariance, we have

$$\mathrm{Cov}_\theta(W,W') = \mathrm{Cov}_\theta(W, a(\theta)W + b(\theta)) = \mathrm{Cov}_\theta(W, a(\theta)W) = a(\theta)\mathrm{Var}_\theta(W),$$

but $\mathrm{Cov}_\theta(W,W') = \mathrm{Var}_\theta(W)$ since we had equality above. Hence $a(\theta) = 1$ and, since $\mathrm{E}_\theta[W'] = \tau(\theta)$, we must have $b(\theta) = 0$ and $W = W'$, showing that W is unique. □

THEOREM 9.42. *If $\mathrm{E}_\theta[W] = \tau(\theta)$, W is the best unbiased estimator of $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0. (This is Theorem 7.3.20 from Casella & Berger; the following proof is given there.)*

PROOF. [proof goes here] □

THEOREM 9.43 (Lehmann-Scheffé Theorem). *Let T be a complete sufficient statistic for a parameter θ, and let $\phi(T)$ be any estimator based only on T. Then $\phi(T)$ is the unique best unbiased estimator of its expected value. (This is Theorem 7.3.23 from Casella & Berger.)*

PROOF. [proof goes here] □

If T is complete and sufficient, and $h(X)$ is unbiased for $\tau(\theta)$, then $\phi(T) = \mathrm{E}[h(X)|T]$ is the best unbiased estimator of $\tau(\theta)$.

EXAMPLE 9.44. Let $X_1, \ldots, X_n$ be iid with pdf

$$f(x|\theta) = \frac{2x}{\theta^2}, \quad 0 < x \le \theta.$$

(1) Show that the MLE of θ is complete and sufficient.

The likelihood function is given by

$$\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(x_i|\theta) = \prod_{i=1}^{n} \frac{2x_i}{\theta^2} I_{\{0<x_i\leq\theta\}} = \frac{2^n}{\theta^{2n}} \left(\prod_{i=1}^{n} x_i\right) I_{\left\{0<x_{(n)}\leq\theta\right\}},$$

which can be written as

$$\mathcal{L}(\theta|\mathbf{x}) = \begin{cases} \frac{2^n}{\theta^{2n}} \prod_{i=1}^{n} x_i, & \theta \geq x_{(n)} \\ 0, & \text{otherwise} \end{cases}.$$

When $\theta \geq X_{(n)}$, i.e., when $\mathcal{L}(\theta|\mathbf{x}) > 0$, $\mathcal{L}(\theta|\mathbf{x})$ is decreasing in θ. It follows that $\mathcal{L}(\theta|\mathbf{x})$ attains its maximum at $\theta = X_{(n)}$, i.e., the MLE of θ is $\hat{\theta} = X_{(n)}$. We will now show that $\hat{\theta}$ is sufficient. We can write the joint density of $\mathbf{X} = X_1, \ldots, X_n$ as

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \underbrace{\left(\prod_{i=1}^{n} x_i\right)}_{h(\mathbf{x})} \underbrace{\frac{2^n}{\theta^{2n}} I_{\left\{0<x_{(n)}\leq\theta\right\}}}_{g(T(\mathbf{x})|\theta)},$$

so it follows from theorem 8.11 that $T(\mathbf{X}) = X_{(n)}$ is sufficient for θ. We will now find a pdf for $T(\mathbf{X})$. The cdf of X is given by

$$F_X(t|\theta) = \int_0^t f_X(x|\theta)\,\mathrm{d}\,x = \int_0^t \frac{2x}{\theta^2}\,\mathrm{d}\,x = \left[\frac{x^2}{\theta^2}\right|_0^t = \frac{t^2}{\theta^2},$$

so it follows from theorem 5.6 that a pdf for $T(\mathbf{X}) = X_{(n)}$ is given by

$$\begin{aligned} f_{X_{(n)}}(x) &= \frac{n!}{(n-1)!\,(n-n)!} f_X(x) \left[F_X(x)\right]^{n-1} \left[1 - F_X(x)\right]^{n-n} \\ &= \frac{n\,(n-1)!}{(n-1)!0!} \left[\frac{2x}{\theta^2}\right] \left[\frac{x^2}{\theta^2}\right]^{n-1} \left[1 - \frac{x^2}{\theta^2}\right]^0 \\ &= \frac{2nx}{\theta^2} \left(\frac{x^{2n-2}}{\theta^{2n-2}}\right) \\ &= \frac{2nx^{2n-1}}{\theta^{2n}} \end{aligned}$$

where $0 < x \leq \theta$ and $f_{X_{(n)}}(x) = 0$ otherwise. To show that $\hat{\theta}$ is complete, suppose that $g(T(\mathbf{X}))$ is a function satisfying $\mathrm{E}\left[g(T(\mathbf{X}))\right] = 0$, so that we have

$$0 = \mathrm{E}\left[g(T(\mathbf{X}))\right] = \mathrm{E}\left[g\left(X_{(n)}\right)\right] = \int_0^\theta g(x) \cdot f_{X_{(n)}}(x)\,\mathrm{d}\,x = \int_0^\theta g(x)\,\frac{2nx^{2n-1}}{\theta^{2n}}\,\mathrm{d}\,x.$$

By assumption, $\mathrm{E}\left[g(T)\right] = 0$, i.e., it is constant as a function of θ, so that its derivative with respect to θ is zero. Then, we have

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta} \mathrm{E}\left[g(T(\mathbf{X}))\right] \\ &= \frac{\partial}{\partial\theta} \int_0^\theta g(x)\,\frac{2nx^{2n-1}}{\theta^{2n}}\,\mathrm{d}\,x \\ &= \frac{\partial}{\partial\theta} \left[\frac{2n}{\theta^{2n}} \int_0^\theta g(x) \cdot x^{2n-1}\,\mathrm{d}\,x\right] \\ &= 2n\theta^{-2n-1}(-2n) \int_0^\theta g(x) \cdot x^{2n-1}\,\mathrm{d}\,x + \frac{2n}{\theta^{2n}} \frac{\partial}{\partial\theta} \int_0^\theta g(x) \cdot x^{2n-1}\,\mathrm{d}\,x \\ &= -\frac{2n}{\theta} \int_0^\theta g(x)\,\frac{2nx^{2n-1}}{\theta^{2n}}\,\mathrm{d}\,x + \frac{2n}{\theta^{2n}} \left[g(\theta) \cdot \theta^{2n-1}\right] \\ &= -\frac{2n}{\theta} \cdot 0 + \frac{2n}{\theta} g(\theta) \end{aligned}$$

$$= \frac{2n}{\theta} g(\theta).$$

We have $n, \theta > 0$, so it follows that $\mathrm{E}[g(T)] = 0$ if and only if $g(\theta) = 0$ for all θ, i.e.,

$$\mathrm{E}[g(T)] = 0 \quad \forall\theta \implies P(\{g(T) = 0\}) = 1 \quad \forall\theta,$$

so by definition $T(\mathbf{X})$ is a complete statistic.

(2) Find the UMVUE of θ.

The expected value of $\hat{\theta} = X_{(n)}$ is given by

$$\begin{aligned}
\mathrm{E}\left[\hat{\theta}\right] &= \mathrm{E}\left[X_{(n)}\right] \\
&= \int_0^\theta x \cdot f_{X_{(n)}}(x)\,\mathrm{d}x \\
&= \int_0^\theta x \cdot \frac{2nx^{2n-1}}{\theta^{2n}}\,\mathrm{d}x \\
&= \frac{2n}{\theta^{2n}} \int_0^\theta x^{2n}\,\mathrm{d}x \\
&= \frac{2n}{\theta^{2n}} \left[\frac{1}{2n+1} x^{2n+1}\right|_0^\theta \\
&= \frac{2n}{\theta^{2n}} \left[\frac{\theta^{2n+1}}{2n+1} - 0\right] \\
&= \frac{2n}{2n+1}\theta,
\end{aligned}$$

so it follows that

$$\tilde{\theta} = \frac{2n+1}{2n} X_{(n)}$$

is unbiased for θ. The estimator $\tilde{\theta}$ is based only on $T(\mathbf{X}) = X_{(n)}$, which is a complete sufficient statistic, so it follows from theorem 9.43 that $\tilde{\theta}$ is the UMVUE for θ.

(3) Show that the variance of the UMVUE of θ is less than the Cramér-Rao lower bound. Why is it so?

The variance of $\tilde{\theta}$ is given by

$$\mathrm{Var}\left(\tilde{\theta}\right) = \mathrm{Var}\left(\frac{2n+1}{2n} X_{(n)}\right) = \left(\frac{2n+1}{2n}\right)^2 \mathrm{Var}\left(X_{(n)}\right).$$

The expected value of $X_{(n)}^2$ is given by

$$\begin{aligned}
\mathrm{E}\left[X_{(n)}^2\right] &= \int_0^\theta x^2 \cdot f_{X_{(n)}}(x)\,\mathrm{d}x \\
&= \int_0^\theta x^2 \cdot \frac{2nx^{2n-1}}{\theta^{2n}}\,\mathrm{d}x \\
&= \frac{2n}{\theta^{2n}} \int_0^\theta x^{2n+1}\,\mathrm{d}x \\
&= \frac{2n}{\theta^{2n}} \left[\frac{1}{2n+2} x^{2n+2}\right|_0^\theta \\
&= \frac{2n}{\theta^{2n}} \left[\frac{\theta^{2n+2}}{2n+2} - 0\right] \\
&= \frac{2n\theta^2}{2n+2} \\
&= \frac{2n\theta^2}{2(n+1)}
\end{aligned}$$

$$= \frac{n\theta^2}{n+1},$$

so that we have

$$\begin{aligned}
\operatorname{Var}\left(X_{(n)}\right) &= \operatorname{E}\left[X_{(n)}^2\right] - \left(\operatorname{E}\left[X_{(n)}\right]\right)^2 \\
&= \frac{n\theta^2}{n+1} - \left(\frac{2n\theta}{2n+1}\right)^2 \\
&= \frac{n\theta^2}{n+1} - \frac{4n^2\theta^2}{(2n+1)^2} \\
&= \frac{n\theta^2(2n+1)^2 - 4n^2\theta^2(n+1)}{(n+1)(2n+1)^2} \\
&= \frac{n\theta^2\left[4n^2+4n+1-4n(n+1)\right]}{(n+1)(2n+1)^2} \\
&= \frac{n\theta^2\left(4n^2+4n+1-4n^2-4n\right)}{(n+1)(2n+1)^2} \\
&= \frac{n\theta^2}{(n+1)(2n+1)^2}
\end{aligned}$$

and

$$\operatorname{Var}\left(\tilde{\theta}\right) = \left(\frac{2n+1}{2n}\right)^2 \operatorname{Var}\left(X_{(n)}\right) = \frac{(2n+1)^2}{4n^2}\frac{n\theta^2}{(n+1)(2n+1)^2} = \frac{\theta^2}{4n(n+1)}.$$

The Cramér-Rao lower bound for the variance of an unbiased estimator $W(\mathbf{X})$ of $\tau(\theta) = \theta$ that satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\,\theta}\operatorname{E}_\theta\left[W(\mathbf{X})\right] = \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\left[W(\mathbf{x})\,f(\mathbf{x}|\theta)\right]\mathrm{d}\,\mathbf{x} \quad \text{and} \quad \operatorname{Var}_\theta\left(W(\mathbf{X})\right) < \infty$$

is given by

$$\operatorname{Var}\left(W(\mathbf{X})\right) \geq \frac{\left[\tau'(\theta)\right]^2}{\operatorname{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]}.$$

In the case that $0 < X_{(n)} \leq \theta$, we have

$$\begin{aligned}
\operatorname{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right] &= \operatorname{E}\left[\left(\frac{\partial}{\partial\theta}\log\frac{2^n}{\theta^{2n}}\left(\prod_{i=1}^n X_i\right)\right)^2\right] \\
&= \operatorname{E}\left[\left(\frac{\partial}{\partial\theta}\left(\log\frac{2^n}{\theta^{2n}} + \log\prod_{i=1}^n X_i\right)\right)^2\right] \\
&= \operatorname{E}\left[\left(\frac{\partial}{\partial\theta}\left(n\log 2 - 2n\log\theta + \sum_{i=1}^n \log X_i\right)\right)^2\right] \\
&= \operatorname{E}\left[\left(0 - \frac{2n}{\theta} + 0\right)^2\right] \\
&= \operatorname{E}\left[\left(-\frac{2n}{\theta}\right)^2\right] \\
&= \operatorname{E}\left[\frac{4n^2}{\theta^2}\right] \\
&= \frac{4n^2}{\theta^2}.
\end{aligned}$$

We have $\tau(\theta) = \theta$, so that $\tau'(\theta) = 1$, so that the Cramér-Rao lower bound is given by

$$\operatorname{Var}(W(\mathbf{X})) \geq \frac{[\tau'(\theta)]^2}{\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log f(\mathbf{X}|\theta)\right)^2\right]} = \frac{1}{\frac{4n^2}{\theta^2}} = \frac{\theta^2}{4n^2}.$$

Noting that $n, \theta > 0$, we will check that the variance of $\tilde{\theta}$ is less than the CRLB.

$$\frac{\theta^2}{4n(n+1)} < \frac{\theta^2}{4n^2} \implies \left(\frac{\theta^2}{4n(n+1)}\right)\frac{4n}{\theta^2} < \left(\frac{\theta^2}{4n^2}\right)\frac{4n}{\theta^2} \implies \frac{1}{n+1} < \frac{1}{n} \implies n < n+1$$

We see that n is indeed less than $n+1$, so it follows that $\operatorname{Var}\left(\tilde{\theta}\right)$ is less than the Cramér-Rao lower bound. This occurs because $f(\mathbf{x}|\theta)$ does not satisfy the regularity conditions of the theorem, i.e., the interchange of differentiation and integration, because the bounds of the support of $f(\mathbf{x}|\theta)$ depend on θ.

9.2.4. Loss function optimality. Mean squared error is a special case of a function called a loss function. The study of the performance, and the optimality, of estimators evaluated through loss functions is a branch of decision theory. The loss function is a nonnegative function that generally increases as the distance between an estimator $\hat{\theta}$ and θ increases. If θ is real-valued, two commonly used loss functions are

$$\text{absolute error loss,} \qquad L\left(\theta, \hat{\theta}\right) = \left|\hat{\theta} - \theta\right|$$

and

$$\text{squared error loss,} \qquad L\left(\theta, \hat{\theta}\right) = \left(\hat{\theta} - \theta\right)^2.$$

In a loss function, the quality of an estimator is quantified in its risk function. For an estimator $T(\mathbf{x})$ of θ, the risk function is

$$\text{risk} = R(\theta, T) = \mathrm{E}_\theta\left[L(\theta, T)\right] = \int_{\mathcal{X}} L(\theta, T(\mathbf{x}))\, f(\mathbf{x}|\theta)\, \mathrm{d}\,\mathbf{x}.$$

At a given θ, the risk function is the average loss that will be incurred if the estimator $T(\mathbf{x})$ is used (averaged over $f(x|\theta)$). For two different estimators T_1 and T_2, if $R(\theta, T_1) < R(\theta, T_2)$ for all $\theta \in \Theta$, then T_1 is the preferred estimator. More typically, the two risk functions will cross. Then the judgment as to which estimator is better may not be so clear-cut.

EXAMPLE 9.45. Let $X \sim \mathcal{N}(\theta, 1)$ and assume we use a squared error loss. Consider two estimators

$$\hat{\theta}_1 = X \qquad \hat{\theta}_2 = 3.$$

The risk function for $\hat{\theta}_1$ is

$$R\left(\theta, \hat{\theta}_1\right) = \mathrm{E}\left[(X-\theta)^2\right] = \mathrm{E}\left[X^2 - 2\theta X + \theta^2\right] = \mathrm{E}\left[X^2\right] - 2\theta\,\mathrm{E}[X] + \mathrm{E}\left[\theta^2\right].$$

We have

$$\operatorname{Var}(X) = \mathrm{E}\left[X^2\right] - (\mathrm{E}[X])^2 \Leftrightarrow \mathrm{E}\left[X^2\right] = \operatorname{Var}(X) + (\mathrm{E}[X])^2 = 1 + \theta^2,$$

so that

$$R\left(\theta, \hat{\theta}_1\right) = \mathrm{E}\left[X^2\right] - 2\theta\,\mathrm{E}[X] + \mathrm{E}\left[\theta^2\right] = 1 + \theta^2 - 2\theta(\theta) + \theta^2 = 1 + 2\theta^2 - 2\theta^2 = 1.$$

The risk function for $\hat{\theta}_2$ is

$$R\left(\theta, \hat{\theta}_2\right) = \mathrm{E}\left[(3-\theta)^2\right] = (3-\theta)^2.$$

If $2 < \theta < 4$, then $R\left(\theta, \hat{\theta}_2\right) < R\left(\theta, \hat{\theta}_1\right)$, otherwise $R\left(\theta, \hat{\theta}_1\right) < R\left(\theta, \hat{\theta}_2\right)$. Neither estimator uniformly dominates the other.

Under squared error loss, the risk function is the MSE

$$R_T(\theta) = \mathrm{E}\left[(T(\mathbf{X}) - \theta)^2\right] = \operatorname{MSE}(T(\mathbf{X})) = \operatorname{Var}(T(\mathbf{X})) + [\operatorname{Bias}(T(\mathbf{X}))]^2.$$

The risk function for squared error loss clearly indicates that a good estimator should have both a small variance and a small bias. Restricting the set of allowable estimators to the set of unbiased estimators would not be typical in decision theory (in that case, minimizing the risk would just be minimizing the variance).

A decision theoretic analysis would be more comprehensive in that both the variance and bias are in the risk and will be considered simultaneously.

9.2.4.1. *Bayes risk.* In a Bayesian analysis, we would use the prior distribution to compute an average risk

$$\begin{aligned} r(T) &= \int_{\Theta} R(\theta, T)\, \pi(\theta)\, \mathrm{d}\,\theta \\ &= \int_{\Theta} \left[\int_{\mathcal{X}} L(\theta, T(\mathbf{x}))\, f(\mathbf{x}|\theta)\, \mathrm{d}\, x \right] \pi(\theta)\, \mathrm{d}\,\theta \\ &= \int_{\mathcal{X}} \left[\int_{\Theta} L(\theta, T(\mathbf{x}))\, \pi(\theta|\mathbf{x})\, \mathrm{d}\,\theta \right] f(\mathbf{x})\, \mathrm{d}\,\mathbf{x} \\ &= \int_{\mathcal{X}} r(T, \mathbf{x})\, f(\mathbf{x})\, \mathrm{d}\,\mathbf{x}, \end{aligned}$$

where $\pi(\theta|\mathbf{x})$ is the posterior distribution of θ and $f(\mathbf{x})$ is the marginal distribution of $\mathbf{X}$, i.e.,

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\, \pi(\theta)}{f(\mathbf{x})} \implies \pi(\theta)\, f(\mathbf{x}|\theta) = \pi(\theta|\mathbf{x})\, f(\mathbf{x}).$$

$r(T, \mathbf{X})$ is the expected value of the loss function with respect to the posterior distribution, called the posterior expected loss. $r(T, \mathbf{X})$ is a function only of $\mathbf{x}$ and not a function of θ. Thus, for each $\mathbf{x}$, if we choose the estimator $T(\mathbf{x})$ to minimize the posterior expected loss, we will minimize the Bayes risk.

EXAMPLE 9.46. Let $X_1, \ldots, X_n$ be iid Bernoulli (θ) random variables and consider a Beta (α, β) prior for θ.

(1) Find the Bayes estimator for a squared error loss.

The loss function is given by$L(\theta, T) = (T - \theta)^2$, so that the Bayes risk is given by

$$r(T) = \int_{\mathcal{X}} r(T, \mathbf{x})\, f(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

Then, the posterior expected loss is given by

$$r(T, \mathbf{x}) = \int_{\Theta} L(\theta, T(\mathbf{x}))\, \pi(\theta|\mathbf{x})\, \mathrm{d}\theta.$$

The density of the prior distribution is given by

$$\pi(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\, \Gamma(\beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1}, \qquad 0 < \theta < 1.$$

From example 8.13, the joint pmf of $\mathbf{X} = X_1, \ldots, X_n$ is given by

$$f_{\mathbf{X}}(\mathbf{x}|\theta) = \theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i},$$

so that the posterior distribution is given by

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{f(\mathbf{x}|\theta)\, \pi(\theta)}{\int_{\Theta} f(\mathbf{x}|\theta)\, \pi(\theta)\, \mathrm{d}\theta} \\ &= \frac{\left[\theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \right] \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \right]}{\int_0^1 \left[\theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i} \right] \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1} (1 - \theta)^{\beta - 1} \right] \mathrm{d}\theta} \\ &= \frac{\theta^{\sum_{i=1}^n x_i + \alpha - 1} (1 - \theta)^{n - \sum_{i=1}^n x_i + \beta - 1}}{\int_0^1 \theta^{\sum_{i=1}^n x_i + \alpha - 1} (1 - \theta)^{n - \sum_{i=1}^n x_i + \beta - 1} \mathrm{d}\theta}. \end{aligned}$$

Let $\alpha^* = \alpha + \sum_{i=1}^n x_i$ and let $\beta^* = \beta + n - \sum_{i=1}^n x_i$, so that we have

$$\begin{aligned} \pi(\theta|\mathbf{x}) &= \frac{\theta^{\alpha^* - 1} (1 - \theta)^{\beta^* - 1}}{\int_0^1 \theta^{\alpha^* - 1} (1 - \theta)^{\beta^* - 1} \mathrm{d}\theta} \\ &= \frac{\theta^{\alpha^* - 1} (1 - \theta)^{\beta^* - 1}}{\frac{\Gamma(\alpha^*)\Gamma(\beta^*)}{\Gamma(\alpha^* + \beta^*)} \int_0^1 \frac{\Gamma(\alpha^* + \beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)} \theta^{\alpha^* - 1} (1 - \theta)^{\beta^* - 1} \mathrm{d}\theta}. \end{aligned}$$

We recognize the integrand as the pdf of a Beta (α^*, β^*) random variable, which integrates to 1, so that we have

$$\pi(\theta|\mathbf{x}) = \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1},$$

i.e., $\pi(\theta|\mathbf{x}) \sim \text{Beta}(\alpha^*, \beta^*)$. Then, the posterior expected loss is given by

$$\begin{aligned}
r(T,\mathbf{x}) &= \int_\Theta L(\theta, T(\mathbf{x}))\,\pi(\theta|\mathbf{x})\,\mathrm{d}\theta \\
&= \int_0^1 (t-\theta)^2 \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta \\
&= \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\int_0^1 \left(t^2-2\theta t+\theta^2\right)\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta \\
&= \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\left[t^2\int_0^1 \theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta\right. \\
&\quad \left. -2t\int_0^1 \theta^{\alpha^*}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta + \int_0^1 \theta^{\alpha^*+1}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta\right] \\
&= \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\left[t^2\frac{\Gamma(\alpha^*)\Gamma(\beta^*)}{\Gamma(\alpha^*+\beta^*)}\int_0^1 \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\theta^{\alpha^*-1}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta\right. \\
&\quad -2t\frac{\Gamma(\alpha^*+1)\Gamma(\beta^*)}{\Gamma(\alpha^*+1+\beta^*)}\int_0^1 \frac{\Gamma(\alpha^*+1+\beta^*)}{\Gamma(\alpha^*+1)\Gamma(\beta^*)}\theta^{\alpha^*}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta \\
&\quad \left. +\frac{\Gamma(\alpha^*+2)\Gamma(\beta^*)}{\Gamma(\alpha^*+2+\beta^*)}\int_0^1 \frac{\Gamma(\alpha^*+2+\beta^*)}{\Gamma(\alpha^*+2)\Gamma(\beta^*)}\theta^{\alpha^*+1}(1-\theta)^{\beta^*-1}\,\mathrm{d}\theta\right] \\
&= \frac{\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)\Gamma(\beta^*)}\left[t^2\frac{\Gamma(\alpha^*)\Gamma(\beta^*)}{\Gamma(\alpha^*+\beta^*)}\cdot 1 - 2t\frac{\Gamma(\alpha^*+1)\Gamma(\beta^*)}{\Gamma(\alpha^*+1+\beta^*)}\cdot 1 + \frac{\Gamma(\alpha^*+2)\Gamma(\beta^*)}{\Gamma(\alpha^*+2+\beta^*)}\cdot 1\right] \\
&= t^2 - 2t\frac{\Gamma(\alpha^*+\beta^*)\Gamma(\alpha^*+1)}{\Gamma(\alpha^*)\Gamma(\alpha^*+1+\beta^*)} + \frac{\Gamma(\alpha^*+\beta^*)\Gamma(\alpha^*+2)}{\Gamma(\alpha^*)\Gamma(\alpha^*+2+\beta^*)} \\
&= t^2 - 2t\frac{\alpha^*\Gamma(\alpha^*)\Gamma(\alpha^*+\beta^*)}{\Gamma(\alpha^*)(\alpha^*+\beta^*)\Gamma(\alpha^*+\beta^*)} + \frac{\Gamma(\alpha^*+\beta^*)(\alpha^*+1)\Gamma(\alpha^*+1)}{\Gamma(\alpha^*)(\alpha^*+\beta^*+1)\Gamma(\alpha^*+\beta^*+1)} \\
&= t^2 - 2t\frac{\alpha^*}{\alpha^*+\beta^*} + \frac{\Gamma(\alpha^*+\beta^*)(\alpha^*+1)\alpha^*\Gamma(\alpha^*)}{\Gamma(\alpha^*)(\alpha^*+\beta^*+1)(\alpha^*+\beta^*)\Gamma(\alpha^*+\beta^*)} \\
&= t^2 - 2t\frac{\alpha^*}{\alpha^*+\beta^*} + \frac{\alpha^*(\alpha^*+1)}{(\alpha^*+\beta^*)(\alpha^*+\beta^*+1)}.
\end{aligned}$$

Taking the derivative with respect to t gives

$$\frac{\partial}{\partial t}r(T,\mathbf{x}) = \frac{\partial}{\partial t}\left[t^2 - 2t\frac{\alpha^*}{\alpha^*+\beta^*} + \frac{\alpha^*(\alpha^*+1)}{(\alpha^*+\beta^*)(\alpha^*+\beta^*+1)}\right] = 2t - 2\frac{\alpha^*}{\alpha^*+\beta^*}.$$

Setting this equal to zero, we have

$$2\hat{t} = 2\frac{\alpha^*}{\alpha^*+\beta^*} \Leftrightarrow \hat{t} = \frac{\alpha^*}{\alpha^*+\beta^*} = \frac{\alpha+\sum_{i=1}^n x_i}{\alpha+\sum_{i=1}^n x_i + (\beta+n-\sum_{i=1}^n x_i)} = \frac{\alpha+\sum_{i=1}^n x_i}{\alpha+\beta+n},$$

so it follows that the Bayes estimator of θ with respect to a Beta (α, β) prior distribution is $\hat{t} = (\alpha+\sum_{i=1}^n X_i)/(\alpha+\beta+n)$.

(2) Show that this estimator corresponds to the posterior mean.

We found that the posterior distribution $\pi(\theta|\mathbf{x})$ has the distribution of a Beta (α^*, β^*) random variable. The expected value (mean) of a Beta (γ, ψ) random variable is given by $\gamma/(\gamma+\psi)$, so it follows that

$$\mathrm{E}[\pi(\theta|\mathbf{x})] = \frac{\alpha^*}{\alpha^*+\beta^*} = \frac{\alpha+\sum_{i=1}^n x_i}{\alpha+\sum_{i=1}^n x_i + (\beta+n-\sum_{i=1}^n x_i)} = \frac{\alpha+\sum_{i=1}^n x_i}{\alpha+\beta+n}$$

is the posterior mean, which agrees with the estimator we found.

We can find an explicit formula for the Bayes estimator for specific loss functions. For a squared error loss function, i.e., $L\left(\theta,\hat{\theta}\right)=\left(\hat{\theta}-\theta\right)^2$, the Bayes estimator is the posterior mean. For an absolute error loss function, i.e., $L\left(\theta,\hat{\theta}\right)=\left|\theta-\hat{\theta}\right|$, the Bayes estimator is the posterior median. For a zero-one loss function, i.e.,

$$L\left(\theta,\hat{\theta}\right)=\begin{cases}0, & \text{if } \theta=\hat{\theta},\\ 1, & \text{if } \theta\neq\hat{\theta}\end{cases},$$

the Bayes estimator is the posterior mode.

CHAPTER 10

Hypothesis testing

DEFINITION 10.1. A *hypothesis* is a statement about a population parameter.

DEFINITION 10.2. The two complementary hypotheses in a hypothesis testing problem are called the *null hypothesis* and the *alternative hypothesis*. They are denoted by H_0 and H_1, respectively.

If θ denotes a population parameter, the general format of the null and alternative hypotheses if $H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, which Θ_0 is some subset of the parameter space and Θ_0^c is its complement. Hypotheses that specify only one possible distribution for the sample $\mathbf{X}$ are called simple hypotheses. In most realistic problems, the hypotheses of interest specify more than one possible distribtion for the sample. Such hypotheses are called *composite hypotheses.* Suppose $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Then

$$\begin{aligned}
&H_0 : \theta = \frac{1}{2}, \quad H_1 : \theta = \frac{1}{4} && \text{are called simple hypotheses,}\\
&H_0 : \theta \leq \frac{1}{2}, \quad H_1 : \theta > \frac{1}{2} && \text{are called one-sided composite hypotheses, and}\\
&H_0 : \theta = \frac{1}{2}, \quad H_1 : \theta \neq \frac{1}{2} && \text{are called two-sided composite hypotheses.}
\end{aligned}$$

DEFINITION 10.3. A *hypothesis testing procedure* or *hypothesis test* is a rule that specifies:

(1) For which sample values the decision is made to accept H_0 as true.
(2) For which sample values H_0 is rejected and H_1 is accepted as true.

The subset of the sample space for which H_0 will be rejected is called the *rejection region* or *critical region.* The complement of the rejection region is called the *acceptance region.*

Typically, a hypothesis test is specified in terms of a *test statistic* $W(\mathbf{X})$, a function of the sample. If the rejection region is denoted as R, then a test may reject H_0 if $W(\mathbf{X}) \in R$ and fail to reject H_0 if $W(\mathbf{X}) \in R^c$.

10.1. Methods of finding tests

10.1.1. Likelihood ratio tests. The likelihood ratio method of hypothesis testing is related to maximum likelihood estimators.

DEFINITION 10.4. The *likelihood ratio test statistic* for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}.$$

A *likelihood ratio test* (LRT) is any test that has a rejection region of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

The numerator of the LRT statistic, $\sup_{\Theta_0} L(\theta|\mathbf{x})$, is the maximized likelihood under the null hypothesis. The denominator, $\sup_{\Theta} L(\theta|\mathbf{x})$, is the unrestricted maximized likelihood.

EXAMPLE 10.5 (Normal LRT with known variance). Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} N(\theta, 1)$. Specify the LRT for

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

The likelihood function is given by

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i|\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi \cdot 1}} e^{-(x_i-\theta)^2/(2\cdot 1)} = (2\pi)^{-n/2} e^{\sum_{i=1}^{n} -(x_i-\theta)^2/2}.$$

The LRT test statistic is

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}.$$

Under the null hypothesis, we have $\Theta_0 = \{\theta : \theta = \theta_0\}$, i.e., only one value of θ is specified by H_0, so that the numerator is

$$\sup_{\Theta_0} L(\theta|\mathbf{x}) = \sup_{\theta=\theta_0} (2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right\} = (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_0)^2\right\}.$$

Under the alternative hypothesis, we have $\Theta = \{\theta : \theta \neq \theta_0\}$, so we must maximize $L(\theta|\mathbf{x})$ over the parameter space $\theta \in \mathbb{R}$, i.e., find the MLE of θ. The log-likelihood is given by

$$\begin{aligned}
\log L(\theta|\mathbf{x}) &= \log(2\pi)^{-n/2} e^{\sum_{i=1}^{n} -(x_i-\theta)^2/2} \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2 \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}\left(x_i^2 - 2\theta x_i + \theta^2\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\left(\sum_{i=1}^{n}x_i^2 - \sum_{i=1}^{n}2\theta x_i + \sum_{i=1}^{n}\theta^2\right) \\
&= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}x_i^2 + \theta\sum_{i=1}^{n}x_i - \frac{n\theta^2}{2}.
\end{aligned}$$

Taking the derivative with respect to θ gives

$$\begin{aligned}
\frac{\partial}{\partial\theta}\log L(\theta|\mathbf{x}) &= \frac{\partial}{\partial\theta}\left[-\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}x_i^2 + \theta\sum_{i=1}^{n}x_i - \frac{n\theta^2}{2}\right] \\
&= 0 - 0 + \sum_{i=1}^{n}x_i - n\theta \\
&= \sum_{i=1}^{n}x_i - n\theta.
\end{aligned}$$

Setting this equal to zero, we have

$$n\hat{\theta} = \sum_{i=1}^{n}x_i \implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}.$$

We will evaluate the second derivative of $\log L(\theta|\mathbf{x})$ with respect to θ at $\theta = \hat{\theta}$ to verify that this is a maximum.

$$\frac{\partial^2}{\partial\theta^2}\left[\log L(\theta|\mathbf{x})\right]_{\theta=\hat{\theta}} = \frac{\partial}{\partial\theta}\left[\sum_{i=1}^{n}x_i - n\theta\right|_{\theta=\hat{\theta}} = \left[0-n\right]_{\theta=\hat{\theta}} = -n$$

We have $n > 0$, so that $-n < 0$. It follows that $\hat{\theta} = \bar{X}$ is the MLE of θ, so that

$$\sup_{\Theta} L(\theta|\mathbf{x}) = \sup_{\theta=\bar{x}} (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta)^2\right\} = (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}.$$

Then, we have

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})} \\
&= \frac{(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\theta_0)^2\right\}}{(2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2\right\}}
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i-\theta_0\right)^2\right\}\exp\left\{\frac{1}{2}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right\}\\
&= \exp\left\{\frac{1}{2}\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2-\sum_{i=1}^{n}\left(x_i-\theta_0\right)^2\right]\right\}\\
&= \exp\left\{\frac{1}{2}\left[\sum_{i=1}^{n}\left(x_i^2-2\bar{x}x_i+\bar{x}^2\right)-\sum_{i=1}^{n}\left(x_i^2-2\theta_0 x_i+\theta_0^2\right)\right]\right\}\\
&= \exp\left\{\frac{1}{2}\left(\sum_{i=1}^{n}x_i^2-2\bar{x}\sum_{i=1}^{n}x_i+n\bar{x}^2-\sum_{i=1}^{n}x_i^2+2\theta_0\sum_{i=1}^{n}x_i-n\theta_0^2\right)\right\}\\
&= \exp\left\{\frac{1}{2}\left(-2n\bar{x}^2+n\bar{x}^2+2\theta_0 n\bar{x}-n\theta_0^2\right)\right\}\\
&= \exp\left\{-\frac{n}{2}\left(\theta_0^2-2\theta_0\bar{x}+\bar{x}^2\right)\right\}\\
&= \exp\left\{-\frac{n}{2}\left(\theta_0-\bar{x}\right)^2\right\}.
\end{aligned}
$$

Then, the likelihood ratio test rejects H_0 if

$$
c \geq \lambda(\mathbf{x}) = \exp\left\{-\frac{n}{2}\left(\theta_0-\bar{x}\right)^2\right\} \Leftrightarrow \log c \geq -\frac{n}{2}\left(\theta_0-\bar{x}\right)^2 \Leftrightarrow -\frac{2}{n}\log c \leq \left(\theta_0-\bar{x}\right)^2 \Leftrightarrow \left|\theta_0-\bar{x}\right| \geq \sqrt{-\frac{2}{n}\log c}
$$

for some $c \in (0,1]$.

EXAMPLE 10.6 (Exponential LRT). Let $X_1,\ldots,X_n \overset{\text{iid}}{\sim} \text{Exp}(\theta)$, i.e.,

$$
X \sim f(x|\theta) = \frac{1}{\theta}\mathrm{e}^{-x/\theta}.
$$

Specify the LRT for

$$
H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0.
$$

The likelihood function is given by

$$
\mathcal{L}(\theta|\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i|\theta) = \prod_{i=1}^{n}\frac{1}{\theta}\mathrm{e}^{-x_i/\theta} = \theta^{-n}\mathrm{e}^{\sum_{i=1}^{n}-x_i/\theta},
$$

so that the log-likelihood is given by

$$
\log\mathcal{L}(\theta|\mathbf{x}) = \log\theta^{-n}\mathrm{e}^{\sum_{i=1}^{n}-x_i/\theta} = -n\log\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i.
$$

Taking the derivative with respect to θ gives

$$
\frac{\partial}{\partial\theta}\log\mathcal{L}(\theta|\mathbf{x}) = \frac{\partial}{\partial\theta}\left[-n\log\theta-\frac{1}{\theta}\sum_{i=1}^{n}x_i\right] = -\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i.
$$

Setting this equal to zero, we have

$$
\frac{n}{\hat{\theta}} = \frac{1}{\hat{\theta}^2}\sum_{i=1}^{n}x_i \implies n\hat{\theta}^2 = \hat{\theta}\sum_{i=1}^{n}x_i \implies \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}.
$$

We will evaluate the second derivative of $\log\mathcal{L}(\theta|\mathbf{x})$ with respect to θ at $\theta=\hat{\theta}$ to verify that this is a maximum.

$$
\frac{\partial^2}{\partial\theta^2}\left[\mathcal{L}(\theta|\mathbf{x})\right]_{\theta=\hat{\theta}} = \frac{\partial}{\partial\theta}\left[-\frac{n}{\theta}+\frac{1}{\theta^2}\sum_{i=1}^{n}x_i\right|_{\theta=\hat{\theta}} = \left[\frac{n}{\theta^2}-\frac{2}{\theta^3}\sum_{i=1}^{n}x_i\right|_{\theta=\hat{\theta}} = \frac{n}{\bar{x}^2}-\frac{2}{\bar{x}^3}n\bar{x} = -\frac{n}{\bar{x}^2}
$$

We have $n > 0$ and $x \geq 0$, so that $\bar{x}^2 \geq 0$, so that $-n/\bar{x}^2 < 0$ in the case that at least one $x_i > 0$. It follows that $\hat{\theta} = \bar{X}$ is the MLE for θ. The derivative of the log-likelihood can be written as

$$\frac{\partial}{\partial\theta} \log \mathcal{L}(\theta|\mathbf{x}) = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = \frac{n\bar{x}}{\theta^2} - \frac{n}{\theta} = \frac{n}{\theta}\left(\frac{\bar{x}}{\theta} - 1\right).$$

The log-likelihood function attains its maximum at $\theta = \bar{x}$, so in the case that $\theta < \bar{x}$, we will have $\bar{x}/\theta > 1$, so that $(\bar{x}/\theta) - 1 > 0$, i.e., the derivative will be positive, so $\log L(\theta|\mathbf{x})$ is increasing in θ. In the case that $\theta > \bar{x}$, we will have $\bar{x}/\theta < 1$, so that $(\bar{x}/\theta) - 1 < 0$, i.e., the derivative will be negative, so $\log L(\theta|\mathbf{x})$ is decreasing in θ. Under the null hypothesis, we have $\Theta_0 = \{\theta : \theta \leq \theta_0\}$. If $\theta_0 \leq \bar{x}$, then $L(\theta|\mathbf{x})$ will be maximized at $\theta = \theta_0$, i.e.,

$$\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x}) = \sup_{\theta=\theta_0} \theta^{-n} \mathrm{e}^{\sum_{i=1}^{n} -x_i/\theta} = \theta_0^{-n} \mathrm{e}^{-\sum_{i=1}^{n} x_i/\theta_0}.$$

If $\theta_0 > \bar{x}$, then $L(\theta|\mathbf{x})$ will be maximized at $\theta = \bar{x}$, i.e.,

$$\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x}) = \sup_{\theta=\bar{x}} \theta^{-n} \mathrm{e}^{\sum_{i=1}^{n} -x_i/\theta} = \bar{x}^{-n} \mathrm{e}^{-\sum_{i=1}^{n} x_i/\bar{x}}.$$

The LRT test statistic is given by

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\Theta} \mathcal{L}(\theta|\mathbf{x})},$$

so in the case that $\theta_0 > \bar{x}$, we have

$$\frac{\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\Theta} \mathcal{L}(\theta|\mathbf{x})} = \frac{\bar{x}^{-n} \mathrm{e}^{-\sum_{i=1}^{n} x_i/\bar{x}}}{\bar{x}^{-n} \mathrm{e}^{-\sum_{i=1}^{n} x_i/\bar{x}}} = 1.$$

In the case that $\theta_0 \leq \bar{x}$, we have

$$\begin{aligned}
\frac{\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\Theta} \mathcal{L}(\theta|\mathbf{x})} &= \frac{\theta_0^{-n} \exp\left\{-\frac{1}{\theta_0}\sum_{i=1}^{n} x_i\right\}}{\bar{x}^{-n} \exp\left\{-\frac{1}{\bar{x}}\sum_{i=1}^{n} x_i\right\}} \\
&= \left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left\{-\frac{1}{\theta_0}\sum_{i=1}^{n} x_i + \frac{1}{\bar{x}}\sum_{i=1}^{n} x_i\right\} \\
&= \left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left\{-\frac{n\bar{x}}{\theta_0} + \frac{n\bar{x}}{\bar{x}}\right\} \\
&= \left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left\{-n\left(\frac{\bar{x}}{\theta_0} - 1\right)\right\},
\end{aligned}$$

i.e.,

$$\lambda(\mathbf{x}) = \begin{cases} 1, & \theta_0 > \bar{x} \\ \left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left\{-n\left(\frac{\bar{x}}{\theta_0} - 1\right)\right\}, & \theta_0 \leq \bar{x} \end{cases}.$$

In the case that $\theta_0 > \bar{x}$, the LRT will never reject H_0. In the case that $\theta_0 \leq \bar{x}$, the LRT rejects H_0 if

$$c \geq \lambda(\mathbf{x}) = \left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left\{-n\left(\frac{\bar{x}}{\theta_0} - 1\right)\right\} \implies \log c \geq \log\left(\frac{\bar{x}}{\theta_0}\right)^n \exp\left\{-n\left(\frac{\bar{x}}{\theta_0} - 1\right)\right\} = n\log\frac{\bar{x}}{\theta_0} - n\left(\frac{\bar{x}}{\theta_0} - 1\right).$$

THEOREM 10.7. *If $T(\mathbf{X})$ is a sufficient statistic for θ and $\lambda^*(t)$ and $\lambda(\mathbf{x})$ are the LRT statistics based on T and $\mathbf{X}$, respectively, then $\lambda^*(T(\mathbf{x})) = \lambda(\mathbf{x})$ for every $\mathbf{x}$ in the sample space. (This is Theorem 8.2.4 from Casella & Berger; the following proof is given there.)*

PROOF. From theorem 8.11, the pdf or pmf of $\mathbf{X}$ can be written as $f(\mathbf{x}|\theta) = g(T(\mathbf{x})|\theta)\, h(\mathbf{x})$, where $g(t|\theta)$ is the pdf or pmf of T and $h(\mathbf{x})$ does not depend on θ. Thus

$$\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\sup_{\Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\Theta} \mathcal{L}(\theta|\mathbf{x})} \\
&= \frac{\sup_{\Theta_0} f(\mathbf{x}|\theta)}{\sup_{\Theta} f(\mathbf{x}|\theta)} \\
&= \frac{\sup_{\Theta_0} g(T(\mathbf{x})|\theta)\, h(\mathbf{x})}{\sup_{\Theta} g(T(\mathbf{x})|\theta)\, h(\mathbf{x})}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\sup_{\Theta_0} g\left(T\left(\mathbf{x}\right)|\theta\right)}{\sup_{\Theta} g\left(T\left(\mathbf{x}\right)|\theta\right)} \\
&= \frac{\sup_{\Theta_0} \mathcal{L}^*\left(\theta|T\left(\mathbf{x}\right)\right)}{\sup_{\Theta} \mathcal{L}^*\left(\theta|T\left(\mathbf{x}\right)\right)} \\
&= \lambda^*\left(T\left(\mathbf{x}\right)\right).
\end{aligned}
$$

The third equality follows from the Factorization Theorem (because $T\left(\mathbf{x}\right)$ is sufficient for θ). The fourth equality follows from the fact that $h\left(\mathbf{x}\right)$ does not depend on θ, so the terms in the numerator and denominator cancel. The fifth equality follows from the fact that $g\left(T\left(\mathbf{x}\right)|\theta\right)$ is the pdf or pmf of T, and the resulting ratio is just the LRT test statistic based on T. □

EXAMPLE 10.8. Let $X_1, \ldots, X_n$, H_0, and H_1 be as in example 10.5. Then, the LRT test statistic is

$$\lambda\left(\mathbf{x}\right) = \exp\left\{-\frac{n}{2}\left(\theta_0 - \bar{x}\right)^2\right\},$$

so that the LRT rejects $H_0 : \theta = \theta_0$ if

$$|\theta_0 - \bar{x}| \geq \sqrt{-\frac{2}{n}\log c}.$$

From the proof of theorem 5.2, the joint density of $\mathbf{X}$ can be written as

$$
\begin{aligned}
f_{\mathbf{X}}\left(\mathbf{x}|\theta\right) &= \left(2\pi\right)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i - \theta\right)^2\right\} \\
&= \left(2\pi\right)^{-n/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + \sum_{i=1}^{n}\left(\bar{x} - \theta\right)^2\right]\right\} \\
&= \left(2\pi\right)^{-n/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + n\left(\bar{x} - \theta\right)^2\right]\right\}.
\end{aligned}
$$

Consider the statistic $T\left(\mathbf{X}\right) = \bar{X}$. From theorem 5.3, $T\left(\mathbf{X}\right)$ has a $N\left(\theta, 1/n\right)$ distribution, i.e.,

$$f_{T(\mathbf{X})}\left(t|\theta\right) = \frac{1}{\sqrt{2\pi/n}} \exp\left\{-\frac{1}{2}\frac{\left(t-\theta\right)^2}{1/n}\right\} = \sqrt{\frac{n}{2\pi}} \exp\left\{-\frac{n}{2}\left(t-\theta\right)^2\right\}.$$

Then, we have

$$
\begin{aligned}
\frac{f_{\mathbf{X}}\left(\mathbf{x}|\theta\right)}{f_{T(\mathbf{X})}\left(t|\theta\right)} &= \frac{\left(2\pi\right)^{-n/2} \exp\left\{-\left(1/2\right)\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + n\left(\bar{x} - \theta\right)^2\right]\right\}}{n^{1/2}\left(2\pi\right)^{-1/2} \exp\left\{-\left(n/2\right)\left(t-\theta\right)^2\right\}} \\
&= n^{-1/2}\left(2\pi\right)^{-(n-1)/2} \exp\left\{-\frac{1}{2}\left[\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 + n\left(\bar{x} - \theta\right)^2\right] + \frac{n}{2}\left(t-\theta\right)^2\right\} \\
&= n^{-1/2}\left(2\pi\right)^{-(n-1)/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 - \frac{n}{2}\left(\bar{x} - \theta\right)^2 + \frac{n}{2}\left(\bar{x} - \theta\right)^2\right\} \\
&= n^{-1/2}\left(2\pi\right)^{-(n-1)/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right\},
\end{aligned}
$$

which does not depend on θ, so it follows from theorem 8.6 that $T\left(\mathbf{X}\right) = \bar{X}$ is sufficient for θ. The log-likelihood of θ given $T\left(\mathbf{X}\right) = T\left(\mathbf{x}\right)$ is

$$\log \mathcal{L}^*\left(\theta|T\left(\mathbf{x}\right)\right) = \log\sqrt{\frac{n}{2\pi}} \exp\left\{-\frac{n}{2}\left(t-\theta\right)^2\right\} = \frac{1}{2}\log\frac{n}{2\pi} - \frac{n}{2}\left(\bar{x} - \theta\right)^2.$$

Taking the derivative with respect to θ gives

$$\frac{\partial}{\partial\theta}\log \mathcal{L}^*\left(\theta|T\left(\mathbf{x}\right)\right) = \frac{\partial}{\partial\theta}\left[\frac{1}{2}\log\frac{n}{2\pi} - \frac{n}{2}\left(\bar{x} - \theta\right)^2\right] = 0 - n\left(\bar{x} - \theta\right)\cdot -1 = n\left(\bar{x} - \theta\right).$$

Setting this equal to zero, we have

$$n\left(\bar{x}-\hat{\theta}\right)=0 \implies n\hat{\theta}=n\bar{x} \implies \hat{\theta}=\bar{x}.$$

We will evaluate the second derivative of $\log \mathcal{L}^*\left(\theta|T\left(\mathbf{x}\right)\right)$ with respect to θ at $\theta=\hat{\theta}$ to verify that this is a maximum.

$$\frac{\partial^2}{\partial\theta^2}\left[\log \mathcal{L}^*\left(\theta|T\left(\mathbf{x}\right)\right)\right|_{\theta=\hat{\theta}}=\frac{\partial}{\partial\theta}\left[n\left(\bar{x}-\theta\right)\right|_{\theta=\hat{\theta}}=\left[0-n\right|_{\theta=\hat{\theta}}=-n$$

We have $-n<0$, so it follows that $\hat{\theta}=\bar{X}$ is the MLE of $T\left(\mathbf{X}\right)=\bar{X}$. Then, the LRT test statistic based on $T\left(\mathbf{X}\right)$ is

$$\lambda^*\left(T\left(\mathbf{x}\right)\right)=\frac{\sup_{\Theta_0}\mathcal{L}^*\left(\theta|t\right)}{\sup_{\Theta}\mathcal{L}^*\left(\theta|t\right)}=\frac{\sqrt{n/2\pi}\exp\left\{-\left(n/2\right)\left(\bar{x}-\theta_0\right)^2\right\}}{\sqrt{n/2\pi}\exp\left\{-\left(n/2\right)\left(\bar{x}-\bar{x}\right)^2\right\}}=\exp\left\{-\frac{n}{2}\left(\bar{x}-\theta_0\right)^2\right\},$$

so that the LRT rejects H_0 if

$$c\geq\lambda^*\left(\mathbf{x}\right)=\exp\left\{-\frac{n}{2}\left(\bar{x}-\theta_0\right)^2\right\} \implies \log c\geq-\frac{n}{2}\left(\bar{x}-\theta_0\right)^2 \implies -\frac{2}{n}\log c\leq\left(\bar{x}-\theta_0\right)^2 \implies \left|\bar{x}-\theta_0\right|\geq\sqrt{-\frac{2}{n}\log c}.$$

We have$\left|\bar{x}-\theta_0\right|=\left|\theta_0-\bar{x}\right|$, so we see that the LRT statistics based on $\mathbf{X}$ and $T\left(\mathbf{X}\right)$ are equal for every $\mathbf{x}$.

EXAMPLE 10.9 (Normal LRT with unknown variance). Let $X_1,\ldots,X_n\overset{\text{iid}}{\sim}\mathcal{N}\left(\mu,\sigma^2\right)$ with $\left(\mu,\sigma^2\right)$ unknown. Specify the LRT for

$$H_0:\mu=\mu_0 \quad \text{vs.} \quad H_1:\mu\neq\mu_0.$$

The likelihood function is given by

$$\mathcal{L}\left(\mu,\sigma^2|\mathbf{x}\right)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right\}=\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\}.$$

Under the alternative hypothesis, we have $\Theta=\left\{\left(\mu,\sigma^2\right):\mu\neq\mu_0,\sigma^2>0\right\}$, so we must maximize $\left(\mu,\sigma^2\right)$ over the parameter space, i.e., find the MLEs of μ and σ^2. The log-likelihood is given by

$$\log\mathcal{L}\left(\mu,\sigma^2|\mathbf{x}\right)=\log\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right\}=-\frac{n}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2.$$

Taking the derivative with respect to μ gives

$$\begin{aligned}\frac{\partial}{\partial\mu}\log\mathcal{L}\left(\mu,\sigma^2|\mathbf{x}\right)&=\frac{\partial}{\partial\mu}\left[-\frac{n}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right]\\&=0-\frac{1}{2\sigma^2}\sum_{i=1}^{n}2\left(x_i-\mu\right)\cdot-1\\&=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right).\end{aligned}$$

Setting this equal to zero, we have

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)=0 \implies \sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\hat{\mu}=0 \implies n\bar{x}=n\hat{\mu} \implies \hat{\mu}=\bar{x}.$$

Taking the derivative of the log-likelihood with respect to σ^2 gives

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log\mathcal{L}\left(\mu,\sigma^2|\mathbf{x}\right)&=\frac{\partial}{\partial\sigma^2}\left[-\frac{n}{2}\log 2\pi\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\right]\\&=-\frac{n}{2}\frac{1}{2\pi\sigma^2}\left(2\pi\right)+\frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2\end{aligned}$$

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

Setting this equal to zero, we have

$$\frac{1}{2\left(\hat{\sigma}^2\right)^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = \frac{n}{2\hat{\sigma}^2} \implies 2n\left(\hat{\sigma}^2\right)^2 = 2\hat{\sigma}^2\sum_{i=1}^{n}\left(x_i-\mu\right)^2 \implies \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu\right)^2.$$

Substituting $\mu = \hat{\mu} = \bar{x}$ into our expression for $\hat{\sigma}^2$, we have

$$\begin{aligned}
\sup_{\Theta}\mathcal{L}\left(\mu,\sigma^2|\mathbf{x}\right) &= \mathcal{L}\left(\hat{\mu},\hat{\sigma}^2|\mathbf{x}\right) \\
&= \left(2\pi\hat{\sigma}^2\right)^{-n/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(x_i-\hat{\mu}\right)^2\right\} \\
&= (2\pi)^{-n/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)^{-n/2}\exp\left\{-\frac{1}{\frac{2}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right\} \\
&= (2\pi)^{-n/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right)^{-n/2}\exp\left\{-\frac{n}{2}\right\}.
\end{aligned}$$

Under the null hypothesis, we have $\Theta_0 = \left\{\left(\mu,\sigma^2\right) : \mu = \mu_0, \sigma^2 > 0\right\}$, i.e., only one value of μ is specified by H_0 and no restrictions are placed on the value of σ^2, so that we have

$$\begin{aligned}
\sup_{\Theta_0}\mathcal{L}\left(\mu,\sigma^2|\mathbf{x}\right) &= \mathcal{L}\left(\mu_0,\hat{\sigma}^2|\mathbf{x}\right) \\
&= \left(2\pi\hat{\sigma}^2\right)^{-n/2}\exp\left\{-\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2\right\} \\
&= (2\pi)^{-n/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu_0\right)\right)^{-n/2}\exp\left\{-\frac{1}{\frac{2}{n}\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2}\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2\right\} \\
&= (2\pi)^{-n/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\mu_0\right)\right)^{-n/2}\exp\left\{-\frac{n}{2}\right\}.
\end{aligned}$$

Then, the LRT test statistic is given by

$$\begin{aligned}
\lambda\left(\mathbf{x}\right) &= \frac{\sup_{\Theta_0} L\left(\mu,\sigma^2|\mathbf{x}\right)}{\sup_{\Theta} L\left(\mu,\sigma^2|\mathbf{x}\right)} \\
&= \frac{(2\pi)^{-n/2}\left[(1/n)\sum_{i=1}^{n}\left(x_i-\mu_0\right)\right]^{-n/2}\mathrm{e}^{-n/2}}{(2\pi)^{-n/2}\left[(1/n)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]^{-n/2}\mathrm{e}^{-n/2}} \\
&= \frac{\left[\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\right]^{n/2}}{\left[\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2\right]^{n/2}} \\
&= \left[\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2}\right]^{n/2},
\end{aligned}$$

so that the LRT rejects H_0 if

$$\begin{aligned}
c &\geq \lambda\left(\mathbf{x}\right) \\
&= \left[\frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2}\right]^{n/2}
\end{aligned}$$

$$
\begin{aligned}
\Leftrightarrow c^{2/n} &\geq \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\mu_0\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}+\bar{x}-\mu_0\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left[\left(x_i-\bar{x}\right)^2+2\left(x_i-\bar{x}\right)\left(\bar{x}-\mu_0\right)+\left(\bar{x}-\mu_0\right)^2\right]} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+2\left(\bar{x}-\mu_0\right)\sum_{i=1}^{n}\left(x_i-\bar{x}\right)+n\left(\bar{x}-\mu_0\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+2\left(\bar{x}-\mu_0\right)\left(\sum_{i=1}^{n}x_i-n\bar{x}\right)+n\left(\bar{x}-\mu_0\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+2\left(\bar{x}-\mu_0\right)\left(n\bar{x}-n\bar{x}\right)+n\left(\bar{x}-\mu_0\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-\mu_0\right)^2} \\
\Leftrightarrow \frac{1}{c^{2/n}} &\leq \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2+n\left(\bar{x}-\mu_0\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \\
&= \frac{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}+\frac{n\left(\bar{x}-\mu_0\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \\
&= 1+\frac{n\left(\bar{x}-\mu_0\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} \\
\Leftrightarrow \frac{n\left(\bar{x}-\mu_0\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} &\geq \frac{1}{c^{2/n}}-1 \\
\Leftrightarrow \frac{n\left(\bar{x}-\mu_0\right)^2}{\frac{n-1}{n-1}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2} &\geq \frac{1}{c^{2/n}}-1.
\end{aligned}
$$

Let $S^2=\left[1/\left(n-1\right)\right]\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$, so that we have

$$
\frac{n\left(\bar{x}-\mu_0\right)^2}{\left(n-1\right)S^2} \geq \frac{1}{c^{2/n}}-1 \implies \frac{\left(\bar{x}-\mu_0\right)^2}{S^2/n} \geq \left(n-1\right)\left(\frac{1}{c^{2/n}}-1\right).
$$

Then, the LRT rejects H_0 if

$$
\frac{\left(\bar{x}-\mu_0\right)^2}{S^2/n} \geq \left(n-1\right)\left(\frac{1}{c^{2/n}}-1\right) \implies \sqrt{\frac{\left(\bar{x}-\mu_0\right)^2}{S^2/n}} \geq \sqrt{\left(n-1\right)\left(\frac{1}{c^{2/n}}-1\right)} \implies \frac{\left|\bar{x}-\mu_0\right|}{S/\sqrt{n}} \geq c^*,
$$

where $c^*=\sqrt{\left(n-1\right)\left(1/c^{2/n}-1\right)}$.

From theorem 5.3, $\bar{X}$ has the distribution of a $\mathcal{N}\left(\mu,\sigma^2/n\right)$ random variable. Let $U=\left(\bar{X}-\mu\right)/\left(\sigma/\sqrt{n}\right)$, so that

$$
\begin{aligned}
P\left(\left\{U\leq u\right\}\right) &= P\left(\left\{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\leq u\right\}\right) \\
&= P\left(\left\{\bar{X}-\mu\leq\frac{\sigma u}{\sqrt{n}}\right\}\right) \\
&= P\left(\left\{\bar{X}\leq\frac{\sigma u}{\sqrt{n}}+\mu\right\}\right) \\
&= \int_{-\infty}^{\sigma u/\sqrt{n}+\mu}\frac{1}{\sqrt{2\pi\sigma^2/n}}\exp\left\{-\frac{\left(\bar{x}-\mu\right)^2}{2\sigma^2/n}\right\}\mathrm{d}\,\bar{x}.
\end{aligned}
$$

Let

$$s = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Leftrightarrow \bar{x} = \frac{\sigma s}{\sqrt{n}} + \mu,$$

so that we have

$$\frac{\mathrm{d}\, s}{\mathrm{d}\, \bar{x}} = \frac{1}{\sigma/\sqrt{n}} \implies \mathrm{d}\, \bar{x} = \frac{\sigma}{\sqrt{n}} \mathrm{d}\, s$$

and

$$s\left(\frac{\sigma u}{\sqrt{n}} + \mu\right) = \frac{(\sigma u/\sqrt{n} + \mu) - \mu}{\sigma/\sqrt{n}} = \frac{\sigma u/\sqrt{n}}{\sigma/\sqrt{n}} = u.$$

Then, we have

$$\begin{aligned} P\left(\{U \leq u\}\right) &= \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi\sigma^2/n}} \exp\left\{-\frac{\left(\sigma s/\sqrt{n} + \mu - \mu\right)^2}{2\sigma^2/n}\right\} \frac{\sigma}{\sqrt{n}} \mathrm{d}\, s \\ &= \int_{-\infty}^{u} \frac{\sqrt{n}}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\left(\sigma s/\sqrt{n}\right)^2}{2\sigma^2/n}\right\} \frac{\sigma}{\sqrt{n}} \mathrm{d}\, s \\ &= \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\sigma^2 s^2/n}{2\sigma^2/n}\right\} \mathrm{d}\, s \\ &= \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-s^2/2} \mathrm{d}\, s, \end{aligned}$$

which we recognize as the cdf of a $\mathcal{N}\left(0, 1^2\right)$ random variable, i.e., U has the distribution of a standard normal random variable. Now consider $V = \sqrt{S^2/\sigma^2}$. From theorem 5.3, the quantity $(n-1)\, S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom, so it follows that $V = \sqrt{S^2/\sigma^2} \sim \sqrt{\chi^2_{n-1}/\left(n-1\right)}$. Then, we have

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} = \frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \frac{\sigma}{\sigma} = \frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \frac{\sigma}{S} = \frac{|\bar{x} - \mu_0| \,/\, (\sigma/\sqrt{n})}{S/\sigma} = \frac{|\bar{x} - \mu_0| \,/\, (\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}} = \frac{U}{V},$$

where U and V are independent. We wish to find the distribution of $U/\sqrt{V/p}$, where $V \sim \chi^2_p$. The joint pdf of $U$and V is

$$f_{U,V}\left(u, v\right) = f_U\left(u\right) f_V\left(v\right) = \left[\frac{1}{\sqrt{2\pi}} \mathrm{e}^{-u^2/2}\right] \left[\frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} v^{(p/2)-1} \mathrm{e}^{-v/2}\right], \quad u \in \mathbb{R}, \quad v > 0.$$

Let

$$\mathcal{A} = \left\{(u, v) : f_{U,V}\left(u, v\right) > 0\right\},$$

and make the transformation

$$t = \frac{u}{\sqrt{v/p}}, \quad w = v \implies u = t\sqrt{\frac{v}{p}} = t\sqrt{\frac{w}{p}},$$

so that we have

$$\mathcal{B} = \left\{(t, w) : t = u/\sqrt{v/p}, w = v, (u, v) \in \mathcal{A}\right\}.$$

Then, $f_{T,W}\left(t, w\right)$ will be positive on $\mathcal{B}$. We have $u \in \mathbb{R}$ and $p, v > 0$, so the quantity $t = u/\sqrt{v/p} \in \mathbb{R}$, and $w > 0$. The Jacobian of the transformation is

$$J = \begin{vmatrix} \dfrac{\partial u}{\partial t} & \dfrac{\partial u}{\partial w} \\ \dfrac{\partial v}{\partial t} & \dfrac{\partial v}{\partial w} \end{vmatrix} = \begin{vmatrix} \sqrt{w/p} & 0 \\ 0 & 1 \end{vmatrix} = \sqrt{w/p},$$

and the joint pdf of T and W is given by

$$f_{T,W}\left(t, w\right) = f_{U,V}\left(t\sqrt{\frac{w}{p}}, w\right) |J|$$

$$
\begin{aligned}
&= \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\left(t\sqrt{w/p}\right)^2/2} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} w^{(p/2)-1} \mathrm{e}^{-w/2} \left|\sqrt{\frac{w}{p}}\right| \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2}} \mathrm{e}^{-(1/2)t^2 w/p} w^{(p/2)-1} \mathrm{e}^{-w/2} w^{1/2} p^{-1/2} \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \mathrm{e}^{-(1/2)t^2 w/p - w/2} w^{(p/2)-1/2} \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \mathrm{e}^{-(1/2)\left(t^2 w/p + w\right)} w^{(p-1)/2+1-1} \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \mathrm{e}^{-(1/2)\left(t^2/p+1\right)w} w^{(p+1)/2-1},
\end{aligned}
$$

so that the marginal pdf of T is given by

$$
\begin{aligned}
f_T(t) &= \int_0^\infty f_{T,W}(t,w)\,\mathrm{d}\,w \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \mathrm{e}^{-(1/2)\left(t^2/p+1\right)w} w^{(p+1)/2-1}\,\mathrm{d}\,w \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \int_0^\infty \mathrm{e}^{-(1/2)\left(t^2/p+1\right)w} w^{(p+1)/2-1}\,\mathrm{d}\,w.
\end{aligned}
$$

We recognize the integrand as the kernel of a Gamma $\left((p+1)/2, 2/\left(t^2/p+1\right)\right)$ random variable, so we write

$$
\begin{aligned}
f_T(t) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \Gamma\left(\frac{p+1}{2}\right) \left(\frac{2}{t^2/p+1}\right)^{(p+1)/2} \\
&\quad \cdot \int_0^\infty \frac{1}{\Gamma\left((p+1)/2\right) \left[2/\left(t^2/p+1\right)\right]^{(p+1)/2}} \mathrm{e}^{-(1/2)\left(t^2/p+1\right)w} w^{(p+1)/2-1}\,\mathrm{d}\,w \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \Gamma\left(\frac{p+1}{2}\right) \left(\frac{2}{t^2/p+1}\right)^{(p+1)/2} \cdot 1 \\
&= \frac{1}{\sqrt{2\pi}} \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} p^{1/2}} \Gamma\left(\frac{p+1}{2}\right) \left(\frac{2}{t^2/p+1}\right)^{(p+1)/2},
\end{aligned}
$$

which from definition 5.4 is the pdf of a $t_p = t_{n-1}$ distribution.

10.2. Methods of evaluating tests

10.2.1. Error probabilities and the power function. A hypothesis test of $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ might make one of two types of errors. These two types of errors traditionally are known as Type I Error and Type II Error. If $\theta \in \Theta_0$ but the hypothesis test incorrectly decides to reject H_0, then the test has made a *Type I Error*. If, on the other hand, $\theta \in \Theta_0^c$ but the test decides to accept H_0, a *Type II Error* has been made.

TABLE 1. Two types of errors in hypothesis testing

		Decision	
		Accept H_0	Reject H_0
Truth	H_0	Correct decision	Type I Error
	H_1	Type II Error	Correct decision

Suppose R denotes the rejection region for a test. Then for $\theta \in \Theta_0$, the test will make a mistake if $\mathbf{x} \in R$, so the probability of a Type I Error is $P_\theta(\{\mathbf{X} \in R\})$. For $\theta \in \Theta_0^c$, the probability of a Type II Error is

$P_\theta(\{\mathbf{X} \in R^c\})$. We have

$$P_\theta(\{\mathbf{X} \in R\}) = \begin{cases} \text{probability of a Type I Error} & \text{if } \theta \in \Theta_0 \\ \text{one minus the probability of a Type II Error} & \text{if } \theta \in \Theta_0^c \end{cases}.$$

DEFINITION 10.10. The *power function* of a hypothesis test with rejection region R is the function of θ defined by $\beta(\theta) = P_\theta(\{\mathbf{X} \in R\})$.

The ideal power function is 0 for all $\theta \in \Theta_0$ and 1 for all $\theta \in \Theta_0^c$. Except in trivial situations, this ideal cannot be attained. Qualitatively, a good test has a power function near 1 for most $\theta \in \Theta_0^c$ and near 0 for most $\theta \in \Theta_0$. Figure 10.2.1 shows the relationships among the critical region, Type I Error, Type II Error, and the power function.

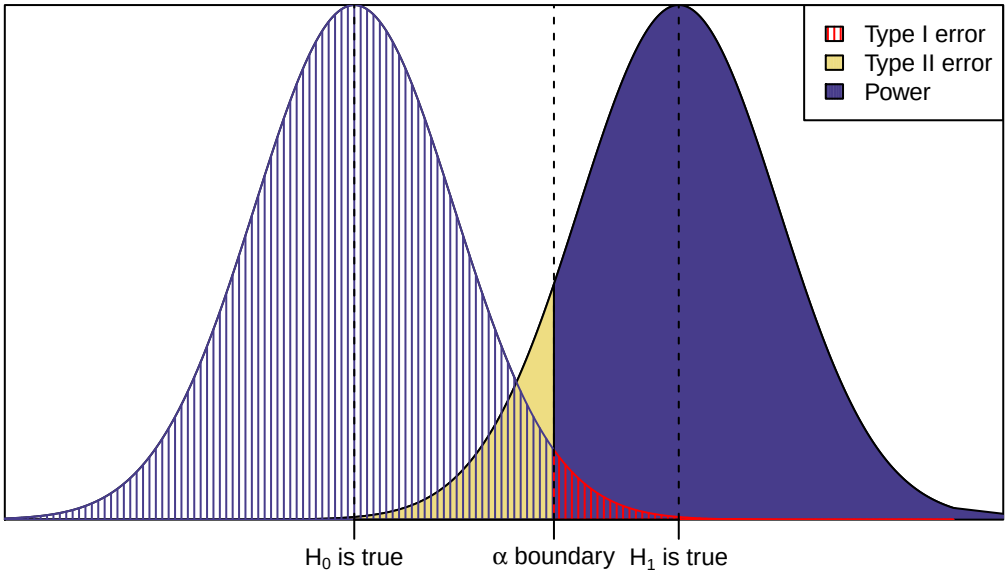


FIGURE 10.2.1. Type I and Type II errors

EXAMPLE 10.11. Let $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 5^2)$. Consider the hypothesis test

$$H_0 : \theta \leq 17 \quad \text{vs.} \quad H_1 : \theta > 17.$$

We will reject H_0 if $\bar{X} > 17 + 5/\sqrt{n}$. Find the probabilities of Type I and Type II Errors.
The parameter spaces associated with the null and alternative hypotheses are

$$\Theta_0 = \{\theta : \theta \leq 17\} \quad \text{and} \quad \Theta_0^c = \{\theta : \theta > 17\}.$$

The probability of a Type I error is equal to the probability that we reject H_0 given that $\theta \in \Theta_0$, i.e., $P(\{\mathbf{X} \in R | \theta \in \Theta_0\})$. The power function for this test is $\beta(\theta) = P_\theta\left(\left\{\bar{X} > 17 + 5/\sqrt{n}\right\}\right)$. From theorem 5.3, $\bar{X}$ has the distribution of a $\mathcal{N}\left(\theta, (5/\sqrt{n})^2\right)$ random variable, so it follows that

$$Z = \frac{\bar{X} - \theta}{5/\sqrt{n}}$$

has the distribution of a $\mathcal{N}(0, 1^2)$ random variable, and that $\bar{X} = (5/\sqrt{n})Z + \theta$. Then, we have

$$\begin{aligned} \beta(\theta) &= P_\theta(\{\mathbf{X} \in R\}) \\ &= P_\theta\left(\left\{\bar{X} > 17 + \frac{5}{\sqrt{n}}\right\}\right) \\ &= P_\theta\left(\left\{\frac{5}{\sqrt{n}}Z + \theta > 17 + \frac{5}{\sqrt{n}}\right\}\right) \end{aligned}$$

$$
\begin{aligned}
&= P_\theta\left(\left\{\frac{5}{\sqrt{n}}Z > 17 + \frac{5}{\sqrt{n}} - \theta\right\}\right)\\
&= P_\theta\left(\left\{Z > \left(17 + \frac{5}{\sqrt{n}} - \theta\right)\frac{\sqrt{n}}{5}\right\}\right)\\
&= P_\theta\left(\left\{Z > \frac{17\sqrt{n}}{5} + 1 - \frac{\theta\sqrt{n}}{5}\right\}\right)\\
&= P_\theta\left(\left\{Z > \frac{17\sqrt{n} + 5 - \theta\sqrt{n}}{5}\right\}\right)\\
&= P_\theta\left(\left\{Z > \frac{17 - \theta + 5/\sqrt{n}}{5/\sqrt{n}}\right\}\right)\\
&= 1 - P_\theta\left(\left\{Z \leq \frac{17 - \theta + 5/\sqrt{n}}{5/\sqrt{n}}\right\}\right)\\
&= 1 - \Phi\left(\frac{17 - \theta + 5/\sqrt{n}}{5/\sqrt{n}}\right),
\end{aligned}
$$

where $\Phi(z)$ is the cdf of Z. It follows from theorem 2.4 that $\Phi(z)$ is a nondecreasing function of z, that $\lim_{z\to-\infty}\Phi(z) = 0$, and that $\lim_{z\to\infty}\Phi(z) = 1$. It follows that as θ increases from $-\infty$ to ∞, $\Phi(z)$ increases from 0 to 1. Then, the probability of a Type I Error is given by

$$
\begin{aligned}
\alpha &= P(\{\mathbf{X} \in R | \theta \in \Theta_0\})\\
&= \sup_{\theta\in\Theta_0}\beta(\theta)\\
&= \sup_{\theta\leq 17}\left[1 - \Phi\left(\frac{17 - \theta + 5/\sqrt{n}}{5/\sqrt{n}}\right)\right]\\
&= 1 - \Phi\left(\frac{17 - 17 + 5/\sqrt{n}}{5/\sqrt{n}}\right)\\
&= 1 - \Phi(1)\\
&\approx 1 - 0.8413447\\
&= 0.1586553.
\end{aligned}
$$

The probability of a Type II Error is equal to the probability that we fail to reject H_0 given that $\theta \in \Theta_0^c$, i.e., $P(\{\mathbf{X} \in R^c | \theta \in \Theta_0^c\})$. The probability that $\mathbf{X} \in R^c$ is given by

$$
P_\theta(\{\mathbf{X} \in R^c\}) = 1 - P_\theta(\{\mathbf{X} \in R\}) = 1 - \beta(\theta) = 1 - \left(1 - \Phi\left(\frac{17 - \theta + 5/\sqrt{n}}{5/\sqrt{n}}\right)\right) = \Phi\left(\frac{17 - \theta + 5/\sqrt{n}}{5/\sqrt{n}}\right).
$$

To calculate the probability of a Type II Error, we would require knowledge of the true value of θ.

DEFINITION 10.12. For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a *size* α *test* if $\sup_{\theta\in\Theta_0}\beta(\theta) = \alpha$.

If the data are continuous, it is possible to achieve any particular α. If the data are discrete, though, it may not be possible to achieve any particular α.

DEFINITION 10.13. For $0 \leq \alpha \leq 1$, a test with power function $\beta(\theta)$ is a *level* α *test* if $\sup_{\theta\in\Theta_0}\beta(\theta) \leq \alpha$.

Level α tests have Type I Error probabilities at most α for all $\theta \in \Theta_0$.

EXAMPLE 10.14. Let $X_1, \ldots, X_{10}$ be a random sample (of size $n = 10$) from the distribution of a Bernoulli (θ) random variable. Consider the hypothesis test

$$
H_0 : \theta = 0.5 \quad \text{vs.} \quad H_1 : \theta \neq 0.5.
$$

We will reject H_0 if $\sum_{i=1}^n X_i \geq 8$. Find the level of this test.

Noting that the random variable $Y = \sum_{i=1}^n X_i$ has a Binomial (n, θ) distribution, the power function for this test is

$$
\beta(\theta) = P_\theta(\{\mathbf{X} \in R\}) = P_\theta\left(\left\{\sum_{i=1}^n X_i \geq 8\right\}\right) = P_\theta(\{Y \geq 8\}) = \sum_{y=8}^{10}\binom{10}{y}\theta^y(1-\theta)^{10-y}.
$$

The level α of this test is given by

$$\begin{aligned}
\alpha &= \sup_{\theta \in \Theta_0} \beta(\theta) \\
&= \beta\left(\frac{1}{2}\right) \\
&= \sum_{y=8}^{10} \binom{10}{y} \left(\frac{1}{2}\right)^y \left(1-\frac{1}{2}\right)^{10-y} \\
&= \binom{10}{8}\left(\frac{1}{2}\right)^8\left(\frac{1}{2}\right)^2 + \binom{10}{9}\left(\frac{1}{2}\right)^9\left(\frac{1}{2}\right)^1 + \binom{10}{10}\left(\frac{1}{2}\right)^{10}\left(\frac{1}{2}\right)^0 \\
&= 45\frac{1}{2^{10}} + 10\frac{1}{2^{10}} + 1\frac{1}{2^{10}} \\
&= \frac{56}{1024} \\
&\approx 0.0546875.
\end{aligned}$$

Note that it is not possible to construct a test of H_0 versus H_1 of size $\alpha = 0.05$. If we change the rejection region such that we will reject H_0 if $\sum_{i=1}^n X_i \geq 7$, then we will have a level $\alpha = 0.171875$ test. If we change the rejection region such that we will reject H_0 if $\sum_{i=1}^n X_i \geq 9$, then we will have a level $\alpha = 0.0107422$ test.

EXAMPLE 10.15. Let $X_1, \ldots, X_n \sim \mathcal{N}\left(\theta, 1^2\right)$. Consider the hypothesis test

$$H_0 : \theta = \theta_0 \quad \text{vs.} \quad H_1 : \theta \neq \theta_0.$$

From example 10.5, the likelihood ratio test rejects H_0 if

$$\left|\theta_0 - \bar{X}\right| \geq \sqrt{-\frac{2}{n}\log c} = \frac{\sqrt{-2\log c}}{\sqrt{n}} \implies \left|\theta_0 - \bar{X}\right|\sqrt{n} \geq \sqrt{-2\log c} \implies \frac{\left|\theta_0 - \bar{X}\right|}{1/\sqrt{n}} \geq k,$$

where $c \in (0, 1]$, so that $-2\log c \geq 0$, so that $k = \sqrt{-2\log c} \geq 0$. Specify a size $\alpha = 0.05$ test.
A size α test is given by

$$\begin{aligned}
\alpha &= \sup_{\theta \in \Theta_0} \beta(\theta) \\
&= \sup_{\theta \in \Theta_0} P_\theta(\{\mathbf{X} \in R\}) \\
&= \sup_{\theta \in \Theta_0} P_\theta\left(\left\{\frac{\left|\theta_0 - \bar{X}\right|}{1/\sqrt{n}} \geq k\right\}\right).
\end{aligned}$$

From theorem 5.3, $\bar{X}$ has the distribution of a $\mathcal{N}\left(\theta, \left(1/\sqrt{n}\right)^2\right)$ random variable, so it follows that

$$Z = \frac{\bar{X} - \theta}{1/\sqrt{n}}$$

has the distribution of a $\mathcal{N}\left(0, 1^2\right)$ random variable, and that $\bar{X} = \left(1/\sqrt{n}\right)Z + \theta$. Then, we have

$$\begin{aligned}
\alpha &= \sup_{\theta \in \Theta_0} P_\theta\left(\left\{\frac{\left|\theta_0 - \bar{X}\right|}{1/\sqrt{n}} \geq k\right\}\right) \\
&= \sup_{\theta \in \Theta_0} P_\theta\left(\left\{\frac{\left|\theta_0 - \left[\left(1/\sqrt{n}\right)Z + \theta\right]\right|}{1/\sqrt{n}} \geq k\right\}\right) \\
&= P_\theta\left(\left\{\frac{\left|\theta_0 - \left(1/\sqrt{n}\right)Z - \theta_0\right|}{1/\sqrt{n}} \geq k\right\}\right) \\
&= P_\theta\left(\left\{\frac{\left(1/\sqrt{n}\right)|-Z|}{1/\sqrt{n}} \geq k\right\}\right) \\
&= P_\theta(\{|-Z| \geq k\}) \\
&= P_\theta(\{Z \geq k\} \cup \{-Z \geq k\})
\end{aligned}$$

$$= P_\theta(\{Z \geq k\}) + P(\{Z \leq -k\}),$$

where the final equality follows from the fact that $\{Z \geq k\}$ and $\{Z \leq -k\}$ are disjoint. From the symmetry of $\Phi(z)$, the cdf of Z, we have $P_\theta(\{Z \geq k\}) = P_\theta(\{Z \leq -k\})$, and it follows that

$$\alpha = P_\theta(\{Z \geq k\}) + P_\theta(\{Z \leq -k\}) = P_\theta(\{Z \leq -k\}) + P_\theta(\{Z \leq -k\}) = 2\Phi(-k).$$

For $\alpha = 0.05$, we have

$$0.05 = 2\Phi(-k) \implies \Phi(-k) = 0.025 \implies k \approx 1.959964.$$

10.2.2. Most powerful tests.

DEFINITION 10.16. Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$, is a *uniformly most powerful* (UMP) *class $\mathcal{C}$ test* if $\beta(\theta) \geq \beta'(\theta)$ for every $\theta \in \Theta_0^c$ and every $\beta'(\theta)$ that is a power function of a test in class $\mathcal{C}$.

The class $\mathcal{C}$ will be the class of all level α tests. The test described in the definition above is then called a UMP level α test.

THEOREM 10.17 (Neyman-Pearson Lemma). *Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(\mathbf{x}|\theta_i)$, $i = 0, 1$, using a test with rejection region R that satisfies*

$$\mathbf{x} \in R \quad \text{if} \quad f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)$$

(10.2.1) *and*

$$\mathbf{x} \in R^c \quad \text{if} \quad f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0),$$

for some $k \geq 0$, and

$$\alpha = P_{\theta_0}(\{\mathbf{X} \in R\}). \tag{10.2.2}$$

Then

(a) *(Sufficiency) Any test that satisfies (10.2.1) and (10.2.2) is a UMP level α test.*
(b) *(Necessity) If there exists a test satifying (10.2.1) and (10.2.2) with $k > 0$, then every UMP level α test is a size α test (satisfies (10.2.2)) and every UMP level α test satisfies (10.2.1) except perhaps on a set A satisfying $P_{\theta_0}(\{\mathbf{X} \in A\}) = P_{\theta_1}(\{\mathbf{X} \in A\}) = 0$.*

(This is Theorem 8.3.12 from Casella & Berger; the following proof is given there.)

PROOF. We will prove the theorem for the case that $f(\mathbf{x}|\theta_0)$ and $f(\mathbf{x}|\theta_1)$ are pdfs of continuous random variables. The proof for discrete random variables can be accomplished by replacing integrals with sums.

Note first that any test satisfying (10.2.2) is a size α and, hence, a level α test because $\sup_{\theta \in \Theta_0} P_\theta(\{\mathbf{X} \in R\}) = P_{\theta_0}(\mathbf{X} \in R) = \alpha$, since Θ_0 has only one point.

To ease notation, we define a *test function*, a function on the sample space that is 1 if $\mathbf{x} \in R$ and 0 if $\mathbf{x} \in R^c$. That is, it is the indicator function of the rejection region. Let $\phi(\mathbf{x})$ be the test function of a test satisfying (10.2.1) and (10.2.2). Let $\phi'(\mathbf{x})$ be the test function of any other level α test, and let $\beta(\theta)$ and $\beta'(\theta)$ be the power functions corresponding to the tests ϕ and ϕ', respectively. Because $0 \leq \phi'(\mathbf{x}) \leq 1$, (10.2.1) implies that $(\phi(\mathbf{x}) - \phi'(\mathbf{x}))(f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0)) \geq 0$ for every $\mathbf{x}$ (since $\phi = 1$ if $f(\mathbf{x}|\theta_1) > kf(\mathbf{x}|\theta_0)$ and $\phi = 0$ if $f(\mathbf{x}|\theta_1) < kf(\mathbf{x}|\theta_0)$). Thus

(10.2.3)

$$\begin{aligned}
0 &\leq \int [\phi(\mathbf{x}) - \phi'(\mathbf{x})][f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0)]\,\mathrm{d}\mathbf{x} \\
&= \int [\phi(\mathbf{x}) f(\mathbf{x}|\theta_1) - \phi(\mathbf{x}) kf(\mathbf{x}|\theta_0) - \phi'(\mathbf{x}) f(\mathbf{x}|\theta_1) + \phi'(\mathbf{x}) kf(\mathbf{x}|\theta_0)]\,\mathrm{d}\mathbf{x} \\
&= \int \phi(\mathbf{x}) f(\mathbf{x}|\theta_1)\,\mathrm{d}\mathbf{x} - \int \phi(\mathbf{x}) kf(\mathbf{x}|\theta_0)\,\mathrm{d}\mathbf{x} - \int \phi'(\mathbf{x}) f(\mathbf{x}|\theta_1)\,\mathrm{d}\mathbf{x} + \int \phi'(\mathbf{x}) kf(\mathbf{x}|\theta_0)\,\mathrm{d}\mathbf{x} \\
&= \left[\int_R 1 \cdot f(\mathbf{x}|\theta_1)\,\mathrm{d}\mathbf{x} + \int_{R^c} 0 \cdot f(\mathbf{x}|\theta_1)\,\mathrm{d}\mathbf{x}\right] - \left[\int_R 1 \cdot kf(\mathbf{x}|\theta_0)\,\mathrm{d}\mathbf{x} + \int_{R^c} 0 \cdot kf(\mathbf{x}|\theta_0)\,\mathrm{d}\mathbf{x}\right]
\end{aligned}$$

$$
\begin{aligned}
&- \left[\int_{R'} 1 \cdot f(\mathbf{x}|\theta_1)\,\mathrm{d}\,\mathbf{x} + \int_{(R')^c} 0 \cdot f(\mathbf{x}|\theta_1)\,\mathrm{d}\,\mathbf{x}\right] + \left[\int_{R'} 1 \cdot kf(\mathbf{x}|\theta_0)\,\mathrm{d}\,\mathbf{x} + \int_{(R')^c} 0 \cdot kf(\mathbf{x}|\theta_0)\,\mathrm{d}\,\mathbf{x}\right] \\
&= \left[\int_R f(\mathbf{x}|\theta_1)\,\mathrm{d}\,\mathbf{x} + 0\right] - \left[\int_R kf(\mathbf{x}|\theta_0)\,\mathrm{d}\,\mathbf{x} + 0\right] - \left[\int_{R'} f(\mathbf{x}|\theta_1)\,\mathrm{d}\,\mathbf{x} + 0\right] + \left[\int_{R'} kf(\mathbf{x}|\theta_0)\,\mathrm{d}\,\mathbf{x}\right] \\
&= \int_R f(\mathbf{x}|\theta_1)\,\mathrm{d}\,\mathbf{x} - k\int_R f(\mathbf{x}|\theta_0)\,\mathrm{d}\,\mathbf{x} - \int_{R'} f(\mathbf{x}|\theta_1)\,\mathrm{d}\,\mathbf{x} + k\int_{R'} f(\mathbf{x}|\theta_0)\,\mathrm{d}\,\mathbf{x} \\
&= P_{\theta_1}(\{\mathbf{X} \in R\}) - kP_{\theta_0}(\{\mathbf{X} \in R\}) - P_{\theta_1}(\{\mathbf{X} \in R'\}) + kP_{\theta_0}(\{\mathbf{X} \in R'\}) \\
&= \beta(\theta_1) - k\beta(\theta_0) - \beta'(\theta_1) + k\beta'(\theta_0) \\
&= \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)).
\end{aligned}
$$

Statement (a) is proved by noting that, since ϕ' is a level α test and ϕ is a size α test,

$$\beta(\theta_0) - \beta'(\theta_0) = \sup_{\theta \in \Theta_0} \beta(\theta) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \geq 0.$$

Thus (10.2.3) and $k \geq 0$ imply that

$$0 \leq \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)) \leq \beta(\theta_1) - \beta'(\theta_1),$$

showing that $\beta(\theta_1) \geq \beta'(\theta_1)$ and hence ϕ has greater power than ϕ'. Since ϕ' was an arbitrary level α test and θ_1 is the only point in Θ_0^c, ϕ is a UMP level α test.

To prove statement (b), let ϕ' now be the test function for any UMP level α test. By part (a), ϕ, the test satisfying (10.2.1) and (10.2.2), is also a UMP level α test, thus $\beta(\theta_1) = \beta'(\theta_1)$. This fact, (10.2.3), and $k > 0$ imply

$$0 \leq \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0)) = -k(\beta(\theta_0) - \beta'(\theta_0)) \implies \beta(\theta_0) - \beta'(\theta_0) \leq 0.$$

Now, since ϕ' is a level α test, $\sup_{\theta \in \Theta_0} \beta'(\theta) = \beta'(\theta_0) \leq \alpha$. We also have

$$\beta(\theta_0) - \beta'(\theta_0) = \alpha - \beta'(\theta_0) \leq 0 \implies \alpha \leq \beta'(\theta_0),$$

so it follows that $\beta'(\theta_0) = \alpha$, that is, ϕ' is a size α test, and this also implies that (10.2.3) is an equality in this case. But the nonnegative integrated $(\phi(\mathbf{x}) - \phi'(\mathbf{x}))(f(\mathbf{x}|\theta_1) - kf(\mathbf{x}|\theta_0))$ will have a zero integral only if ϕ' satisfies (10.2.1), except perhaps on a set A with $\int_A f(\mathbf{x}|\theta_i)\,\mathrm{d}\mathbf{x} = 0$. This implies that the last assertion in statement (b) is true. □

COROLLARY 10.18. *Consider the hypothesis problem posed in theorem 10.17. Suppose $T(\mathbf{X})$ is a sufficient statistic for θ and $g(t|\theta_i)$ is the pdf or pmf of T corresponding to θ_i, $i = 0, 1$. Then any test based on T with rejection region S (a subset of the sample space of T) is a UMP level α test if it satisfies*

$$
\begin{aligned}
&t \in S \quad \text{if} \quad g(t|\theta_1) > kg(t|\theta_0) \\
&\text{and} \\
&t \in S^c \quad \text{if} \quad g(t|\theta_1) < kg(t|\theta_0),
\end{aligned}
\tag{10.2.4}
$$

for some $k \geq 0$, where

$$\alpha = P_{\theta_0}(\{T \in S\}). \tag{10.2.5}$$

(This is Corollary 8.3.13 from Casella & Berger; the following proof is given there.)

PROOF. In terms of the original sample $\mathbf{X}$, the test based on T has the rejection region $R = \{\mathbf{x} : T(\mathbf{x}) \in S\}$. By theorem 8.11, the pdf or pmf of $\mathbf{X}$ can be written as $f(\mathbf{x}|\theta_i) = g(T(\mathbf{x})|\theta_i)\,h(\mathbf{x})$, $i = 0, 1$, for some nonnegative function $h(\mathbf{x})$. Multiplying the inequalities in (10.2.4) by this nonnegative function, we see that R satisfies

$$\mathbf{x} \in R \text{ if } f(\mathbf{x}|\theta_1) = g(T(\mathbf{x})|\theta_1)\,h(\mathbf{x}) > kg(T(\mathbf{x})|\theta_0)\,h(\mathbf{x}) = kf(\mathbf{x}|\theta_0)$$

and

$$\mathbf{x} \in R^c \text{ if } f(\mathbf{x}|\theta_1) = g(T(\mathbf{x})|\theta_1)\,h(\mathbf{x}) < kg(T(\mathbf{x})|\theta_0)\,h(\mathbf{x}) = kf(\mathbf{x}|\theta_0).$$

Also, by (10.2.5),

$$P_{\theta_0}(\{\mathbf{X} \in R\}) = P_{\theta_0}(T(\mathbf{X}) \in S) = \alpha.$$

So, by the sufficient part of theorem 10.17, the test based on T is a UMP level α test. □

When we derive a test that satisfies the inequality (10.2.1), and hence is a UMP level α test, it is usually easier to rewrite the inequalities as $f(\mathbf{x}|\theta_1)/f(\mathbf{x}|\theta_0) > k$. (We must be careful about dividing by 0.)

EXAMPLE 10.19. In example 10.15, we found that a size $\alpha = 0.05$ test rejects H_0 for $k \approx 1.96$. Find an expression for the minimum sample size such that the power of the test is 0.90.

The power of a test is equal to the probability that it rejects the null hypothesis given that the alternative hypothesis is true. A size $\alpha = 0.05$ test will reject H_0 if

$$\frac{\left|\theta_0 - \bar{X}\right|}{1/\sqrt{n}} \geq k \Leftrightarrow \left|\theta_0 - \bar{X}\right| \geq \frac{1.96}{\sqrt{n}},$$

so that the power of such a test is given by power $= P\left(\left\{\left|\theta_0 - \bar{X}\right| \geq 1.96/\sqrt{n} | \theta = \theta_1\right\}\right)$. Recalling that $\bar{X} \sim \mathcal{N}\left(\theta, (1/\sqrt{n})^2\right)$ and letting

$$Z = \frac{\bar{X} - \theta}{1/\sqrt{n}} \sim \mathcal{N}\left(0, 1^2\right),$$

we have

$$\begin{aligned}
\text{power} &= P\left(\left\{\left|\theta_0 - \bar{X}\right| \geq \frac{1.96}{\sqrt{n}} | \theta = \theta_1\right\}\right) \\
&= P\left(\left\{\left|\theta_0 - \left(\frac{1}{\sqrt{n}}Z + \theta\right)\right| \geq \frac{1.96}{\sqrt{n}} | \theta = \theta_1\right\}\right) \\
&= P\left(\left\{\left|\theta_0 - \frac{1}{\sqrt{n}}Z - \theta_1\right| \geq \frac{1.96}{\sqrt{n}}\right\}\right) \\
&= P\left(\left\{\left|-\frac{1}{\sqrt{n}}Z + (\theta_0 - \theta_1)\right| \geq \frac{1.96}{\sqrt{n}}\right\}\right) \\
&= P\left(\left\{-\frac{1}{\sqrt{n}}Z + (\theta_0 - \theta_1) \geq \frac{1.96}{\sqrt{n}}\right\} \cup \left\{-\left(-\frac{1}{\sqrt{n}}Z + (\theta_0 - \theta_1)\right) \geq \frac{1.96}{\sqrt{n}}\right\}\right) \\
&= P\left(\left\{-\frac{1}{\sqrt{n}}Z \geq \frac{1.96}{\sqrt{n}} - (\theta_0 - \theta_1)\right\} \cup \left\{\frac{1}{\sqrt{n}}Z \geq \frac{1.96}{\sqrt{n}} + (\theta_0 - \theta_1)\right\}\right) \\
&= P\left(\left\{Z \leq -1.96 + (\theta_0 - \theta_1)\sqrt{n}\right\} \cup \left\{Z \geq 1.96 + (\theta_0 - \theta_1)\sqrt{n}\right\}\right) \\
&= P\left(\left\{Z \leq -1.96 + (\theta_0 - \theta_1)\sqrt{n}\right\}\right) + P\left(\left\{Z \geq 1.96 + (\theta_0 - \theta_1)\sqrt{n}\right\}\right).
\end{aligned}$$

From the symmetry of $\Phi(z)$, the cdf of Z, we have $P(\{Z \geq k\}) = P(\{Z \leq -k\})$, so it follows that

$$\begin{aligned}
\text{power} &= P\left(\left\{Z \leq -1.96 + (\theta_0 - \theta_1)\sqrt{n}\right\}\right) + P\left(\left\{Z \geq 1.96 + (\theta_0 - \theta_1)\sqrt{n}\right\}\right) \\
&= P\left(\left\{Z \leq -1.96 + (\theta_0 - \theta_1)\sqrt{n}\right\}\right) + P\left(\left\{Z \leq -\left(1.96 + (\theta_0 - \theta_1)\sqrt{n}\right)\right\}\right) \\
&= \Phi\left(-1.96 + (\theta_0 - \theta_1)\sqrt{n}\right) + \Phi\left(-1.96 - (\theta_0 - \theta_1)\sqrt{n}\right).
\end{aligned}$$

This function is plotted below.

For a desired power of 0.90, we have

$$0.90 = \Phi\left(-1.96 + (\theta_0 - \theta_1)\sqrt{n}\right) + \Phi\left(-1.96 - (\theta_0 - \theta_1)\sqrt{n}\right).$$

For specified values of θ_0 and θ_1, it is then possible to choose n such that the resulting expression is (approximately) equal to 0.90. (Because the function is monotone and n is discrete, there will be a single value of n such that the power of the function evaluated at n is greater than or equal to 0.90, but the power of the function evaluated at $n-1$ is less than 0.90.) We can now find the required sample size, implemented in the code below.

```
sample.size <- function(power, size, theta_diff, tails = 2) {
        z_crit <- qnorm(size / tails)
        n <- 1
        x <- 0
        while (x < power) {
                x <- pnorm(z_crit + theta_diff * sqrt(n)) +
                        pnorm(z_crit - theta_diff * sqrt(n))
```

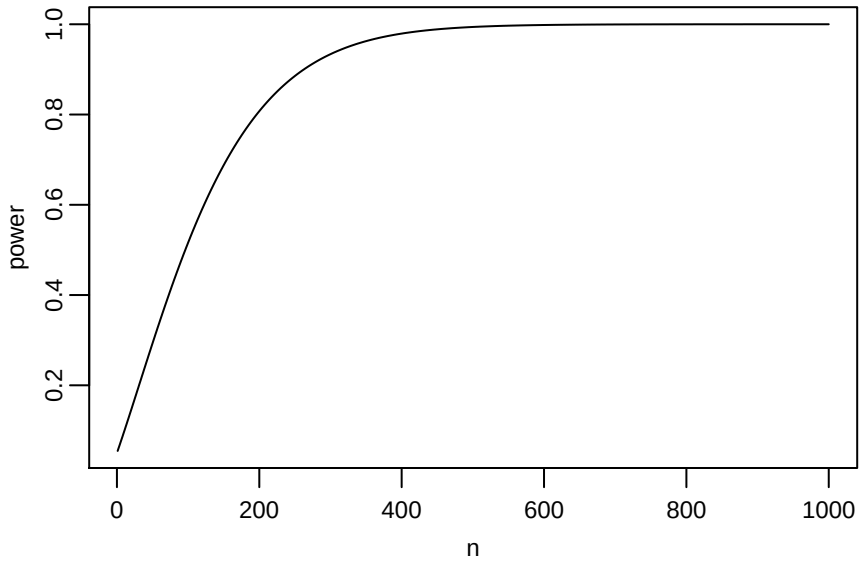


FIGURE 10.2.2. power function for $H_0 - H_1 = 0.2$

```
        n <- n + 1
        }
        return(n - 1)
}
```

Suppose that $\theta_0 - \theta_1 = 0.2$. Then, the required sample size such that the test has a power of 0.90 is 263.

EXAMPLE 10.20. Suppose that X is a discrete random variable whose distributions under H_0 and under H_1 are shown below. Use Neyman-Pearson to find the most powerful test of H_0 versus H_1 with $\alpha = 0.04$.

x	1	2	3	4	5	6	7
$f(x\|H_0)$	0.01	0.01	0.01	0.01	0.01	0.01	0.94
$f(x\|H_1)$	0.06	0.05	0.04	0.03	0.02	0.01	0.79

From theorem 10.17, the uniformly most powerful test of level α of H_0 versus H_1 will be one that rejects H_0 if

$$\frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > k \qquad \text{when } \mathbf{x} \in R$$

for some $k \geq 0$, where R denotes the rejection region, and $\alpha = P(\{\mathbf{X} \in R\})$. Noting that we have $\mathbf{X} = X$, the distribution of the ratio of the pmfs is

$$\frac{f(x|H_1)}{f(x|H_0)} = \begin{cases} 0.06/0.01 = 6, & x = 1 \\ 0.05/0.01 = 5, & x = 2 \\ 0.04/0.01 = 4, & x = 3 \\ 0.03/0.01 = 3, & x = 4 \\ 0.02/0.01 = 2, & x = 5 \\ 0.01/0.01 = 1, & x = 6 \\ 0.79/0.94 = 79/94, & x = 7 \end{cases},$$

so we must choose k such that the corresponding rejection region satisfies $P(\{X \in R\}) = 0.04$. Recalling that the size of a test is equal to the probability that it rejects H_0 given that H_0 is true and the power of a test is equal to the probability that it rejects H_0 when H_1 is true, the following table shows the possible choices

for k. Note that the outcomes are mutually exclusive, so we can simply add the corresponding probabilities, e.g.,

$$\begin{aligned} P(\{X \in \{1,2\} | H_0\}) &= P(\{X=1|H_0\} \cup \{X=2|H_0\}) \\ &= P(\{X=1|H_0\}) + P(\{X=2|H_0\}) \\ &= 0.01 + 0.01 \\ &= 0.02. \end{aligned}$$

k	R	size $= P(\{X \in R|H_0\})$	power $= P(\{X \in R|H_1\})$
> 6	$\emptyset$	0	0
$5 < k < 6$	$\{x : x = 1\}$	0.01	0.06
$4 < k < 5$	$\{x : x \in \{1,2\}\}$	0.02	0.11
$3 < k < 4$	$\{x : x \in \{1,2,3\}\}$	0.03	0.15
$2 < k < 3$	$\{x : x \in \{1,2,3,4\}\}$	0.04	0.18
$1 < k < 2$	$\{x : x \in \{1,2,3,4,5\}\}$	0.05	0.20
$79/94 < k < 1$	$\{x : x \in \{1,2,3,4,5,6\}\}$	0.06	0.21
$k < 79/94$	$\{x : x \in \{1,2,3,4,5,6,7\}\}$	1	1

Choosing $k \in (2,3)$ will give a test of the desired size ($\alpha = 0.04$), so it follows that a UMP level α test rejects H_0 if $X \in \{1,2,3,4\}$.

EXAMPLE 10.21. Let $X \sim \text{Binomial}(2,\theta)$. We want to test $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$. Find the power and size for different choices of rejection region.

From theorem 10.17, the uniformly most powerful test of level α of H_0 versus H_1 will be one that rejects H_0 if

$$\frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > k \qquad \text{when } \mathbf{x} \in R$$

for some $k \geq 0$, where R denotes the rejection region, and $\alpha = P_{\theta_0}(\{\mathbf{X} \in R\})$. Noting that we have $\mathbf{X} = X$, the ratio of the pmfs is

$$\frac{f(x|\theta_1)}{f(x|\theta_0)} = \frac{\binom{2}{x}\theta_1^x(1-\theta_1)^{2-x}}{\binom{2}{x}\theta_0^x(1-\theta_0)^{2-x}} = \frac{(3/4)^x(1/4)^{2-x}}{(1/2)^x(1/2)^{2-x}} = \frac{(3/4)^x}{(1/4)^x}\frac{(1/4)^2}{(1/2)^2} = 3^x\left(\frac{1}{2}\right)^2 = \frac{3^x}{4},$$

so a UMP test of level α will reject H_0 if $f(x|\theta_1)/f(x|\theta_0) = 3^x/4 > k$. Noting that

$$\frac{3^x}{4} = \begin{cases} 1/4, & x = 0 \\ 3/4, & x = 1 \\ 9/4, & x = 2 \end{cases}$$

and

$$\begin{aligned} P(\{X=0|H_0\}) &= \binom{2}{0}\left(\frac{1}{2}\right)^0\left(\frac{1}{2}\right)^2 = \frac{1}{4} \\ P(\{X=1|H_0\}) &= \binom{2}{1}\left(\frac{1}{2}\right)^1\left(\frac{1}{2}\right)^1 = \frac{1}{2} \\ P(\{X=2|H_0\}) &= \binom{2}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^0 = \frac{1}{4} \\ P(\{X=0|H_1\}) &= \binom{2}{0}\left(\frac{3}{4}\right)^0\left(\frac{1}{4}\right)^2 = \frac{1}{16} \end{aligned}$$

$$P(\{X=1|H_1\}) = \binom{2}{1}\left(\frac{3}{4}\right)^1\left(\frac{1}{4}\right)^1 = \frac{3}{8}$$
$$P(\{X=2|H_1\}) = \binom{2}{2}\left(\frac{3}{4}\right)^2\left(\frac{1}{4}\right)^0 = \frac{9}{16},$$

the possible choices for k are shown in the table below.

k	R	size $= P(\{X \in R\|H_0\})$	power $= P(\{X \in R\|H_1\})$
$> 9/4$	$\emptyset$	0	0
$3/4 < k < 9/4$	$\{x : x = 2\}$	$1/4$	$9/16$
$1/4 < k < 3/4$	$\{x : x \in \{1,2\}\}$	$3/4$	$15/16$
$k < 1/4$	$\{x : x \in \{0,1,2\}\}$	1	1

Note that it is not possible to construct a UMP size $\alpha = 0.05$ test (we will never reject H_0). Other level tests are possible, but the size of such a test may be unacceptably large, or the power of such a test unacceptably small.

EXAMPLE 10.22. Let X be a random sample from

$$f(x|\theta) = \theta x^{\theta-1} \qquad 0 < x < 1, \quad \theta > 0.$$

(a) Consider $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$ and reject H_0 if $X \geq 1/2$. Find the power and size of this test.

The power function for this test is given by

$$\beta(\theta) = P_\theta(\{X \in R\}) = P_\theta(\{X \geq 1/2\}) = \int_{1/2}^{1} f(x|\theta)\,\mathrm{d}\,x = \int_{1/2}^{1} \theta x^{\theta-1}\,\mathrm{d}\,x = x^\theta\Big|_{1/2}^{1} = 1^\theta - \left(\frac{1}{2}\right)^\theta = 1 - \frac{1}{2^\theta}.$$

This function may be evaluated at some particular value of θ, e.g., $\theta = 2$, to determine the power. The size of this test is given by

$$\alpha = \sup_{\theta\in\Theta_0} \beta(\theta) = \sup_{\theta\leq 1}\left(1 - \frac{1}{2^\theta}\right).$$

As θ increases from 0 to 1, $1/2^\theta$ decreases from 1 to $1/2$, so it follows that $\beta(\theta)$ attains its maximum under H_0 at $\theta = 1$, i.e., $\alpha = \sup_{\theta\in\Theta_0}\beta(\theta) = 1/2$.

(b) Consider $H_0 : \theta = 2$ versus $H_1 : \theta = 1$. Find the UMP size α test.

From theorem 10.17, the uniformly most powerful test of level α of H_0 versus H_1 will be one that rejects H_0 if

$$\frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > k \qquad \text{when } \mathbf{x} \in R$$

for some $k \geq 0$, where R denotes the rejection region, and $\alpha = P_{\theta_0}(\{\mathbf{X} \in R\})$. Noting that we have $\mathbf{X} = X$, the ratio of the pdfs is

$$\frac{f(x|H_1)}{f(x|H_0)} = \frac{f(x|\theta=1)}{f(x|\theta=2)} = \frac{x^0}{2x} = \frac{1}{2x},$$

so a UMP test of level α will reject H_0 if

$$\frac{f(x|H_1)}{f(x|H_0)} = \frac{1}{2x} > k \implies x < \frac{1}{2k}.$$

The power function for this test is

$$\beta(\theta) = P_\theta(\{X \in R\}) = P_\theta\left(X < \frac{1}{2k}\right) = \int_0^{1/(2k)} \theta x^{\theta-1}\,\mathrm{d}\,x = x^\theta\Big|_0^{1/(2k)} = \left(\frac{1}{2k}\right)^\theta - 0^\theta = \left(\frac{1}{2k}\right)^\theta.$$

Then, the size of this test is

$$\alpha = \sup_{\theta\in\Theta_0} \beta(\theta) = \sup_{\theta=2} \left(\frac{1}{2k}\right)^{\theta} = \left(\frac{1}{2k}\right)^{2} \implies \frac{1}{2k} = \sqrt{\alpha} \implies k = \frac{1}{2\sqrt{\alpha}},$$

so that the rejection region for the UMP size α test is

$$R = \left\{x : x < \frac{1}{2k}\right\} = \left\{x : x < \frac{1}{2\left(1/\left(2\sqrt{\alpha}\right)\right)}\right\} = \left\{x : x < \sqrt{\alpha}\right\}.$$

DEFINITION 10.23. A family of pdfs or pmfs $\{g(t|\theta) : \theta \in \Theta\}$ for a univariate random variable T with real-valued parameter θ has a *monotone likelihood ratio* (MLR) if, for every $\theta_2 > \theta_1$, $g(t|\theta_2)/g(t|\theta_1)$ is a monotone (nonincreasing or nondecreasing) function of t on $\{t : g(t|\theta_1) > 0 \text{ or } g(t|\theta_2) > 0\}$. Note that $c/0$ is defined as ∞ if $0 < c$.

Consider the family of pdfs $\mathcal{N}\left(\theta, 1^2\right)$, and let $\theta_2 > \theta_1$. Then, the ratio of the pdfs of $\mathcal{N}\left(\theta_2, 1^2\right)$ and $\mathcal{N}\left(\theta_1, 1^2\right)$ random variables is

$$\begin{aligned}
\frac{1/\left(\sqrt{2\pi}\right)\exp\left\{-\left(x-\theta_2\right)^2/2\right\}}{1/\left(\sqrt{2\pi}\right)\exp\left\{-\left(x-\theta_1\right)^2/2\right\}} &= \exp\left\{-\frac{\left(x-\theta_2\right)^2}{2}\right\}\exp\left\{\frac{\left(x-\theta_1\right)^2}{2}\right\} \\
&= \exp\left\{-\frac{\left(x-\theta_2\right)^2}{2} + \frac{\left(x-\theta_1\right)^2}{2}\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\left(x-\theta_2\right)^2 - \left(x-\theta_1\right)^2\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[x^2 - 2\theta_2 x + \theta_2^2 - \left(x^2 - 2\theta_1 x + \theta_1^2\right)\right]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(-2\theta_2 x + \theta_2^2 + 2\theta_1 x - \theta_1^2\right)\right\} \\
&= \exp\left\{-\frac{1}{2}\left[2x\left(\theta_1 - \theta_2\right) + \theta_2^2 - \theta_1^2\right]\right\} \\
&= \exp\left\{x\left(\theta_2 - \theta_1\right) + \frac{\theta_1^2 - \theta_2^2}{2}\right\} \\
&= \mathrm{e}^{x\left(\theta_2-\theta_1\right)}\mathrm{e}^{\left(\theta_1^2-\theta_2^2\right)/2}.
\end{aligned}$$

We have $\theta_1, \theta_2 \in \mathbb{R}$, so that $\mathrm{e}^{\left(\theta_1^2-\theta_2^2\right)/2} > 0$. We have $\theta_2 > \theta_1$, so that $\theta_2 - \theta_1 > 0$, and it follows that the ratio of the pdfs is nondecreasing in x.

Consider the family of pdfs Poisson(λ), and let $\lambda_2 > \lambda_1$. Then, the ratio of the pdfs of Poisson(λ_2) and Poisson(λ_1) random variables is

$$\frac{\mathrm{e}^{-\lambda_2}\lambda_2^x/x!}{\mathrm{e}^{-\lambda_1}\lambda_1^x/x!} = \frac{\lambda_2^x}{\lambda_1^x}\mathrm{e}^{\lambda_1-\lambda_2} = \left(\frac{\lambda_2}{\lambda_1}\right)^x \mathrm{e}^{\lambda_1-\lambda_2},$$

which is nondecreasing in x because $\lambda_2 > \lambda_1$, so that $\lambda_2/\lambda_1 > 1$ (note that $\lambda_1, \lambda_2 \geq 0$).

DEFINITION 10.24. A cdf F is *stochastically greater* than a cdf G if $F(x) \leq G(x)$ for all x, with strict inequality for some x.

This definition implies that if $X \sim F$, $Y \sim G$, then $P(\{X > x\}) \geq P(\{Y > x\})$ for all x, with strict inequality for some x. In other words, F gives more probability to greater values. Suppose that $F \sim \mathcal{N}\left(1, 1^2\right)$ and $G \sim \mathcal{N}\left(0, 1^2\right)$, whose plots are shown in figure 10.2.3.

EXAMPLE 10.25. Suppose that $X \sim F$, let f be a pdf for X such that f has a nondecreasing MLR, and let $\theta_2 > \theta_1$. Then,

$$\frac{\mathrm{d}}{\mathrm{d}\,x}\left[F(x|\theta_2) - F(x|\theta_1)\right] = f(x|\theta_2) - f(x|\theta_1) = \frac{f(x|\theta_2)\,f(x|\theta_1)}{f(x|\theta_1)} - f(x|\theta_1) = f(x|\theta_1)\left(\frac{f(x|\theta_2)}{f(x|\theta_1)} - 1\right).$$

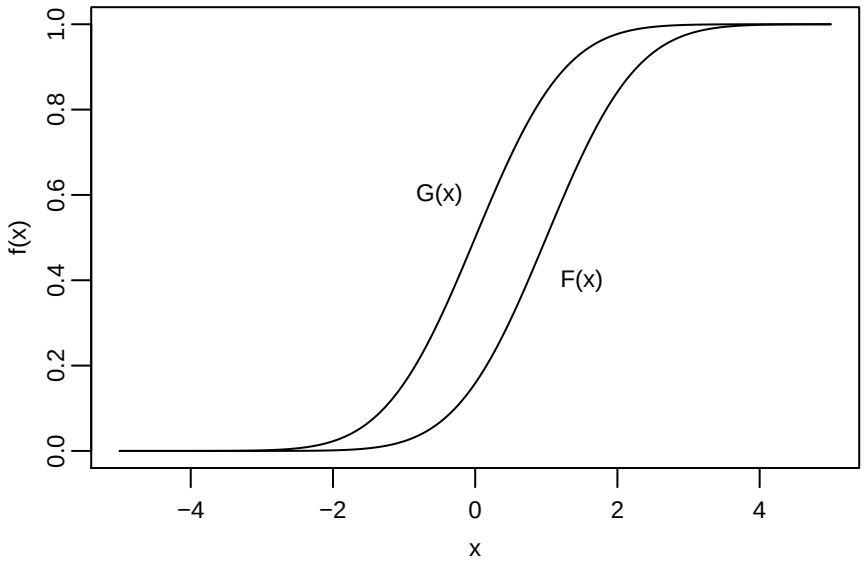


FIGURE 10.2.3. F is stochastically greater than G

Because f has a nondecreasing MLR, $f(x|\theta_2)/f(x|\theta_1)$ is increasing. Because f is a pdf, we have $f \geq 0$ for all x. It follows that, as x increases, $f(x|\theta_1)[(f(x|\theta_2)/f(x|\theta_1)) - 1]$ can only change sign from negative to positive (when the ratio exceeds 1), which implies that any interior extremum is a minimum. Thus, $F(x|\theta_2) - F(x|\theta_1)$ is maximized when $x = \pm\infty$, in which case it is equal to 0 (see theorem 2.4), i.e.,

$$\sup_{x\in\mathbb{R}}[F(x|\theta_2) - F(x|\theta_1)] = 0 \implies F(x|\theta_2) - F(x|\theta_1) \leq 0 \implies F(x|\theta_2) \leq F(x|\theta_1).$$

It follows that $F(x|\theta_2)$ is stochastically greater than $F(x|\theta_1)$. Further, for some constant c, we have

$$\begin{aligned} & F(c|\theta_2) \leq F(c|\theta_1) \\ \implies & P_{\theta_2}(\{X \leq c\}) \leq P_{\theta_1}(\{X \leq c\}) \\ \implies & 1 - P_{\theta_2}(\{X > c\}) \leq 1 - P_{\theta_1}(\{X > c\}) \\ \implies & P_{\theta_1}(\{X > c\}) \leq P_{\theta_2}(\{X > c\}). \end{aligned}$$

THEOREM 10.26 (Karlin-Rubin). *Consider testing $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$. Suppose that T is a sufficient statistic for θ and the family of pdfs or pmfs $\{g(t|\theta) : \theta \in \Theta\}$ of T has a nondecreasing MLR. Then for any t_0, the test that rejects H_0 if and only if $T > t_0$ is a UMP level α test, where $\alpha = P_{\theta_0}(\{T > t_0\})$. (This is Theorem 8.3.17 from Casella & Berger; the following proof is given there.)*

PROOF. Let $\beta(\theta) = P_\theta(\{T > t_0\})$ be the power function of the test. Fix $\theta' > \theta_0$ and consider testing $H_0' : \theta = \theta_0$ versus $H_1' : \theta = \theta'$. Since the family of pdfs or pmfs of T has a nondecreasing MLR, it follows from example 10.25 that$\beta(\theta)$ is nondecreasing, so

i. $\sup_{\theta\leq\theta_0}\beta(\theta) = \beta(\theta_0) = \alpha$, and this is a level α test.
ii. If we define

$$k' = \inf_{t\in T}\frac{g(t|\theta')}{g(t|\theta_0)},$$

where $\mathcal{T} = \{t : t > t_0 \text{ and either } g(t|\theta') > 0 \text{ or } g(t|\theta_0) > 0\}$, it follows that

$$T > t_0 \implies \frac{g(t|\theta')}{g(t|\theta_0)} > k'.$$

Together with corollary 10.18, (i) and (ii) imply that $\beta(\theta') \geq \beta^*(\theta')$, where $\beta^*(\theta)$ is the power function for any other level α test of H_0', that is, any test satisfying $\beta(\theta_0) \leq \alpha$. However, any level α test of H_0 satisfies $\beta^*(\theta_0) \leq \sup_{\theta\in\Theta_0}\beta^*(\theta) \leq \alpha$. Thus, $\beta(\theta') \geq \beta^*(\theta')$ for any level α test of H_0. Since θ' was arbitrary, the test is a UMP level α test.

By an analogous argument, it can be shown that under the conditions of theorem 10.26, the test that rejects $H_0 : \theta \geq \theta_0$ in favor of $H_1 : \theta < \theta_0$ if and only if $T < t_0$ is a UMP level $\alpha = P_{\theta_0}(\{T < t_0\})$ test. □

EXAMPLE 10.27. Let $X_1, \ldots, X_n \sim \mathcal{U}(0, \theta)$. Consider the hypothesis test

$$H_0 : \theta \leq \theta_0 \quad \text{vs.} \quad H_1 : \theta > \theta_0.$$

From example 8.8, $X_{(n)}$ is sufficient for θ, and a pdf for $X_{(n)}$ is given by

$$f_{X_{(n)}}(t) = nt^{n-1}\left(\frac{1}{\theta^n}\right) I_{(0,\theta)}(t).$$

Let $\theta_2 > \theta_1$, so that the ratio of the pdfs of $f_{X_{(n)}}(t|\theta_2)$ and $f_{X_{(n)}}(t|\theta_1)$ is

$$\frac{f_{X_{(n)}}(t|\theta = \theta_2)}{f_{X_{(n)}}(t|\theta = \theta_1)} = \frac{\left(nt^{n-1}/\theta_2^n\right) I_{(0,\theta_2)}(t)}{\left(nt^{n-1}/\theta_1^n\right) I_{(0,\theta_1)}(t)} = \left(\frac{\theta_1}{\theta_2}\right)^n \frac{I_{(0,\theta_2)}(t)}{I_{(0,\theta_1)}(t)} = \begin{cases} (\theta_1/\theta_2)^n, & t < \theta_1 \\ \infty, & \theta_1 < t < \theta_2 \end{cases},$$

which is nondecreasing in t. Therefore, this family has MLR, and by theorem 10.26, the UMP level α test rejects H_0 if $X_{(n)} > k$. We have

$$\begin{aligned} \alpha &= P_{\theta_0}\left(\{X_{(n)} > k\}\right) \\ &= \int_k^\infty f_{X_{(n)}}(t|\theta = \theta_0)\,\mathrm{d}\,t \\ &= \int_k^\infty \frac{nt^{n-1}}{\theta_0^n} I_{(0,\theta_0)}(t)\,\mathrm{d}\,t \\ &= \int_k^{\theta_0} \frac{nt^{n-1}}{\theta_0^n}\mathrm{d}t + \int_{\theta_0}^\infty 0\,\mathrm{d}\,t \\ &= \frac{t^n}{\theta_0^n}\Big|_k^{\theta_0} + 0 \\ &= \frac{\theta_0^n}{\theta_0^n} - \frac{k^n}{\theta_0^n} \\ &= 1 - \frac{k^n}{\theta_0^n}, \end{aligned}$$

so that

$$\alpha = 1 - \frac{k^n}{\theta_0^n} \implies 1 - \alpha = \frac{k^n}{\theta_0^n} \implies k^n = \theta_0^n(1-\alpha) \implies k = \left[\theta_0^n(1-\alpha)\right]^{1/n} = \theta_0(1-\alpha)^{1/n}.$$

Then, the UMP level α test rejects H_0 if $X_{(n)} > \theta_0(1-\alpha)^{1/n}$.

EXAMPLE 10.28. X is a single observation from

$$f(x|\theta) = 2\theta x + 1 - \theta, \qquad 0 \leq x \leq 1, \qquad -1 \leq \theta \leq 1.$$

(1) Find the most powerful test of $H_0 : \theta = 0$ vs. $H_1 : \theta = 1$.

From theorem 10.17, the uniformly most powerful test of level α of H_0 versus H_1 will be one that rejects H_0 if

$$\frac{f(\mathbf{x}|H_1)}{f(\mathbf{x}|H_0)} > k \qquad \text{when } \mathbf{x} \in R$$

for some $k \geq 0$, where R denotes the rejection region, and $\alpha = P_{\theta_0}(\{\mathbf{X} \in R\})$. Noting that we have $\mathbf{X} = X$, the ratio of the pdfs is

$$\frac{f(x|\theta = 1)}{f(x|\theta = 0)} = \frac{2x + 1 - 1}{0 + 1 - 0} = 2x,$$

so we will reject H_0 if $2x > k \Rightarrow x > k/2$. We have

$$\begin{aligned} \alpha &= P_{\theta_0}(\{\mathbf{X} \in R\}) \\ &= P_{\theta_0}\left(\left\{X > \frac{k}{2}\right\}\right) \end{aligned}$$

$$
\begin{aligned}
&= \int_{k/2}^{\infty} f\left(x|\theta=\theta_0\right) \mathrm{d}\, x \\
&= \int_{k/2}^{1} 2\theta_0 x + 1 - \theta_0 \,\mathrm{d}\, x + \int_{1}^{\infty} 0 \,\mathrm{d}\, x \\
&= \int_{k/2}^{1} 0 + 1 - 0 \,\mathrm{d}\, x + 0 \\
&= \int_{k/2}^{1} 1 \,\mathrm{d}\, x \\
&= x|_{k/2}^{1} \\
&= 1 - \frac{k}{2},
\end{aligned}
$$

so that

$$\alpha = 1 - \frac{k}{2} \implies \frac{k}{2} = 1 - \alpha.$$

Then, the UMP level α test rejects H_0 if $X > 1 - \alpha$.

(2) Consider $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$ and reject if $X > 1/2$. Find the power and size of this test.

The power function for this test is

$$
\begin{aligned}
\beta\left(\theta\right) &= P_\theta\left(\left\{X \in R\right\}\right) \\
&= P_\theta\left(\left\{X > \frac{1}{2}\right\}\right) \\
&= \int_{1/2}^{1} 2\theta x + 1 - \theta \,\mathrm{d}\, x \\
&= \theta x^2 + x - \theta x|_{1/2}^{1} \\
&= \theta + 1 - \theta - \left(\frac{\theta}{4} + \frac{1}{2} - \frac{\theta}{2}\right) \\
&= 1 - \left(\frac{1}{2} - \frac{\theta}{4}\right) \\
&= \frac{1}{2} + \frac{\theta}{4}.
\end{aligned}
$$

The power of a test is equal to the probability that it rejects the null hypothesis given that the alternative hypothesis is true. Then, the power of this test for a particular alternative $H_1 : \theta = \theta_1$ is

$$P_{\theta_1}\left(\left\{X \in R\right\}\right) = P\left(\left\{X > \frac{1}{2}|\theta = \theta_1\right\}\right) = \beta\left(\theta_1\right) = \frac{1}{2} + \frac{\theta_1}{4}.$$

Noting that the power function for this test is increasing in θ, the size of this test is

$$\alpha = \sup_{\theta \in \Theta_0} \beta\left(\theta\right) = \sup_{\theta \leq 0} \beta\left(\theta\right) = \beta\left(0\right) = \frac{1}{2} + 0 = \frac{1}{2}.$$

(3) Find the UMP test for $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$.

Let $g\left(T\left(x\right)|\theta\right) = 2\theta x + 1 - \theta$. Then, setting $h\left(x\right) = 1$, it follows from theorem 8.11 that $T\left(X\right) = X$ is sufficient for θ. Let $\theta_2 > \theta_1$, so that the ratio of the pdfs of $f\left(x|\theta_2\right)$ and f$(x|\theta_1)$ is

$$\frac{f\left(x|\theta=\theta_2\right)}{f\left(x|\theta=\theta_1\right)} = \frac{2\theta_2 x + 1 - \theta_2}{2\theta_1 x + 1 - \theta_1}.$$

Taking the derivative with respect to x gives

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\, x}\frac{2\theta_2 x + 1 - \theta_2}{2\theta_1 x + 1 - \theta_1} &= 2\theta_2\left(2\theta_1 x + 1 - \theta\right)^{-1} + \left(2\theta_2 x + 1 - \theta_2\right)\left[-\left(2\theta_1 x + 1 - \theta_1\right)^{-2}\left(2\theta_1\right)\right] \\
&= \frac{2\theta_2}{2\theta_1 x + 1 - \theta_1} - \frac{2\theta_1\left(2\theta_2 x + 1 - \theta_2\right)}{\left(2\theta_1 x + 1 - \theta_1\right)^2}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{2\theta_2\left(2\theta_1 x + 1 - \theta_1\right) - 2\theta_1\left(2\theta_2 x + 1 - \theta_2\right)}{\left(2\theta_1 x + 1 - \theta_1\right)^2} \\
&= \frac{2\left(2\theta_1\theta_2 x + \theta_2 - \theta_1\theta_2 - 2\theta_1\theta_2 x - \theta_1 + \theta_1\theta_2\right)}{\left(2\theta_1 x + 1 - \theta_1\right)^2} \\
&= \frac{2\left(\theta_2 - \theta_1\right)}{\left(2\theta_1 x + 1 - \theta_1\right)^2}.
\end{aligned}
$$

We have $\theta_2 > \theta_1$, so the numerator of this expression is positive. The denominator is also positive, so it follows that this expression is non-negative. Therefore, this ratio is nondecreasing in x and this family of pdfs has a monotone likelihood ratio. It follows from theorem 10.26 that the UMP test rejects H_0 if $X > k$. The size of this test is

$$
\begin{aligned}
\alpha &= \sup_{\theta\in\Theta_0} \beta(\theta) \\
&= \sup_{\theta\leq 0} P_\theta(\{X > k\}) \\
&\qquad (\beta(\theta) \text{ is increasing in } \theta) \\
&= P(\{X > k | \theta = 0\}) \\
&= \int_k^\infty f(x|\theta = 0)\,\mathrm{d}\,x \\
&= \int_k^1 2\cdot 0\cdot x + 1 - 0\,\mathrm{d}\,x + \int_1^\infty 0\,\mathrm{d}\,x \\
&= \int_k^1 1\,\mathrm{d}\,x + 0 \\
&= x|_k^1 \\
&= 1 - k,
\end{aligned}
$$

which implies $k = 1 - \alpha$. Then, the UMP level α test rejects H_0 if $X > 1 - \alpha$.

(4) Find the LRT for $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$.

Noting that $\mathbf{X} = X$, the likelihood function is given by

$$\mathcal{L}(\theta|\mathbf{x}) = 2\theta x + 1 - \theta = 1 + \theta(2x - 1),$$

and the LRT test statistic is

$$
\begin{aligned}
\lambda(\mathbf{x}) &= \frac{\sup_{\Theta_0}\mathcal{L}(\theta|\mathbf{x})}{\sup_{\Theta}\mathcal{L}(\theta|\mathbf{x})} \\
&= \frac{\sup_{\theta=0}\mathcal{L}(\theta|\mathbf{x})}{\sup_{\theta\in\Theta}\mathcal{L}(\theta|\mathbf{x})} \\
&= \frac{1 + 0\cdot(2x-1)}{\sup_{\theta\in\Theta}\mathcal{L}(\theta|x)} \\
&= \frac{1}{\sup_{\theta\in\Theta}\mathcal{L}(\theta|x)}.
\end{aligned}
$$

We see that

$$0 \leq x < \frac{1}{2} \implies 2x - 1 < 0 \quad \text{and} \quad \frac{1}{2} < x \leq 1 \implies 2x - 1 > 0.$$

Then, for $x \in \left[0, \frac{1}{2}\right)$, $\mathcal{L}(\theta|x)$ is decreasing in θ and attains its maximum value at $\hat{\theta} = -1$. For $x \in \left(\frac{1}{2}, 1\right]$, $\mathcal{L}(\theta|x)$ is increasing in θ and attains its maximum value at $\hat{\theta} = 1$. Thus,

$$\sup_{\theta\in\Theta}\mathcal{L}(\theta|x) = \begin{cases} 1 + (-1)(2x-1) = 2 - 2x, & x < \frac{1}{2} \\ 1 + 1(2x-1) = 2x, & x > \frac{1}{2} \end{cases},$$

and the LRT test statistic is

$$\lambda(x) = \begin{cases} \frac{1}{2-2x}, & x < \frac{1}{2} \\ \frac{1}{2x}, & x > \frac{1}{2} \end{cases}.$$

Then, the likelihood ratio test rejects H_0 if

$$c \geq \lambda(x) \implies \text{reject } H_0 \text{ if } \begin{cases} \frac{1}{2-2x} \leq c, & x < \frac{1}{2} \\ \frac{1}{2x} \leq c, & x > \frac{1}{2} \end{cases}$$

for some $c \in (0,1)$. I.e., we will reject H_0 if

$$\frac{1}{2-2x} \leq c \implies 1 \leq c(2-2x) \implies \frac{1}{c} \leq 2-2x \implies \frac{1}{c} - 2 \leq -2x \implies x \leq 1 - \frac{1}{2c}$$

for $x < \frac{1}{2}$ and

$$\frac{1}{2x} \leq c \implies 1 \leq c(2x) \implies x \geq \frac{1}{2c}$$

for $x > \frac{1}{2}$. Then, size α test is given by

$$\begin{aligned}
\alpha &= P_{\theta_0}(\{X \in R\}) \\
&= P_{\theta=0}\left(\left(\left\{X \leq 1 - \frac{1}{2c}\right\} \cap \left\{X < \frac{1}{2}\right\}\right) \cup \left(\left\{X \geq \frac{1}{2c}\right\} \cap \left\{X > \frac{1}{2}\right\}\right)\right) \\
&= P_{\theta=0}\left(\left\{0 \leq X \leq 1 - \frac{1}{2c}\right\} \cup \left\{\frac{1}{2c} \leq X \leq 1\right\}\right) \\
&= P_{\theta=0}\left(\left\{0 \leq X \leq 1 - \frac{1}{2c}\right\}\right) + P_{\theta=0}\left(\left\{\frac{1}{2c} \leq X \leq 1\right\}\right) \\
&= \int_0^{1-1/(2c)} f(x|\theta=0)\,\mathrm{d}\,x + \int_{1/(2c)}^1 f(x|\theta=0)\,\mathrm{d}\,x \\
&= \int_0^{1-1/(2c)} 2 \cdot 0 \cdot x + 1 - 0\,\mathrm{d}\,x + \int_{1/(2c)}^1 2 \cdot 0 \cdot x + 1 - 0\,\mathrm{d}\,x \\
&= \int_0^{1-1/(2c)} 1\,\mathrm{d}\,x + \int_{1/(2c)}^1 1\,\mathrm{d}\,x \\
&= x\Big|_0^{1-1/(2c)} + x\Big|_{1/(2c)}^1 \\
&= 1 - \frac{1}{2c} - 0 + 1 - \frac{1}{2c} \\
&= 2 - \frac{1}{c},
\end{aligned}$$

so that

$$\alpha = 2 - \frac{1}{c} \implies \frac{1}{c} = 2 - \alpha \implies c(2-\alpha) = 1 \implies c = \frac{1}{2-\alpha}.$$

Then, we will reject H_0 if

$$x \leq 1 - \frac{1}{2c} = 1 - \frac{1}{2\left(\frac{1}{2-\alpha}\right)} = 1 - \frac{2-\alpha}{2} = 1 - \left(\frac{2}{2} - \frac{\alpha}{2}\right) = 1 - 1 + \frac{\alpha}{2} = \frac{\alpha}{2}$$

for $x < \frac{1}{2}$ and

$$x \geq \frac{1}{2c} = \frac{1}{2\left(\frac{1}{2-\alpha}\right)} = \frac{2-\alpha}{2} = 1 - \frac{\alpha}{2}$$

for $x > \frac{1}{2}$. Then, the size α LRT rejects H_0 if $X \geq 1 - \frac{\alpha}{2}$ or $X \leq \frac{\alpha}{2}$.

10.2.3. p-values. One method of reporting the results of a hypothesis test is to report the size, α, of the test used and the decision to reject H_0 or accept H_0. If α is small, the decision to reject H_0 is fairly convincing, but if α is large, the decision to reject H_0 is not very convincing because the test has a large probability of incorrectly making that decision. Another way of reporting the results of a hypothesis test is to report the value of a certain kind of test statistic called a *p-value*.

DEFINITION 10.29. A *p-value* $p(\mathbf{X})$ is a test statistic satisfying $0 \leq p(\mathbf{x}) \leq 1$ for every sample point $\mathbf{x}$. Small values of $p(\mathbf{X})$ give evidence that H_1 is true. A p-value is *valid* if, for every $\theta \in \Theta_0$ and every $0 \leq \alpha \leq 1$,

$$P_\theta(\{p(\mathbf{X}) \leq \alpha\}) \leq \alpha.$$

A p-value is the probability of observing a more "extreme" result than was observed under H_0. A p-value can also be viewed as the minimum α-level at which H_0 could have been rejected with the observed data. The most common way to define a valid p-value is given in the following theorem.

THEOREM 10.30. *Let $W(\mathbf{X})$ be a test statistic such that large values of W give evidence that H_1 is true. For each sample point $\mathbf{x}$, define*

$$p(\mathbf{x}) = \sup_{\theta \in \Theta_0} P_\theta(\{W(\mathbf{X}) \geq W(\mathbf{x})\}).$$

Then, $p(\mathbf{X})$ is a valid p-value. (This is Theorem 8.3.27 from Casella & Berger; the following proof is given there.)

PROOF. Fix $\theta \in \Theta_0$. Let $F_\theta(w)$ denote the cdf of $-W(\mathbf{X})$. Define

$$p_\theta(\mathbf{x}) = P_\theta(\{W(\mathbf{X}) \geq W(\mathbf{x})\}) = P_\theta(\{-W(\mathbf{X}) \leq -W(\mathbf{x})\}) = F_\theta(-W(\mathbf{x})).$$

Then the random variable $p_\theta(\mathbf{X})$ is equal to $F_\theta(-W(\mathbf{X}))$. Hence, by theorem 3.3, the distribution of $p_\theta(\mathbf{X})$ is stochastically greater than or equal to a $U(0,1)$ distribution. That is, for every $0 \leq \alpha \leq 1$, $P_\theta(\{p_\theta(\mathbf{X}) \leq \alpha\}) \leq \alpha$. Because $p(\mathbf{x}) = \sup_{\theta' \in \Theta_0} p_{\theta'}(\mathbf{x}) \geq p_\theta(\mathbf{x})$ for every $\mathbf{x}$,

$$P_\theta(\{p(\mathbf{X}) \leq \alpha\}) \leq P_\theta(\{p_\theta(\mathbf{X}) \leq \alpha\}) \leq \alpha.$$

This is true for every $\theta \in \Theta_0$ and for every $0 \leq \alpha \leq 1$; $p(\mathbf{X})$ is a valid p-value. □

EXAMPLE 10.31. Let $n = 25$ and $X_1, \ldots, X_n \sim \mathcal{N}(\theta, 5^2)$. Consider the hypothesis test

$$H_0 : \theta = 17 \quad \text{vs.} \quad H_1 : \theta > 17.$$

Find the p-value of the test that rejects H_0 if $\bar{X} > 17 + 5/\sqrt{n}$

(1) when we observe $\bar{x} = 18$.

By theorem 5.3, the test statistic $\bar{X}$ has the distribution of a

$$\mathcal{N}\left(\theta, \left(\theta/\sqrt{n}\right)^2\right) = \mathcal{N}\left(\theta, \left(5/\sqrt{25}\right)^2\right) = \mathcal{N}(\theta, 1^2)$$

random variable, so it follows that

$$Z = \frac{\bar{X} - \theta}{5/\sqrt{n}} = \frac{\bar{X} - \theta}{5/\sqrt{25}} = \frac{\bar{X} - \theta}{1} = \bar{X} - \theta$$

has the distribution of a $\mathcal{N}(0, 1^2)$ random variable, and that $\bar{X} = Z + \theta$. The size of this test is

$$\begin{aligned}
\alpha &= P_{\theta_0}(\{\mathbf{X} \in R\}) \\
&= P_{\theta_0}\left(\left\{\bar{X} > 17 + \frac{5}{\sqrt{n}}\right\}\right) \\
&= P\left(\left\{\bar{X} > 17 + \frac{5}{\sqrt{n}} | \theta = 17\right\}\right) \\
&= P\left(\left\{\frac{\bar{X} - \theta}{\sigma/\sqrt{n}} > \frac{17 + 5/\sqrt{n} - \theta}{\sigma/\sqrt{n}} | \theta = 17\right\}\right) \\
&= P\left(\left\{Z > \frac{17 + 5/\sqrt{25} - 17}{5/\sqrt{25}}\right\}\right) \\
&= P(\{Z > 1\}) \\
&\approx 0.1586553.
\end{aligned}$$

For a test of size $\alpha = 0.05$, we have

$$0.05 = P_{\theta_0}(\{\bar{X} > c\})$$

$$
\begin{aligned}
&= P\left(\left\{\bar{X} > c|\theta = 17\right\}\right) \\
&= P\left(\left\{\frac{\bar{X}-\theta}{\sigma/\sqrt{n}} > \frac{c-17}{5/\sqrt{25}}\right\}\right) \\
&= P\left(\{Z > c-17\}\right) \\
&= 1 - P\left(\{Z \leq c-17\}\right) \\
&= 1 - \Phi\left(c-17\right),
\end{aligned}
$$

which implies

$$\Phi\left(c-17\right) = 1 - 0.05 = 0.95 \implies c - 17 \approx 1.6448536 \implies c \approx 18.6448536.$$

Then, a size $\alpha = 0.05$ test will reject H_0 if $\bar{X} > 18.6448536$. For our observed $\bar{x} = 18$, the p-value is

$$
\begin{aligned}
p\left(\mathbf{x}\right) &= \sup_{\theta \in \Theta_0} P_\theta\left(\left\{\bar{X} \geq \bar{x}\right\}\right) \\
&= \sup_{\theta = 17} P_\theta\left\{\bar{X} \geq 18\right\} \\
&= P\left(\left\{\bar{X} \geq 18|\theta = 17\right\}\right) \\
&= P\left(\left\{\bar{X} - \theta \geq 18 - 17\right\}\right) \\
&= P\left(\{Z \geq 1\}\right) \\
&= 1 - \Phi\left(1\right) \\
&\approx 0.1586553.
\end{aligned}
$$

At $\alpha = 0.05$, we do not reject H_0 because this p-value is greater than α.

(2) when we observe $\bar{x} = 19$.

The p-value is

$$
\begin{aligned}
p\left(\mathbf{x}\right) &= \sup_{\theta \in \Theta_0} P_\theta\left(\left\{\bar{X} \geq \bar{x}\right\}\right) \\
&= \sup_{\theta = 17} P_\theta\left\{\bar{X} \geq 19\right\} \\
&= P\left(\left\{\bar{X} \geq 19|\theta = 17\right\}\right) \\
&= P\left(\left\{\bar{X} - \theta \geq 19 - 17\right\}\right) \\
&= P\left(\{Z \geq 2\}\right) \\
&= 1 - \Phi\left(2\right) \\
&\approx 0.0227501.
\end{aligned}
$$

At $\alpha = 0.05$, we reject H_0 because this p-value is greater than α.

Part 3

Bayesian statistics

CHAPTER 11

Introduction to Bayesian paradigm

EXAMPLE 11.1 (Paternity dispute). Suppose you are on a jury considering a paternity suit brough by Suzy Smith's mother against Al Edged. Suzy's mother has blood type O and Al Edged is type AB. You have other information as well. You hear testimony concerning whether Al Edged and Suzy's mother had sexual intercourse during the time that conception could have occurred, about the timing and frequency of such intercourse, about Al Edged's fertility, about the possibility that someone else is the father, and so on. You put all this information together in assessing $P(F)$, your probability that Al is Suzy's father. The evidence of interest is Suzy's blood type, which turns out to be B (if it were O, Al Edged would be excluded from paternity). According to Mendelian genetics, $P(B|F) = 1/2$. You also accept the blood bank's estimate $P(B|F^c) = 0.09$. According to Bayes' rule,

$$P(F|B) = \frac{P(B|F)\,P(F)}{P(B|F)\,P(F) + P(B|F^c)\,P(F^c)} = \frac{(1/2)\,P(F)}{(1/2)\,P(F) + 0.09P(F^c)}.$$

The relationship between our prior probability, $P(F)$, and our posterior probability, $P(F|B)$, may be summarized:

$P(F)$	0	0.100	0.250	0.500	0.750	0.900	1
$P(F\|B)$	0	0.382	0.649	0.847	0.943	0.980	1

DEFINITION 11.2. Let $p(y_1, \ldots, y_n)$ be the joint distribution of $Y_1, \ldots, Y_n$ and let $\pi_1, \ldots, \pi_n$ be a permutation of the indices $1, \ldots, n$. If $p(y_1, \ldots, y_n) = p(y_{\pi_1}, \ldots, y_{\pi_n})$ for all permutations, then $Y_1, \ldots, Y_n$ are *exchangeable*.

THEOREM 11.3 (de Finetti's). *Let $Y_1, Y_2, \ldots$ be a sequence of random variables. If for any n, $Y_1, \ldots, Y_n$ are exchangeable, then there exists a prior distribution $p(\theta)$ and sampling model $p(y|\theta)$ such that*

$$p(y_1, \ldots, y_n) = \int_\Theta \left\{ \prod_{i=1}^{n} p(y_i|\theta) \right\} p(\theta)\,\mathrm{d}\,\theta.$$

EXAMPLE 11.4 (Estimating the probability of a rare event). Suppose we are interested in the prevalance of an infectious disease in a small city. The higher the prevalance, the more public health precautions would need to be put into place. A small random sample of 20 individuals from the city is checked for infection. Interest is in θ, the fraction of infected individuals in the city. The data y records the total number of people in the sample who are infected. The parameter space and sample space are then

$$\Theta = [0, 1], \qquad \mathcal{Y} = \{0, 1, \ldots, 20\}.$$

Before the sample is obtained, the number of infected individuals in the sample is unknown. If the value of θ were known, a reasonable sampling model for Y would be a $\mathrm{Bin}(20, \theta)$ probability distribution, i.e., $Y|\theta \sim \mathrm{Bin}(20, \theta)$. If, for example, the true infection rate is $\theta = 0.05$, then the probability that there will be zero infected individuals in the sample is

$$P(Y = 0) = \binom{20}{0} (0.05)^0 (1 - 0.05)^{20} = 0.36.$$

```
n <- 20
x <- 0:n
del = 0.25
par(mgp = c(1.5,0.5,0), mar = c(2.5,3,0.5,2))
```

```
# finish up the code
```

Other studies from various parts of the country indicate that the infection rate in comparable cities ranges from about 0.05 to 0.20, with an average prevalance of 0.10. We will use a prior distribution $p(\theta)$ that has these characteristics and provides computational convenience. Specifically, we will take a Beta distribution

$$\theta \sim \text{Beta}(a,b), \quad p(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}, \quad 0 \leq \theta \leq 1.$$

The expectation $\text{E}[\theta] = a/(a+b)$ and the mode of θ is $(a-1)/(a+b-2)$. We will represent our information about θ with $\text{Beta}(2,20)$ probability distribution. Then,

$$\text{E}[\theta] = \frac{2}{2+20} = \frac{1}{11} \quad \text{and} \quad \text{mode}(\theta) = \frac{2-1}{2+20-2} = \frac{1}{20}.$$

Suppose for our study, a value $Y = y$ is observed. The posterior distribution of θ is given by

$$p(\theta|y) = \frac{p(y|\theta)\,p(\theta)}{\int_\Theta p\left(y|\tilde{\theta}\right) p\left(\tilde{\theta}\right) \text{d}\,\tilde{\theta}}.$$

In deriving posterior densities, an oft-used technique is to try and recognize the kernel of the posterior density of θ. We have

$$\begin{aligned} p(\theta|y) &= \frac{\left[\binom{n}{y}\theta^y(1-\theta)^{n-y}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right]}{\int_\Theta \left[\binom{n}{y}\theta^y(1-\theta)^{n-y}\right]\left[\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1}\right]\text{d}\,\theta} \\ &\propto \theta^{y+a-1}(1-\theta)^{n-y+b-1}, \end{aligned}$$

which we recognize as the kernel of a $\text{Beta}(y+a, n-y+b)$ random variable. Thus,

$$\theta|y \sim \text{Beta}(y+a, n-y+b) \quad \text{and} \quad p(\theta|y) = \frac{\Gamma(n+a-b)}{\Gamma(y+a)\Gamma(n-y+b)}\theta^{y+a-1}(1-\theta)^{n-y+b-1}.$$

Suppose that for our study a value of $Y = 0$ is observed, i.e., none of the sample individuals are infected. The posterior distribution of θ is then $\theta|Y = 0 \sim \text{Beta}(2,40)$. This density is further to the left than the prior distribution, and more peaked as well. The posterior distribution $p(\theta|Y=0)$ provides us with a model for learning about the city-wide infection rate θ. Suppose we are supposed to discuss the results of the survey with a group of city health officials. We might want to present the posterior results corresponding to a variety of prior distributions. The posterior expectation is a weighted average of the sample mean $\bar{y}$ and the prior expectation θ_0, i.e.,

$$\text{E}[\theta|Y=y] = \frac{a+y}{a+b+n} = \frac{n}{a+b+n}\frac{y}{n} + \frac{a+b}{a+b+n}\frac{a}{a+b} = \frac{n}{w+n}\bar{y} + \frac{w}{w+n}\theta_0,$$

where $\theta_0 = a/(a+b)$ and $w = a+b$. θ_0 represents our prior guess at the true value of θ and w represents our confidence in this guess, expressed on the same scale as the sample size. We can compute such a posterior distribution for a wide range of θ_0 and w values to perform a *sensitivity analysis* (an exploration of how posterior information is affected by differences in prior opinion).

11.1. Bayesian learning

The Bayesian perspective does not consider data that could have been observed, but is not. In the Bayesian approach, instead of supposing that θ is a fixed parameter, it is regarded as the realized value of a random variable Θ, with density $\pi(\theta)$ (the *prior* distribution). The prior distribution summarizes any information we have about θ (not related to that provided by the data y). Bayesian inference refers to the updating of prior beliefs into *posterior* beliefs conditional on observed data using Bayes' theorem.

CHAPTER 12

Single-parameter models

Various numerical summaries of the posterior distribution $p(\theta|\mathbf{y})$ may be used, each of which corresponds to minimizing a different loss function. The posterior mean is the posterior expectation of the parameter, the mode is the "most likely" value, and the standard deviation or interquartile range summarize the variation. In addition to point summaries, it is always important to report posterior uncertainty.

DEFINITION 12.1. A Bayesian $100(1-\alpha)\%$ *credible interval* for the parameter θ is constructed by removing the upper and lower $100(\alpha/2)$ percentiles of the posterior distribution.

We can obtain a credible interval by finding (θ_l, θ_u) that satisfy

$$P(\theta < \theta_l|\mathbf{y}) = \alpha/2 \quad \text{and} \quad P(\theta > \theta_u|\mathbf{y}) = \alpha/2.$$

Then, (θ_l, θ_u) is a $100(1-\alpha)\%$ credible interval. The credible interval has the specified probability coverage and has a natural interpretation, e.g., "we have 95% posterior confidence that θ falls in the interval." We can obtain a 95% credible interval for the infection rate from example 11.4 by identifying the values that cut off the upper and lower 2.5% of the posterior distribution.

```
qbeta(c(0.025,0.975), 2, 40)
```

```
## [1] 0.005963118 0.128554020
```

Thus, a 95% credible interval for θ is given by $(0.006, 0.129)$. That is, we are 95% confident *a posteriori* that θ lies in the interval $(0.006, 0.129)$.

DEFINITION 12.2. The $100(1-\alpha)\%$ *highest posterior density (HPD) interval* is defined as the interval that contains $100(1-\alpha)\%$ of the highest area under the posterior density function.

For symmetric, unimodal, and concave distributions, HPD intervals and credible intervals are equivalent. For skewed or multimodal distributions, constructing HPD intervals is not straightforward. The basic idea is to move a horizontal line down across the density, stopping when the posterior probability of the θ values in the region reaches $1-\alpha$.

DEFINITION 12.3. The *posterior predictive distribution* for a future outcome $\tilde{y}$ given the data $\mathbf{y}$ is

$$p(\tilde{y}|\mathbf{y}) = \int_\theta p(\tilde{y}, \theta|\mathbf{y})\,\mathrm{d}\,\theta = \int_\theta p(\tilde{y}|\theta, \mathbf{y})\,p(\theta|\mathbf{y})\,\mathrm{d}\,\theta = \int_\theta p(\tilde{y}|\theta)\,p(\theta|\mathbf{y})\,\mathrm{d}\,\theta.$$

The posterior predictive distribution is an expected value over the posterior distribution, i.e., $p(\tilde{y}|\mathbf{y}) = \mathrm{E}_{\theta|y}[p(\tilde{y}|\theta)]$. Observing $y_1, \ldots, y_n$ gives information about θ, which in turn gives information about $\tilde{y}$ (which is independent of the observed values $\mathbf{y}$).

EXAMPLE 12.4 (Predictive distribution for binary data). Let $y_1, \ldots, y_n$ be the outcomes from a sample of n binary random variables, and let $\tilde{y} \in \{0, 1\}$ be an additional outcome from the same population that has yet to be observed. The predictive distribution for conditionally iid binary variables can be derived as

$$p(\tilde{y} = 1|\mathbf{y}) = \int_\theta p(\tilde{y} = 1|\theta)\,p(\theta|\mathbf{y})\,\mathrm{d}\,\theta = \int_\theta \theta \cdot p(\theta|\mathbf{y})\,\mathrm{d}\,\theta,$$

where the final equality follows from the fact that $\tilde{Y} \sim \text{Bernoulli}(\theta)$, so that $P\left(\left\{\tilde{Y} = 1\right\}\right) = \theta$. Suppose that the posterior distribution of θ is as in example 11.4, i.e.,

$$\theta|\mathbf{y} \sim \text{Beta}\left(a + \sum_{i=1}^n y_i, b + n - \sum_{i=1}^n y_i\right).$$

Then,

$$P\left(\tilde{Y}=1|\mathbf{y}\right)=\mathrm{E}\left[\theta|\mathbf{y}\right]=\frac{a+\sum_{i=1}^{n}y_i}{a+b+n}\quad\text{and}\quad P\left(\tilde{Y}=0|\mathbf{y}\right)=1-\mathrm{E}\left[\theta|\mathbf{y}\right]=\frac{b+n-\sum_{i=1}^{n}y_i}{a+b+n}.$$

12.1. Conjugate prior distributions

Sampling models from exponential families all have conjugate priors. Recall that a one-parameter *exponential family model* is any model with density that can be expressed as

$$p\left(y|\theta\right)=h\left(y\right)g\left(\theta\right)\exp\left\{\phi\left(\theta\right)t\left(y\right)\right\},$$

where $\phi\left(\theta\right)$ is called the *natural parameter* and $t\left(y\right)$ is a *sufficient statistic* for θ. The conjugate prior has the form

$$p\left(\theta\right)\propto g\left(\theta\right)^{n_0}\exp\left\{\phi\left(\theta\right)\nu\right\},$$

where ν represents the prior expected value of $t\left(Y\right)$ and $n_0/\left(n_0+n\right)$ represents how informative the prior is relative to the data. With observed iid data $y_1,\ldots,y_n$, the likelihood for θ is

$$\mathcal{L}\left(\theta\right)=\prod_{i=1}^{n}p\left(y_i|\theta\right)=\left(\prod_{i=1}^{n}h\left(y_i\right)\right)g\left(\theta\right)^{n}\exp\left\{\phi\left(\theta\right)\sum_{i=1}^{n}t\left(y_i\right)\right\}.$$

The posterior distribution then becomes

$$p\left(\theta|y\right)\propto g\left(\theta\right)^{n_0+n}\exp\left\{\phi\left(\theta\right)\left(\nu+\sum_{i=1}^{n}t\left(y_i\right)\right)\right\}.$$

Example 12.5 (Binomial conjugate prior). Suppose $Y\sim\text{Binomial}\left(n,\theta\right)$. Then, example 7.2 implies that

$$p\left(y|\theta\right)=\binom{n}{y}\left(1-\theta\right)^{n}\exp\left\{y\log\frac{\theta}{1-\theta}\right\},$$

so that $g\left(\theta\right)=1-\theta$, $\phi\left(\theta\right)=\log\left(\theta/\left(1-\theta\right)\right)$ is the natural parameter, and $t\left(y\right)=y$ is a sufficient statistic. Then, the conjugate prior for θ is

$$\begin{aligned}p\left(\theta\right)&\propto g\left(\theta\right)^{n_0}\exp\left\{\phi\left(\theta\right)\nu\right\}\\&=\left(1-\theta\right)^{n_0}\exp\left\{\nu\log\frac{\theta}{1-\theta}\right\}\\&=\left(1-\theta\right)^{n_0}\left[\exp\left\{\log\frac{\theta}{1-\theta}\right\}\right]^{\nu}\\&=\left(1-\theta\right)^{n_0}\left(\frac{\theta}{1-\theta}\right)^{\nu}\\&=\theta^{\nu}\left(1-\theta\right)^{n_0}\left(1-\theta\right)^{-\nu}\\&=\theta^{\nu}\left(1-\theta\right)^{n_0-\nu},\end{aligned}$$

which we recognize as the kernel of a $\text{Beta}\left(\nu+1,n_0-\nu+1\right)$ distribution, i.e., the conjugate prior for θ is a beta distribution.

Example 12.6 (Poisson conjugate prior). Suppose $X_1,\ldots,X_n$ is an iid sample from $\text{Poisson}\left(\lambda\right)$. Then, example 8.12 implies that the joint pmf of the X_i's can be written as

$$\begin{aligned}p\left(\mathbf{x}|\theta\right)&=\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)\mathrm{e}^{-\theta n}\theta^{\sum_{i=1}^{n}x_i}\\&=\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)\mathrm{e}^{-\theta n}\exp\left\{\log\theta^{\sum_{i=1}^{n}x_i}\right\}\\&=\left(\prod_{i=1}^{n}\frac{1}{x_i!}\right)\mathrm{e}^{-\theta n}\exp\left\{\left(\sum_{i=1}^{n}x_i\right)\log\theta\right\}.\end{aligned}$$

Then, we have $g(\theta) = \mathrm{e}^{-\theta}$ and $\phi(\theta) = \log\theta$, so that the conjugate prior for θ is

$$p(\theta) \propto \mathrm{e}^{-\theta n_0} \exp\{\nu \log\theta\} = \mathrm{e}^{-\theta n_0} (\exp\{\log\theta\})^{\nu} = \mathrm{e}^{-\theta n_0}\theta^{\nu},$$

which we recognize as the kernel of a Gamma$(\nu+1, n_0)$ random variable. Taking $\theta \sim$ Gamma(α, β), the posterior distribution of θ is

$$p(\theta|\mathbf{x}) = \frac{p(\mathbf{x}|\theta)\,p(\theta)}{\int_\Theta p(\mathbf{x}|\theta)\,p(\theta)\,\mathrm{d}\,\theta} \propto \left[\mathrm{e}^{-\theta n}\theta^{\sum_{i=1}^n x_i}\right]\left[\theta^{\alpha-1}\mathrm{e}^{-\beta\theta}\right] = \theta^{\alpha+\sum_{i=1}^n x_i - 1}\mathrm{e}^{-\theta(\beta+n)},$$

which we recognize as the kernel of a Gamma$(\alpha + \sum_{i=1}^n x_i, \beta + n)$ random variable. The prior mean is α/β, and example 9.13 implies that the maximum likelihood estimate is $\hat{\theta} = \bar{x}$. The posterior mean is then

$$(\theta|\mathbf{x} \sim \text{Gamma}(\alpha + \textstyle\sum_{i=1}^n x_i, \beta + n)) \qquad \begin{aligned} \mathrm{E}[\theta|\mathbf{x}] &= \frac{\alpha + \sum_{i=1}^n x_i}{\beta + n} \\ &= \frac{\alpha}{\beta+n} + \frac{\sum_{i=1}^n x_i}{\beta+n} \\ &= \frac{\beta}{\beta+n}\frac{\alpha}{\beta} + \frac{n}{\beta+n}\bar{x}, \end{aligned}$$

which we see is the average of the prior mean and MLE with weights $\beta/(\beta+n)$ and $n/(\beta+n)$, respectively.

12.2. Noninformative priors

A prior $\pi(\theta)$ is noninformative (vague, flat) if it has minimal impact on the posterior distribution $f(\theta|y)$. Roughly speaking, a prior distribution is noninformative if the prior is "flat" relative to the likelihood function. For example, if $\theta \sim \mathcal{N}(\mu_0, \tau^2)$ and $\tau^2 \to \infty$, then we get a noninformative prior. That is, we can pick τ^2 large enough to obtain a noninformative prior.

EXAMPLE 12.7. Consider $X \sim$ Binomial(n, θ). We know that $\theta \in [0,1]$, and the flat prior on θ is the uniform distribution, $\pi(\theta) = 1$. Consider the parameterization using the log-odds,

$$\rho = \log\frac{\theta}{1-\theta} \implies \mathrm{e}^{\rho} = \frac{\theta}{1-\theta} \implies \theta = \mathrm{e}^{\rho} - \mathrm{e}^{\rho}\theta \implies \theta(1+\mathrm{e}^{\rho}) = \mathrm{e}^{\rho} \implies \theta = \frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}},$$

which maps θ to the real numbers. Then, theorem 3.2 implies that a pdf of ρ is

$$\pi_\rho(\rho) = \pi_\theta\left(\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)\left|\frac{\mathrm{d}}{\mathrm{d}\,\rho}\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right| = 1\left|\frac{\mathrm{e}^{\rho}(1+\mathrm{e}^{\rho}) - \mathrm{e}^{\rho}(0+\mathrm{e}^{p})}{(1+\mathrm{e}^{\rho})^2}\right| = \left|\frac{\mathrm{e}^{\rho}+\mathrm{e}^{2\rho}-\mathrm{e}^{2\rho}}{(1+\mathrm{e}^{\rho})^2}\right| = \frac{\mathrm{e}^{\rho}}{(1+\mathrm{e}^{\rho})^2}.$$

Under this parameterization, the prior distribution $\pi(\rho)$ is no longer flat. This example shows a prior that is noninformative is one parameterization, but becomes informative through a change of variables.

12.3. Improper priors

DEFINITION 12.8. A prior is said to be *improper* if

$$\int_\Theta \pi(\theta)\,\mathrm{d}\,\theta = \infty.$$

If a prior integrates to any finite constant, then it may be normalized to yield a proper prior. Improper priors yield noninformative priors. The posterior obtained from an improper prior may be either proper or improper (though inference cannot be made with improper posterior distributions).

EXAMPLE 12.9 (Normal mean, known variance). Let $Y_1, \ldots, Y_n$ be iid samples from a $\mathcal{N}(\theta, 1^2)$ distribution, and suppose $\pi(\theta) \propto 1$ for $\theta \in \mathbb{R}$. We see that

$$\begin{aligned} \int_\Theta \pi(\theta)\,\mathrm{d}\,\theta &= \int_{-\infty}^{\infty} 1\,\mathrm{d}\,\theta \\ &= \lim_{c\to\infty}\lim_{k\to-\infty}\int_k^c 1\,\mathrm{d}\,\theta \\ &= \lim_{c\to\infty}\lim_{k\to-\infty}\theta|_k^c \end{aligned}$$

$$
\begin{aligned}
&= \lim_{c\to\infty} \lim_{k\to-\infty} (c-k) \\
&= \lim_{c\to\infty} (c-(-\infty)) \\
&= \infty + \infty \\
&= \infty,
\end{aligned}
$$

i.e., the prior distribution is improper. Example 8.14 implies that the joint density of the samples is given by

$$
p(\mathbf{y}|\theta) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\theta\sum_{i=1}^n y_i + n\theta^2\right)\right\},
$$

so that the posterior distribution is

$$
\begin{aligned}
p(\theta|\mathbf{y}) &\propto \frac{\left[\left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\theta\sum_{i=1}^n y_i + n\theta^2\right)\right\}\right]\cdot 1}{\int_\Theta \left[\left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\theta\sum_{i=1}^n y_i + n\theta^2\right)\right\}\right]\cdot 1\,\mathrm{d}\,\theta} \\
&= \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\theta\sum_{i=1}^n y_i + n\theta^2\right)\right\}}{\int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\theta\sum_{i=1}^n y_i + n\theta^2\right)\right\}\mathrm{d}\,\theta}.
\end{aligned}
$$

The denominator becomes

$$
\begin{aligned}
&\int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\theta\sum_{i=1}^n y_i + n\theta^2\right)\right\}\mathrm{d}\,\theta \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 - \frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta n\bar{y}\right)\right\}\mathrm{d}\,\theta \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 - \frac{n}{2\sigma^2}\left(\theta^2 - 2\theta\bar{y}\right)\right\}\mathrm{d}\,\theta \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 - \frac{n}{2\sigma^2}\left(\theta^2 - 2\theta\bar{y} + \bar{y}^2 - \bar{y}^2\right)\right\}\mathrm{d}\,\theta \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 - \frac{n}{2\sigma^2}\left(-\bar{y}^2\right) - \frac{n}{2\sigma^2}\left(\theta^2 - 2\theta\bar{y} + \bar{y}^2\right)\right\}\mathrm{d}\,\theta \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right) - \frac{1}{2\sigma^2/n}\left(\theta-\bar{y}\right)^2\right\}\mathrm{d}\,\theta \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)\right\} \exp\left\{-\frac{\left(\theta-\bar{y}\right)^2}{2\sigma^2/n}\right\}\mathrm{d}\,\theta \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)\right\} \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{\left(\theta-\bar{y}\right)^2}{2\sigma^2/n}\right\}\mathrm{d}\,\theta \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)\right\} n^{-n/2} \int_{-\infty}^{\infty} \frac{\left(2\pi\sigma^2\right)^{-n/2}}{n^{-n/2}} \exp\left\{-\frac{\left(\theta-\bar{y}\right)^2}{2\sigma^2/n}\right\}\mathrm{d}\,\theta \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)\right\} n^{-n/2} \int_{-\infty}^{\infty} \left(2\pi\sigma^2/n\right)^{-n/2} \exp\left\{-\frac{\left(\theta-\bar{y}\right)^2}{2\sigma^2/n}\right\}\mathrm{d}\,\theta.
\end{aligned}
$$

We recognize the integrand as the pdf of a $\mathcal{N}\left(\bar{y}, \sigma^2/n\right)$ random variable, and it follows that

$$
\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - n\bar{y}^2\right)\right\} n^{-n/2} \int_{-\infty}^{\infty} \left(2\pi\sigma^2/n\right)^{-n/2} \exp\left\{-\frac{\left(\theta-\bar{y}\right)^2}{2\sigma^2/n}\right\}\mathrm{d}\,\theta
$$

$$= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)\right\} n^{-n/2} \cdot 1$$
$$= n^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)\right\}.$$

Then, the posterior distribution becomes

$$\begin{aligned}
p(\theta|\mathbf{y}) &\propto \frac{\left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2\right)\right\}}{n^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)\right\}} \\
&= \left(\frac{2\pi\sigma^2}{n}\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2\right)\right\} \exp\left\{-\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)\right]\right\} \\
&= \left(\frac{2\pi\sigma^2}{n}\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\theta\sum_{i=1}^{n} y_i + n\theta^2\right) + \frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right)\right\} \\
&= \left(\frac{2\pi\sigma^2}{n}\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 - \frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta n\bar{y}\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{n} y_i^2 - \frac{1}{2\sigma^2} n\bar{y}^2\right\} \\
&= \left(\frac{2\pi\sigma^2}{n}\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(n\theta^2 - 2\theta n\bar{y} + n\bar{y}^2\right)\right\} \\
&= \left(\frac{2\pi\sigma^2}{n}\right)^{-n/2} \exp\left\{-\frac{(\theta-\bar{y})^2}{2\sigma^2/n}\right\},
\end{aligned}$$

i.e., the posterior distribution of θ is $\theta|\mathbf{y} \sim \mathcal{N}\left(\bar{y}, \sigma^2/n\right)$, which is proper. We have $\sigma^2 = 1$, so that $\theta|\mathbf{y} \sim \mathcal{N}(\bar{y}, 1/n)$.

12.4. Jeffreys prior

Jeffreys (1961) suggested a default procedure for specifying a prior distribution for θ that is invariant under transformation,

$$\pi(\theta) \propto [I(\theta)]^{1/2},$$

where $I(\theta)$ is the expected Fisher information as given in theorem 9.34 (the expectation is taken with respect to the sampling distribution of Y given θ). Jeffreys prior gives an automated scheme for finding a noninformative prior for any parametric model $f(y|\theta)$. Jeffreys prior is improper for many models, although it may be proper for certain models. Jeffreys prior generally works well for one-parameter models, though it is not as straightforward for multiparameter models.

Jeffreys general principle is that any rule for determining the prior density $\pi(\theta)$ should yield an equivalent result if applied to the transformed parameter. Consider a one-to-one transformation of the parameter $\phi = h(\theta)$. Then, theorem 3.2 implies that a pdf of the transformed parameter is

$$\pi_\Phi(\phi) = \pi_\Theta\left(h^{-1}(\phi)\right)\left|\frac{\mathrm{d}}{\mathrm{d}\phi} h^{-1}(\phi)\right| = \pi_\Theta(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|,$$

where the final equality holds for the Jeffreys prior. Using $I(\phi)$, we have

$$I(\phi) = -\mathrm{E}\left[\frac{\partial^2}{\partial\phi^2}\log p(y|\phi)\right] = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log p\left(y|\theta = h^{-1}(\phi)\right)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|^2\right] = I(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|^2,$$

so that

$$\pi_\Phi(\phi) = [I(\phi)]^{1/2} = \left[I(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|^2\right]^{1/2} = [I(\theta)]^{1/2}\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right| = \pi_\Theta(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right|,$$

as obtained with the change of variable formula.

Example 12.10 (Jeffreys prior for normal mean). Suppose $Y \sim \mathcal{N}(\mu, 1)$.

(1) Find Jeffreys prior for μ.

The log-likelihood is

$$\ell(\theta|\mathbf{y}) = \log\left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2\right\} = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\left(\frac{y-\theta}{\sigma}\right)^2.$$

Taking the derivative with respect to θ gives

$$\frac{\partial}{\partial\theta}\ell(\theta|\mathbf{y}) = 0 - \frac{1}{2}\cdot 2\left(\frac{y-\theta}{\sigma}\right)\left(-\frac{1}{\sigma}\right) = \frac{y-\theta}{\sigma^2}.$$

The second derivative with respect to θ is then

$$(\sigma^2 = 1) \qquad \frac{\partial^2}{\partial\theta^2}\ell(\theta|\mathbf{y}) = \frac{\partial}{\partial\theta}\frac{y-\theta}{\sigma^2} = 0 - \frac{1}{\sigma^2} = -1.$$

Then, the Fisher information is

$$I(\theta) = -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta|\mathbf{y})\right] = -\operatorname{E}[-1] = 1,$$

so that the Jeffreys prior for μ is $\pi(\theta) \propto \sqrt{I(\theta)} = \sqrt{1} = 1$.

(2) Let $\psi = \exp(\mu)$. Find Jeffreys prior for ψ.

We have $\psi = \mathrm{e}^{\mu} \implies \mu = \log\psi$. Then, a pdf for ψ is

$$\pi_\Psi(\psi) = \pi_\mu(\log\psi)\left|\frac{\mathrm{d}}{\mathrm{d}\,\psi}\log\psi\right| \propto 1\cdot\left|\frac{1}{\psi}\right| = \frac{1}{\psi}.$$

Noting that $\mu \in \mathbb{R} \implies \psi \in (0,\infty)$, we have

$$\begin{aligned}
\int_0^\infty \frac{1}{\psi}\,\mathrm{d}\,\psi &= \int_0^x \frac{1}{\psi}\,\mathrm{d}\,\psi + \int_x^\infty \frac{1}{\psi}\,\mathrm{d}\,\psi \\
&= \lim_{c\to 0^+}\int_c^x \frac{1}{\psi}\,\mathrm{d}\,\psi + \lim_{c\to\infty}\int_x^c \frac{1}{\psi}\,\mathrm{d}\,\psi \\
&= \lim_{c\to 0^+}\left[\log\psi\right|_c^x + \lim_{c\to\infty}\left[\log\psi\right|_x^c \qquad (\psi > 0 \implies |\psi| = \psi) \\
&= \lim_{c\to 0^+}(\log x - \log c) + \lim_{c\to\infty}(\log c - \log x) \\
&= \log x - \lim_{c\to 0^+}\log c + \lim_{c\to\infty}\log c - \log x \\
&= -(-\infty) + \infty \\
&= \infty,
\end{aligned}$$

i.e., $\pi_\Psi(\psi)$ is improper. Now, the density of Y (parameterized in terms of ψ) is

$$p(y|\psi) = \left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{y-\log\psi}{\sigma}\right)^2\right\},$$

so that the log-likelihood is

$$\ell(\psi|\mathbf{y}) = \log\left(2\pi\sigma^2\right)^{-1/2}\exp\left\{-\frac{1}{2}\left(\frac{y-\log\psi}{\sigma}\right)^2\right\} = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\left(\frac{y-\log\psi}{\sigma}\right)^2.$$

Taking the second derivative with respect to ψ gives

$$\begin{aligned}
\frac{\partial^2}{\partial\psi^2}\ell(\psi|\mathbf{y}) &= \frac{\partial}{\partial\psi}\left[\frac{\partial}{\partial\psi} - \frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2}\left(\frac{y-\log\psi}{\sigma}\right)^2\right] \\
&= \frac{\partial}{\partial\psi}\left[0 - \frac{1}{2}\cdot 2\left(\frac{y-\log\psi}{\sigma}\right)\left(-\frac{1}{\sigma\psi}\right)\right] \\
&= \frac{\partial}{\partial\psi}\frac{y-\log\psi}{\sigma^2\psi} \\
&= \frac{(0-1/\psi)\left(\sigma^2\psi\right) - (y-\log\psi)\left(\sigma^2\right)}{\left(\sigma^2\psi\right)^2}
\end{aligned}$$

$$= \frac{-\sigma^2 - \sigma^2 (y - \log\psi)}{(\sigma^2\psi)^2}$$

$$(\sigma^2 = 1) \qquad = \frac{-1 - (y - \log\psi)}{\psi^2}.$$

Then, the Fisher information is

$$\begin{aligned} I(\psi) &= -\operatorname{E}\left[\frac{\partial^2}{\partial\psi^2}\log p(Y|\psi)\right] \\ &= -\operatorname{E}\left[-\frac{1}{\psi^2} - \frac{1}{\psi^2}(Y - \log\psi)\right] \\ &= -\left(\operatorname{E}\left[-\frac{1}{\psi^2}\right] - \operatorname{E}\left[\frac{1}{\psi^2}(Y - \log\psi)\right]\right) \\ &= -\left(-\frac{1}{\psi^2} - \frac{1}{\psi^2}\operatorname{E}[Y - \log\psi]\right) \\ &= \frac{1}{\psi^2} + \frac{1}{\psi^2}(\operatorname{E}[Y] - \operatorname{E}[\log\psi]) \\ &= \frac{1}{\psi^2}(1 + \operatorname{E}[Y] - \log\psi) \\ (\operatorname{E}[Y] = \log\psi) \qquad &= \frac{1}{\psi^2}(1 + \log\psi - \log\psi) \\ &= \frac{1}{\psi^2}(1 + 0) \\ &= \frac{1}{\psi^2}, \end{aligned}$$

so that Jeffreys prior for ψ is

$$\pi(\psi) \propto \sqrt{I(\psi)} = \sqrt{\frac{1}{\psi^2}} = \frac{1}{\psi},$$

which agrees with the result obtained from the change of variable formula.

EXAMPLE 12.11 (Binomial Jeffreys prior). Suppose $y_1, \ldots, y_n$ are iid Binomial $(1, \theta)$.

(1) Compute Jeffreys prior for θ.
We have $Y_i \sim$ Bernoulli (θ), so that

$$\log p(y|\theta) = \log \theta^y (1-\theta)^{1-y} = y\log\theta + (1-y)\log(1-\theta).$$

Taking the derivative with respect to θ gives

$$\frac{\partial}{\partial\theta} p(y|\theta) = \frac{\partial}{\partial\theta}[y\log\theta + (1-y)\log(1-\theta)] = \frac{y}{\theta} + \frac{1-y}{1-\theta}(-1) = \frac{y}{\theta} - \frac{1-y}{1-\theta}.$$

Then, the second derivative with respect to θ is

$$\frac{\partial^2}{\partial\theta^2} p(y|\theta) = \frac{\partial}{\partial\theta}\left[\frac{y}{\theta} - \frac{1-y}{1-\theta}\right] = -\frac{y}{\theta^2} - \frac{1-y}{(1-\theta)^2}(-1)(-1) = -\frac{y}{\theta^2} - \frac{1-y}{(1-\theta)^2},$$

so that the Fisher information is

$$\begin{aligned} I(\theta) &= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(Y|\theta)\right] \\ &= -\operatorname{E}\left[-\frac{Y}{\theta^2} - \frac{1-Y}{(1-\theta)^2}\right] \\ &= -\left(\operatorname{E}\left[-\frac{Y}{\theta^2}\right] - \operatorname{E}\left[\frac{1-Y}{(1-\theta)^2}\right]\right) \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{\theta^2}\,\mathrm{E}\left[Y\right] - \frac{1}{\left(1-\theta\right)^2}\left(\mathrm{E}\left[1\right] - \mathrm{E}\left[Y\right]\right) \\
(\mathrm{E}\left[Y\right] = \theta) \qquad &= \frac{1}{\theta^2}\theta - \frac{1-\theta}{\left(1-\theta\right)^2} \\
&= \frac{1}{\theta} - \frac{1}{1-\theta} \\
&= \frac{1-\theta-\theta}{\theta\left(1-\theta\right)} \\
&= \frac{1}{\theta\left(1-\theta\right)}.
\end{aligned}
$$

Then, the Jeffreys prior for θ is

$$\pi\left(\theta\right) \propto \sqrt{I\left(\theta\right)} = \sqrt{\frac{1}{\theta\left(1-\theta\right)}} = \theta^{-1/2}\left(1-\theta\right)^{-1/2} = \theta^{1/2-1}\left(1-\theta\right)^{1/2-1},$$

which we recognize as the kernel of a $\mathrm{Beta}\left(1/2, 1/2\right)$ random variable, i.e., $\theta \sim \mathrm{Beta}\left(1/2, 1/2\right)$, which is proper.

(2) Show that Jeffreys prior is invariant to the transformation $\rho = \log\left(\theta/\left(1-\theta\right)\right)$.

From example 12.7, we have

$$\theta = \frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}.$$

Thus, to show that Jeffreys prior is invariant to this transformation, we must show that

$$\left|I\left(\rho\right)\right|^{1/2} = \left|I\left(\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)\right|^{1/2}\left|\frac{\mathrm{d}}{\mathrm{d}\,\rho}\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right|.$$

The density of Y (parameterized in terms of ρ) is

$$p\left(y|\rho\right) = \left(\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)^{y}\left(1-\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)^{1-y}.$$

Then,

$$
\begin{aligned}
\frac{\partial^2}{\partial\rho^2}\log p\left(y|\rho\right) &= \frac{\partial^2}{\partial\rho^2}\log\left(\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)^{y}\left(1-\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)^{1-y} \\
&= \frac{\partial^2}{\partial\rho^2}\left[y\log\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}} + \left(1-y\right)\log\left(\frac{1+\mathrm{e}^{\rho}-\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)\right] \\
&= \frac{\partial^2}{\partial\rho^2}\left[y\left(\log\mathrm{e}^{\rho} - \log\left(1+\mathrm{e}^{\rho}\right)\right) + \left(1-y\right)\left(\log\left(1\right) - \log\left(1+\mathrm{e}^{\rho}\right)\right)\right] \\
&= \frac{\partial^2}{\partial\rho^2}\left[y\rho - y\log\left(1+\mathrm{e}^{\rho}\right) + \left(1-y\right)\left(0 - \log\left(1+\mathrm{e}^{\rho}\right)\right)\right] \\
&= \frac{\partial^2}{\partial\rho^2}\left[y\rho - y\log\left(1+\mathrm{e}^{\rho}\right) - \log\left(1+\mathrm{e}^{\rho}\right) + y\log\left(1+\mathrm{e}^{\rho}\right)\right] \\
&= \frac{\partial^2}{\partial\rho^2}\left[y\rho - \log\left(1+\mathrm{e}^{\rho}\right)\right] \\
&= \frac{\partial}{\partial\rho}\left[y - \frac{1}{1+\mathrm{e}^{\rho}}\left(\mathrm{e}^{\rho}\right)\right] \\
&= 0 - \frac{\mathrm{e}^{\rho}\left(1+\mathrm{e}^{\rho}\right) - \mathrm{e}^{\rho}\left(\mathrm{e}^{\rho}\right)}{\left(1+\mathrm{e}^{\rho}\right)^2} \\
&= -\frac{\mathrm{e}^{\rho} + \mathrm{e}^{2\rho} - \mathrm{e}^{2\rho}}{\left(1+\mathrm{e}^{\rho}\right)^2} \\
&= -\frac{\mathrm{e}^{\rho}}{\left(1+\mathrm{e}^{\rho}\right)^2},
\end{aligned}
$$

so that the Fisher information is

$$I(\rho) = -\operatorname{E}\left[\frac{\partial^2}{\partial\rho^2}\log p(Y|\rho)\right] = -\operatorname{E}\left[-\frac{\mathrm{e}^{\rho}}{(1+\mathrm{e}^{\rho})^2}\right] = \frac{\mathrm{e}^{\rho}}{(1+\mathrm{e}^{\rho})^2}.$$

Then, the Jeffreys prior is

$$\pi_{\rho}(\rho) \propto \sqrt{I(\rho)} = \sqrt{\frac{\mathrm{e}^{\rho}}{(1+\mathrm{e}^{\rho})^2}}. \tag{12.4.1}$$

Now,

$$\begin{aligned}
\left|I\left(\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)\right|^{1/2}\left|\frac{\partial}{\partial\rho}\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right| &= \left[\left(\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)^{-1/2}\left(1-\frac{\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)^{-1/2}\right]\left|\frac{\mathrm{e}^{\rho}(1+\mathrm{e}^{\rho})-\mathrm{e}^{\rho}(\mathrm{e}^{\rho})}{(1+\mathrm{e}^{\rho})^2}\right| \\
&= \frac{(1+\mathrm{e}^{\rho})^{1/2}}{\mathrm{e}^{\rho/2}}\left(\frac{1+\mathrm{e}^{\rho}-\mathrm{e}^{\rho}}{1+\mathrm{e}^{\rho}}\right)^{-1/2}\left|\frac{\mathrm{e}^{\rho}+\mathrm{e}^{2\rho}-\mathrm{e}^{2\rho}}{(1+\mathrm{e}^{\rho})^2}\right| \\
&= \frac{(1+\mathrm{e}^{\rho})^{1/2}}{\mathrm{e}^{\rho/2}}\frac{(1+\mathrm{e}^{\rho})^{1/2}}{1}\left|\frac{\mathrm{e}^{\rho}}{(1+\mathrm{e}^{\rho})^2}\right| \\
&= \frac{(1+\mathrm{e}^{\rho})}{\mathrm{e}^{\rho/2}}\left(\frac{\mathrm{e}^{\rho}}{(1+\mathrm{e}^{\rho})^2}\right) \qquad (\mathrm{e}^{\rho}>0) \\
&= \frac{\mathrm{e}^{\rho}\mathrm{e}^{-\rho/2}}{1+\mathrm{e}^{\rho}} \\
&= \frac{\mathrm{e}^{\rho/2}}{1+\mathrm{e}^{\rho}} \\
&= \frac{\sqrt{\mathrm{e}^{\rho}}}{\sqrt{(1+\mathrm{e}^{\rho})^2}} \\
&= \sqrt{\frac{\mathrm{e}^{\rho}}{(1+\mathrm{e}^{\rho})^2}},
\end{aligned}$$

which agrees with (12.4.1), and the result has been shown.

12.4.1. Likelihood principle and experimental design. Experimental design refers to the method used to collect the data. The likelihood principle refers to the concept that all the information carried in a sample is contained in the likelihood function. Jeffreys prior accounts for the experimental design, thereby violating the likelihood principle. The following examples illustrate this view.

EXAMPLE 12.12. Consider the scenario in which we toss a coin n times and observe r heads. Let $X \sim \text{Binomial}(n,\theta)$, so that the likelihood function is

$$p(X=r|\theta) = \mathcal{L}(\theta) = \binom{n}{r}\theta^r(1-\theta)^{n-r}.$$

In example 12.11, we showed that the Jeffreys prior is given by

$$\pi(\theta) \propto \theta^{-1/2}(1-\theta)^{-1/2},$$

and the corresponding posterior is

$$\theta|X=r \sim \text{Beta}\left(r+\frac{1}{2}, n-r+\frac{1}{2}\right).$$

Now consider the scenario in which we toss a coin until we see r heads and end up tossing it $n = y+r$ times in total. Then, $Y \sim \mathcal{NB}(r,\theta)$, with likelihood function

$$p(Y=y|\theta) = \mathcal{L}(\theta) = \binom{y+r-1}{r-1}\theta^r(1-\theta)^y = \binom{n-1}{r-1}\theta^r(1-\theta)^{n-r}.$$

The difference between the two scenarios is captured in the constant terms $\binom{n-1}{r-1}$ versus $\binom{n}{r}$, which are considered to be exclusively part of the experimental design since they do not affect the shape of the likelihood. In the negative binomial case, the second derivative of the log-likelihood with respect to θ is

$$\begin{aligned}
\frac{\partial^2}{\partial\theta^2}\ell(\theta) &= \frac{\partial^2}{\partial\theta^2}\log\binom{n-1}{r-1}\theta^r(1-\theta)^{n-r} \\
&= \frac{\partial^2}{\partial\theta^2}\left[\log\binom{n-1}{r-1} + r\log\theta + (n-r)\log(1-\theta)\right] \\
&= \frac{\partial}{\partial\theta}\left[0 + \frac{r}{\theta} + \frac{n-r}{1-\theta}(-1)\right] \\
&= -\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2}(-1)(-1) \\
&= -\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2} \\
(y = n-r) \qquad &= -\frac{r}{\theta^2} - \frac{y}{(1-\theta)^2},
\end{aligned}$$

so that the Fisher information is

$$\begin{aligned}
I(\theta) &= -\operatorname{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(Y|\theta)\right] \\
&= -\operatorname{E}\left[-\frac{r}{\theta^2} - \frac{Y}{(1-\theta)^2}\right] \\
&= \frac{r}{\theta^2} + \frac{1}{(1-\theta)^2}\operatorname{E}[Y] \\
&= \frac{r}{\theta^2} + \frac{1}{(1-\theta)^2}\frac{r(1-\theta)}{\theta} \\
&= \frac{r}{\theta^2} + \frac{r}{\theta(1-\theta)} \\
&= \frac{r(1-\theta)+\theta r}{\theta^2(1-\theta)} \\
&= \frac{r}{\theta^2(1-\theta)}.
\end{aligned}$$

Then, the Jeffreys prior is

$$\pi(\theta) \propto \sqrt{I(\theta)} = \sqrt{\frac{r}{\theta^2(1-\theta)}} = \frac{\sqrt{r}}{\sqrt{\theta^2(1-\theta)}} \propto \frac{1}{\theta(1-\theta)^{1/2}} = \theta^{-1}(1-\theta)^{-1/2}.$$

While this resembles the kernel of a beta random variable, we cannot have the shape parameter $\alpha = 0$, and it follows that this is an improper prior. The corresponding posterior is

$$\theta|X = r \sim \operatorname{Beta}\left(r, n-r+\frac{1}{2}\right).$$

In the negative binomial case $r \neq 0$ (one cannot toss a coin until zero heads are observed), whereas in the binomial case r can be zero. Jeffreys prior differs in the two cases, even though the likelihood functions are proportional. We see that Jeffreys prior is affected by the experimental design and can violate the likelihood principle.

CHAPTER 13

Multiparameter models

A model may include several parameters, but often interest lies in making inference about one or a few parameters. The parameters that are not of direct interest are called *nuisance parameters*. The marginal posterior distribution of the parameters of interest can be obtained by integrating the joint posterior density over the parameters that are not of immediate interest.

13.1. Joint and marginal posterior distributions

Suppose $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and we are only interested in inference for θ_1, so θ_2 may be considered a "nuisance" parameter. The joint posterior density is given by

$$p(\theta_1, \theta_2|y) \propto p(y|\theta_1, \theta_2)\, p(\theta_1, \theta_2).$$

The marginal posterior density of θ_1 is derived by averaging over θ_2, i.e.,

$$p(\theta_1|y) = \int_{\Theta_2} p(\theta_1, \theta_2|y)\, \mathrm{d}\,\theta_2.$$

Alternatively, the marginal posterior density of θ_1 can by obtained from

$$p(\theta_1|y) = \int_{\Theta_2} p(\theta_1|\theta_2, y)\, p(\theta_2|y)\, \mathrm{d}\,\theta_2.$$

This integral is not often evaluated explicitly, but it suggests an important practical strategy:

(1) first draw θ_2 from its posterior distribution, $p(\theta_2|y)$,
(2) then draw θ_1 from its conditional posterior distribution, $p(\theta_1|\theta_2, y)$, given the drawn value of θ_2.

EXAMPLE 13.1 (Normal data with noninformative prior). Suppose $y_1, \ldots, y_n$ are iid $\mathcal{N}(\mu, \sigma^2)$ where (μ, σ^2) are both unknown. We begin by computing Jeffreys prior for (μ, σ^2). In the multiparameter case, the Fisher information is a square matrix whose entries are the negative expectations of the second partial derivatives of the log-likelihood, i.e.,

$$\mathbf{I}(\boldsymbol{\theta}) = -\,\mathrm{E}\begin{bmatrix} \dfrac{\partial^2}{\partial\mu^2}\ell(\boldsymbol{\theta}) & \dfrac{\partial^2}{\partial\mu\,\partial(\sigma^2)}\ell(\boldsymbol{\theta}) \\ \dfrac{\partial^2}{\partial(\sigma^2)\,\partial\mu}\ell(\boldsymbol{\theta}) & \dfrac{\partial^2}{\partial(\sigma^2)^2}\ell(\boldsymbol{\theta}) \end{bmatrix},$$

and Jeffreys prior is $\pi(\boldsymbol{\theta}) \propto [\det \mathbf{I}(\boldsymbol{\theta})]^{1/2}$. For $n = 1$ observation, the log-likelihood is

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}(y-\mu)^2,$$

so that

$$\frac{\partial}{\partial\mu}\ell(\boldsymbol{\theta}) = 0 - \frac{1}{2\sigma^2}\cdot 2(y-\mu)(-1) = \frac{y-\mu}{\sigma^2}$$

$$\frac{\partial}{\partial(\sigma^2)}\ell(\boldsymbol{\theta}) = -\frac{1}{2}\frac{1}{2\pi\sigma^2}(2\pi) - \frac{1}{2(\sigma^2)^2}(-1)(y-\mu)^2 = -\frac{1}{2\sigma^2} + \frac{(y-\mu)^2}{2(\sigma^2)^2}.$$

Then, the second partial derivatives are

$$\frac{\partial^2}{\partial\mu^2}\ell(\boldsymbol{\theta}) = 0 - \frac{1}{\sigma^2} = -\frac{1}{\sigma^2}$$

$$\frac{\partial^2}{\partial\mu\partial\left(\sigma^2\right)} = 0 + \frac{2\left(y-\mu\right)}{2\left(\sigma^2\right)^2}\left(-1\right) = -\frac{y-\mu}{\left(\sigma^2\right)^2}$$
$$\frac{\partial^2}{\partial\left(\sigma^2\right)\partial\mu}\ell\left(\boldsymbol{\theta}\right) = -\frac{y-\mu}{\left(\sigma^2\right)^2}$$
$$\frac{\partial^2}{\partial\left(\sigma^2\right)^2}\ell\left(\boldsymbol{\theta}\right) = \frac{1}{2\left(\sigma^2\right)^2} + \left(-2\right)\frac{\left(y-\mu\right)^2}{2\left(\sigma^2\right)^3} = \frac{1}{2\left(\sigma^2\right)^2} - \frac{\left(y-\mu\right)^2}{\left(\sigma^2\right)^3},$$

so that the Fisher information is

$$\begin{aligned}
\mathbf{I}\left(\boldsymbol{\theta}\right) &= -\operatorname{E}\begin{bmatrix} -1/\sigma^2 & \left(Y-\mu\right)/\left(\sigma^2\right)^2 \\ \left(Y-\mu\right)/\left(\sigma^2\right)^2 & 1/2\left(\sigma^2\right)^2 - \left(Y-\mu\right)^2/\left(\sigma^2\right)^3 \end{bmatrix} \\
&= -\begin{bmatrix} \operatorname{E}\left[-1/\sigma^2\right] & \operatorname{E}\left[\left(Y-\mu\right)/\left(\sigma^2\right)^2\right] \\ \operatorname{E}\left[\left(Y-\mu\right)/\left(\sigma^2\right)^2\right] & \operatorname{E}\left[1/2\left(\sigma^2\right)^2 - \left(Y-\mu\right)^2/\left(\sigma^2\right)^3\right] \end{bmatrix} \\
&= -\begin{bmatrix} -1/\sigma^2 & \left(1/\left(\sigma^2\right)^2\right)\operatorname{E}\left[Y-\mu\right] \\ \left(1/\left(\sigma^2\right)^2\right)\operatorname{E}\left[Y-\mu\right] & \operatorname{E}\left[1/2\left(\sigma^2\right)^2\right] - \operatorname{E}\left[\left(Y-\mu\right)^2/\left(\sigma^2\right)^3\right] \end{bmatrix} \\
&= -\begin{bmatrix} -1/\sigma^2 & \left(1/\left(\sigma^2\right)^2\right)\left(\operatorname{E}\left[Y\right] - \operatorname{E}\left[\mu\right]\right) \\ \left(1/\left(\sigma^2\right)^2\right)\left(\operatorname{E}\left[Y\right] - \operatorname{E}\left[\mu\right]\right) & 1/2\left(\sigma^2\right)^2 - \left(1/\left(\sigma^2\right)^3\right)\operatorname{E}\left[\left(Y-\mu\right)^2\right] \end{bmatrix}.
\end{aligned}$$

Now, the expected value of Y is just its mean μ, so it follows that $\operatorname{E}\left[Y\right] - \operatorname{E}\left[\mu\right] = \mu - \mu = 0$. Similarly, the expected value of $\operatorname{E}\left[\left(Y-\mu\right)^2\right]$ is just the variance of Y, i.e., σ^2. Then,

$$\begin{aligned}
\mathbf{I}\left(\boldsymbol{\theta}\right) &= -\begin{bmatrix} -1/\sigma^2 & 0 \\ 0 & 1/2\left(\sigma^2\right)^2 - \left(1/\left(\sigma^2\right)^3\right)\sigma^2 \end{bmatrix} \\
&= -\begin{bmatrix} -1/\sigma^2 & 0 \\ 0 & 1/2\left(\sigma^2\right)^2 - 1/\left(\sigma^2\right)^2 \end{bmatrix} \\
&= -\begin{bmatrix} -1/\sigma^2 & 0 \\ 0 & -1/2\left(\sigma^2\right)^2 \end{bmatrix} \\
&= \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/2\left(\sigma^2\right)^2 \end{bmatrix},
\end{aligned}$$

so that Jeffreys prior is

$$\pi\left(\boldsymbol{\theta}\right) \propto \left[\det \mathbf{I}\left(\boldsymbol{\theta}\right)\right]^{1/2} = \left[\frac{1}{\sigma^2}\left(\frac{1}{2\left(\sigma^2\right)^2}\right) - 0\right]^{1/2} = \left[\frac{1}{2\left(\sigma^2\right)^3}\right]^{1/2} = \left[2\left(\sigma^2\right)^3\right]^{-1/2} \propto \left(\sigma^2\right)^{-3/2}.$$

Now consider the prior distribution $p\left(\mu,\sigma^2\right) \propto 1/\sigma^2$. The posterior distribution of $\left(\mu,\sigma^2\right)$ is

$$\begin{aligned}
p\left(\mu,\sigma^2|\mathbf{y}\right) &\propto p\left(\mathbf{y}|\mu,\sigma^2\right)p\left(\mu,\sigma^2\right) \\
&\propto \left[\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^n\left(y_i-\mu\right)^2}{2\sigma^2}\right\}\right]\frac{1}{\sigma^2} \\
&= \left(2\pi\right)^{-n/2}\left(\sigma^2\right)^{-n/2}\left(\sigma^2\right)^{-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\mu\sum_{i=1}^n y_i + n\mu^2\right)\right\} \\
&= \left(2\pi\right)^{-n/2}\left(\sigma^2\right)^{-n/2-1}\exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^n y_i^2 - 2\mu n\bar{y} + n\mu^2\right)\right\} \\
&\propto \left(\sigma^2\right)^{-(n/2+1)}\exp\left\{-\frac{1}{2\sigma^2}\left(n\mu^2 - 2\mu n\bar{y} + n\bar{y}^2 - n\bar{y}^2 + \sum_{i=1}^n y_i^2\right)\right\}
\end{aligned}$$

$$= \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[n\left(\mu^2 - 2\mu\bar{y} + \bar{y}^2\right) - n\bar{y}^2 + \sum_{i=1}^{n} y_i^2\right]\right\}$$
$$= \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[n\left(\mu - \bar{y}\right)^2 - n\bar{y}^2 + \sum_{i=1}^{n} y_i^2\right]\right\}.$$

It follows from theorem 5.2 that

$$\sum_{i=1}^{n} y_i^2 - n\bar{y}^2 = \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2,$$

and therefore that

$$p\left(\mu, \sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[n\left(\mu - \bar{y}\right)^2 + \sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2\right]\right\}$$
$$= \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\frac{1}{n-1}\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2 + n\left(\mu - \bar{y}\right)^2\right]\right\}$$
$$= \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n\left(\mu - \bar{y}\right)^2\right]\right\},$$

where s^2 is the sample variance. The conditional posterior distribution $p\left(\mu|\sigma^2, \mathbf{y}\right)$ is equivalent to deriving the posterior for μ when σ^2 is known. Then, example 12.9 implies that

$$\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}\left(\bar{y}, \frac{\sigma^2}{n}\right).$$

The marginal posterior distribution of σ^2, $p\left(\sigma^2|\mathbf{y}\right)$, is obtained by integrating $p\left(\mu, \sigma^2|\mathbf{y}\right)$ over μ, i.e.,

$$p\left(\sigma^2|\mathbf{y}\right) = \int_{\mu} p\left(\mu, \sigma^2|\mathbf{y}\right) \mathrm{d}\,\mu$$
$$\propto \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)s^2 + n\left(\mu - \bar{y}\right)^2\right]\right\} \mathrm{d}\,\mu$$
$$= \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-[(n+2)/2]} \exp\left\{-\frac{1}{2\sigma^2}(n-1)s^2 - \frac{1}{2\sigma^2} n\left(\mu - \bar{y}\right)^2\right\} \mathrm{d}\,\mu$$
$$= \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-[1/2+(n+1)/2]} \exp\left\{-\frac{1}{2\sigma^2}(n-1)s^2\right\} \exp\left\{-\frac{1}{2\sigma^2} n\left(\mu - \bar{y}\right)^2\right\} \mathrm{d}\,\mu$$
$$= \exp\left\{-\frac{1}{2\sigma^2}(n-1)s^2\right\} \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-1/2} \left(\sigma^2\right)^{-(n+1)/2} \exp\left\{-\frac{1}{2\sigma^2} n\left(\mu - \bar{y}\right)^2\right\} \mathrm{d}\,\mu$$
$$= \left(\sigma^2\right)^{-(n+1)/2} \exp\left\{-\frac{1}{2\sigma^2}(n-1)s^2\right\} \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-1/2} \exp\left\{-\frac{\left(\mu - \bar{y}\right)^2}{2\sigma^2/n}\right\} \mathrm{d}\,\mu$$
$$= \left(\sigma^2\right)^{-(n+1)/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} \left(\frac{2\pi/n}{2\pi/n}\right)^{1/2} \int_{-\infty}^{\infty} \left(\sigma^2\right)^{-1/2} \exp\left\{-\frac{\left(\mu - \bar{y}\right)^2}{2\sigma^2/n}\right\} \mathrm{d}\,\mu$$
$$= \left(\sigma^2\right)^{-(n+1)/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} (2\pi/n)^{1/2} \int_{-\infty}^{\infty} \left(2\pi\sigma^2/n\right)^{-1/2} \exp\left\{-\frac{\left(\mu - \bar{y}\right)^2}{2\sigma^2/n}\right\} \mathrm{d}\,\mu.$$

We recognize the integrand as the density of a $\mathcal{N}\left(\bar{y}, \sigma^2/n\right)$ random variable, and it follows that

$$p\left(\sigma^2|\mathbf{y}\right) \propto \left(\sigma^2\right)^{-(n+1)/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\} (2\pi/n)^{1/2} \cdot 1$$
$$\propto \left(\sigma^2\right)^{-(n+1)/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}$$
$$= \left(\sigma^2\right)^{-(n+1+2-2)/2} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}$$

$$= \left(\sigma^2\right)^{-[(n-1)/2+1]} \exp\left\{-\frac{(n-1)\,s^2}{2\sigma^2}\right\},$$

which we recognize as the kernel of an inverse-gamma density with shape parameter $\alpha = (n-1)/2$ and scale parameter $\beta = (n-1)\,s^2/2$, i.e.,

$$\sigma^2|\mathbf{y} \sim \text{Inv-Gamma}\left(\frac{n-1}{2}, \frac{(n-1)}{2}s^2\right).$$

Note that we can also obtain the marginal posterior density of μ through an application of conditional probability, i.e.,

$$\begin{aligned}
p\left(\mu|\sigma^2, \mathbf{y}\right) &= \frac{p\left(\mu, \sigma^2|\mathbf{y}\right)}{p\left(\sigma^2|\mathbf{y}\right)} \\
&\propto \frac{\left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\left(\mu-\bar{y}\right)^2\right]\right\}}{\left(\sigma^2\right)^{-[(n-1)/2+1]} \exp\left\{-\frac{(n-1)s^2}{2\sigma^2}\right\}} \\
&= \left(\sigma^2\right)^{-(n/2+1)} \left(\sigma^2\right)^{[(n-1)/2+1]} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\left(\mu-\bar{y}\right)^2\right]\right\} \exp\left\{\frac{(n-1)\,s^2}{2\sigma^2}\right\} \\
&= \left(\sigma^2\right)^{-n/2-1+(n-1)/2+1} \exp\left\{-\frac{(n-1)\,s^2}{2\sigma^2} - \frac{n\left(\mu-\bar{y}\right)^2}{2\sigma^2} + \frac{(n-1)\,s^2}{2\sigma^2}\right\} \\
&= \left(\sigma^2\right)^{-n/2+n/2-1/2} \exp\left\{-\frac{\left(\mu-\bar{y}\right)^2}{2\sigma^2/n}\right\} \\
&= \left(\sigma^2\right)^{-1/2} \exp\left\{-\frac{\left(\mu-\bar{y}\right)^2}{2\sigma^2/n}\right\},
\end{aligned}$$

which we recognize as the kernel of a $\mathcal{N}\left(\bar{y}, \sigma^2/n\right)$ random variable.

13.1.1. Sampling from the joint posterior distribution. Consider the joint posterior density from the preceding example. One can simulate a value of $\left(\mu, \sigma^2\right)$ from the joint posterior density by

(1) simulating σ^2 from an Inv-Gamma $\left((n-1)/2, (n-1)\,s^2/2\right)$ distribution, which can be done by taking the inverse of a random sample from a Gamma $\left((n-1)/2, (n-1)\,s^2/2\right)$ distribution (recall that if $X \sim \text{Gamma}(\alpha, \beta)$, then $Y = 1/X \sim \text{Inv-Gamma}(\alpha, \beta)$);
(2) then simulating μ from a $\mathcal{N}\left(\bar{y}, \sigma^2/n\right)$ distribution.

EXAMPLE 13.2. Suppose we are interested in learning about the distribution of completion times for men between ages 20 and 29 running the New York marathon. We observe the time $y_1, \ldots, y_n$ in minutes of $n = 20$ runners, which are saved under the data `marathontimes.txt`. We assume the y_i's represent a random sample from a $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution. Inference about the parameters or functions of the parameters can be obtained from the simulated samples.

```
# code goes here
```

13.1.2. Analytical form of $p(\mu|y)$. The population mean μ is typically the parameter of interest. So, the objective of the Bayesian analysis is the marginal posterior distribution of μ. For a $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution, we have $\sigma^2 > 0$, so that

$$p\left(\mu|\mathbf{y}\right) = \int_0^\infty p\left(\mu, \sigma^2|\mathbf{y}\right) \mathrm{d}\,\sigma^2 \propto \int_0^\infty \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{2\sigma^2}\left[(n-1)\,s^2 + n\left(\mu-\bar{y}\right)^2\right]\right\} \mathrm{d}\,\sigma^2.$$

Let

$$\alpha = \frac{n}{2} \quad \text{and} \quad \beta = \frac{(n-1)\,s^2 + n\left(\mu-\bar{y}\right)^2}{2},$$

so that

$$p\left(\mu|\mathbf{y}\right) \propto \int_0^\infty \left(\sigma^2\right)^{-(\alpha+1)} \exp\left\{-\beta/\sigma^2\right\} \mathrm{d}\,\sigma^2$$

$$= \frac{\Gamma(\alpha)}{\beta^\alpha} \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)} \left(\sigma^2\right)^{-(\alpha+1)} \exp\left\{-\beta/\sigma^2\right\} \mathrm{d}\,\sigma^2.$$

We recognize the integrand as the density of an Inv-Gamma (α, β) random variable, and it follows that

$$\begin{aligned}
p(\mu|\mathbf{y}) &\propto \frac{\Gamma(\alpha)}{\beta^\alpha} \cdot 1 \\
&= \Gamma\left(\frac{n}{2}\right) \left[\frac{(n-1)s^2 + n(\mu-\bar{y})^2}{2}\right]^{-n/2} \\
&= \Gamma\left(\frac{n}{2}\right) 2^{n/2} \left[(n-1)s^2 + n(\mu-\bar{y})^2\right]^{-n/2} \\
&= \Gamma\left(\frac{n}{2}\right) 2^{n/2} \left[(n-1)s^2 \left(1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right)\right]^{-n/2} \\
&= \Gamma\left(\frac{n}{2}\right) 2^{n/2} \left[(n-1)s^2\right]^{-n/2} \left[1 + \frac{n(\mu-\bar{y})^2}{(n-1)s^2}\right]^{-n/2} \\
&= \Gamma\left(\frac{n}{2}\right) 2^{n/2} \left[(n-1)s^2\right]^{-n/2} \left[1 + \frac{(\mu-\bar{y})^2}{(n-1)s^2/n}\right]^{-n/2} \\
&= \Gamma\left(\frac{n}{2}\right) 2^{n/2} \left[(n-1)s^2\right]^{-n/2} \left[1 + \frac{1}{n-1}\left(\frac{\mu-\bar{y}}{s/\sqrt{n}}\right)^2\right]^{-n/2} \\
&\propto \left[1 + \frac{1}{n-1}\left(\frac{\mu-\bar{y}}{s/\sqrt{n}}\right)^2\right]^{-[(n-1)+1]/2},
\end{aligned}$$

which we recognize as the kernel of a t-distribution with $\nu = n-1$ degrees of freedom, location parameter $\bar{y}$, and scale parameter $(s/\sqrt{n})^2$, i.e.,

$$\mu|\mathbf{y} \sim t\left(n-1, \bar{y}, \left(s/\sqrt{n}\right)^2\right).$$

We can equivalently standardize $\mu|\mathbf{y}$, i.e., let

$$Z = \frac{\mu - \bar{y}}{s/\sqrt{n}}.$$

Then, Z has the central t-distribution with $n-1$ degrees of freedom. To generate $\mu|\mathbf{y} \sim t\left(n-1, \bar{y}, (s/\sqrt{n})^2\right)$ in R, we can

(1) draw m samples from t_{n-1} using `rt(m, df = n - 1, ncp = 0)`,
(2) then transform the samples as $\mu = \bar{y} + t_{n-1} s/\sqrt{n}$.

We implement this approach for the marathon times from example 13.2 below.

```
# code goes here
```

13.1.3. Posterior predictive distribution. As above, let $Y_1, \ldots, Y_n$ be iid samples from a $\mathcal{N}(\mu, \sigma^2)$ distribution. Then, the posterior predictive distribution for a future observation z can be written as

$$p(z|\mathbf{y}) = \int_{\sigma^2} \int_\mu p\left(z|\mu, \sigma^2\right) p\left(\mu, \sigma^2|\mathbf{y}\right) \mathrm{d}\,\mu\,\mathrm{d}\,\sigma^2,$$

or can be found more easily using the factorization of the joint density

$$p(z|\mathbf{y}) = \int_{\sigma^2} \int_\mu p\left(z|\mu, \sigma^2\right) p\left(\mu|\sigma^2, \mathbf{y}\right) p\left(\sigma^2|\mathbf{y}\right) \mathrm{d}\,\mu\,\mathrm{d}\,\sigma^2.$$

As in example 13.1, we will consider the noninformative Jeffreys prior $p\left(\mu,\sigma^2\right) \propto 1/\sigma^2$, so that the conditional posterior distribution of μ and marginal posterior distribution of σ^2 are

$$\mu|\sigma^2,\mathbf{y} \sim \mathcal{N}\left(\bar{y},\sigma^2/n\right) \quad \text{and} \quad \sigma^2|\mathbf{y} \sim \text{Inv-Gamma}\left(\frac{n-1}{2},\frac{(n-1)}{2}s^2\right),$$

respectively. Then, noting that Z is similarly drawn from a $\mathcal{N}\left(\mu,\sigma^2\right)$ distribution, we have

$$\begin{aligned}
p\left(z|\mathbf{y}\right) &= \int_{\sigma^2} p\left(\sigma^2|\mathbf{y}\right) \int_{\mu} p\left(z|\mu,\sigma^2\right) p\left(\mu|\sigma^2,\mathbf{y}\right) \mathrm{d}\,\mu\,\mathrm{d}\,\sigma^2 \\
&= \int_{\sigma^2} p\left(\sigma^2|\mathbf{y}\right) \int_{-\infty}^{\infty} \left[\left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{\left(z-\mu\right)^2}{2\sigma^2}\right\}\right] \left[\left(\frac{2\pi\sigma^2}{n}\right)^{-1/2} \exp\left\{-\frac{\left(\mu-\bar{y}\right)^2}{2\sigma^2/n}\right\}\right] \mathrm{d}\,\mu\,\mathrm{d}\,\sigma^2 \\
&= \int_{\sigma^2} \left(\frac{2\pi\sigma^2}{n}\right)^{-1/2} p\left(\sigma^2|\mathbf{y}\right) \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{\left(z-\mu\right)^2}{2\sigma^2} - \frac{\left(\mu-\bar{y}\right)^2}{2\sigma^2/n}\right\} \mathrm{d}\,\mu\,\mathrm{d}\,\sigma^2.
\end{aligned}$$

The inner integral becomes

$$\begin{aligned}
&\int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\left(z-\mu\right)^2 + n\left(\mu-\bar{y}\right)^2\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 - 2\mu z + \mu^2 + n\left(\mu^2 - 2\mu\bar{y} + \bar{y}^2\right)\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 - 2\mu z + \mu^2 + n\mu^2 - 2\mu n\bar{y} + n\bar{y}^2\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\mu^2\left(n+1\right) - 2\mu\left(z+n\bar{y}\right) + z^2 + n\bar{y}^2\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[\mu^2\left(n+1\right) - 2\mu\left(z+n\bar{y}\right) + z^2 + n\bar{y}^2 + \frac{\left(z+n\bar{y}\right)^2}{n+1} - \frac{\left(z+n\bar{y}\right)^2}{n+1}\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{\left(z+n\bar{y}\right)^2}{n+1}\right] - \frac{1}{2\sigma^2}\left[\mu^2\left(n+1\right) - 2\mu\left(z+n\bar{y}\right) + \frac{\left(z+n\bar{y}\right)^2}{n+1}\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{\left(z+n\bar{y}\right)^2}{n+1}\right] - \frac{n+1}{2\sigma^2}\left[\mu^2 - \frac{2\mu\left(z+n\bar{y}\right)}{n+1} + \frac{\left(z+n\bar{y}\right)^2}{\left(n+1\right)^2}\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{\left(z+n\bar{y}\right)^2}{n+1}\right] - \frac{n+1}{2\sigma^2}\left(\mu - \frac{z+n\bar{y}}{n+1}\right)^2\right\} \mathrm{d}\,\mu \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{\left(z+n\bar{y}\right)^2}{n+1}\right]\right\} \int_{-\infty}^{\infty} \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2/\left(n+1\right)}\left(\mu - \frac{z+n\bar{y}}{n+1}\right)^2\right\} \mathrm{d}\,\mu \\
&= \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{\left(z+n\bar{y}\right)^2}{n+1}\right]\right\} \left(n+1\right)^{-1/2} \\
&\quad \times \int_{-\infty}^{\infty} \left(\frac{2\pi\sigma^2}{n+1}\right)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2/\left(n+1\right)}\left(\mu - \frac{z+n\bar{y}}{n+1}\right)^2\right\} \mathrm{d}\,\mu.
\end{aligned}$$

We recognize the integrand as the density of a

$$\mathcal{N}\left(\frac{z+n\bar{y}}{n+1}, \frac{\sigma^2}{n+1}\right)$$

random variable, and it follows that

$$p\left(z|\mathbf{y}\right) = \int_{\sigma^2} \left(\frac{2\pi\sigma^2}{n}\right)^{-1/2} p\left(\sigma^2|\mathbf{y}\right) \left[\exp\left\{-\frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{\left(z+n\bar{y}\right)^2}{n+1}\right]\right\} \left(n+1\right)^{-1/2} \cdot 1\right] \mathrm{d}\,\sigma^2$$

$$
\begin{aligned}
&= \int_0^\infty \left(\frac{2\pi\sigma^2}{n}\right)^{-1/2} \left[\frac{\left[\frac{n-1}{2}s^2\right]^{(n-1)/2}}{\Gamma\left((n-1)/2\right)} \left(\sigma^2\right)^{-[(n-1)/2+1]} \exp\left\{-\frac{(n-1)\,s^2}{2\sigma^2}\right\}\right] \\
&\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right\} (n+1)^{-1/2}\,\mathrm{d}\,\sigma^2 \\
&= \left(2\pi\frac{n+1}{n}\right)^{-1/2} \frac{\left[\frac{n-1}{2}s^2\right]^{(n-1)/2}}{\Gamma\left((n-1)/2\right)} \\
&\quad \times \int_0^\infty \left(\sigma^2\right)^{-1/2} \left(\sigma^2\right)^{-[(n-1)/2+1]} \exp\left\{-\frac{(n-1)\,s^2}{2\sigma^2} - \frac{1}{2\sigma^2}\left[z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right\}\mathrm{d}\,\sigma^2 \\
&= (2\pi)^{-1/2}\left(1+\frac{1}{n}\right)^{-1/2} \frac{\left[\frac{n-1}{2}s^2\right]^{(n-1)/2}}{\Gamma\left((n-1)/2\right)} \\
&\quad \times \int_0^\infty \left(\sigma^2\right)^{-[(n-1)/2+1/2+1]} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right)\right\}\mathrm{d}\,\sigma^2.
\end{aligned}
$$

The integral becomes

$$
\begin{aligned}
&\int_0^\infty \left(\sigma^2\right)^{-[(n-1)/2+1/2+1]} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right)\right\}\mathrm{d}\,\sigma^2 \\
&= \int_0^\infty \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right)\right\}\mathrm{d}\,\sigma^2 \\
&= \int_0^\infty \left(\sigma^2\right)^{-(n/2+1)} \exp\left\{-\frac{1}{\sigma^2}\left(\frac{1}{2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right)\right\}\mathrm{d}\,\sigma^2,
\end{aligned}
$$

which we recognize as the kernel of an

$$
\text{Inv-Gamma}\left(\frac{n}{2}, \frac{1}{2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right)
$$

random variable, and it follows that

$$
\begin{aligned}
p\left(z|\mathbf{y}\right) &= (2\pi)^{-1/2}\left(1+\frac{1}{n}\right)^{-1/2} \frac{\left[\frac{n-1}{2}s^2\right]^{(n-1)/2}}{\Gamma\left((n-1)/2\right)} \Gamma\left(\frac{n}{2}\right)\left(\frac{1}{2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]\right)^{-n/2} \cdot 1 \\
&= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}} 2^{-1/2}\left(1+\frac{1}{n}\right)^{-1/2} 2^{-(n-1)/2}\left[(n-1)\,s^2\right]^{(n-1)/2} \\
&\quad \times 2^{n/2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]^{-n/2} \\
&= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}} 2^{-1/2-(n-1)/2+n/2}\left(1+\frac{1}{n}\right)^{-1/2}\left[(n-1)\,s^2\right]^{-1/2} \\
&\quad \times \left[(n-1)\,s^2\right]^{n/2}\left[(n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right]^{-n/2} \\
&= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi}\sqrt{1+\frac{1}{n}}\sqrt{(n-1)\,s^2}}\left[\frac{1}{(n-1)\,s^2}\left((n-1)\,s^2 + z^2 + n\bar{y}^2 - \frac{(z+n\bar{y})^2}{n+1}\right)\right]^{-n/2} \\
&= \frac{\Gamma\left(\frac{(n-1)+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\,\pi}s\sqrt{1+\frac{1}{n}}}\left[1 + \frac{1}{(n-1)\,s^2}\left(\frac{z^2\,(n+1) + n\bar{y}^2\,(n+1) - (z+n\bar{y})^2}{n+1}\right)\right]^{-n/2}
\end{aligned}
$$

$$= \frac{\Gamma\left(\frac{(n-1)+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\,\pi}s\sqrt{1+\frac{1}{n}}}\left[1+\frac{1}{(n-1)\,s^2}\left(\frac{1}{n+1}\left(nz^2+z^2+n^2\bar{y}^2+n\bar{y}^2-(z+n\bar{y})^2\right)\right)\right]^{-n/2}.$$

The rightmost expression in parentheses becomes

$$\begin{aligned} nz^2+z^2+n^2\bar{y}^2+n\bar{y}^2-(z+n\bar{y})^2 &= nz^2+z^2+n^2\bar{y}^2+n\bar{y}^2-\left(z^2+2nz\bar{y}+n^2\bar{y}^2\right) \\ &= nz^2+n\bar{y}^2-2nz\bar{y} \\ &= n\left(z^2-2z\bar{y}+\bar{y}^2\right) \\ &= n\left(z-\bar{y}\right)^2, \end{aligned}$$

so that

$$\begin{aligned} p\left(z|\mathbf{y}\right) &= \frac{\Gamma\left(\frac{(n-1)+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\,\pi}s\sqrt{1+\frac{1}{n}}}\left[1+\frac{1}{(n-1)\,s^2}\left(\frac{n}{n+1}\left(z-\bar{y}\right)^2\right)\right]^{-n/2} \\ &= \frac{\Gamma\left(\frac{(n-1)+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\,\pi}s\sqrt{1+\frac{1}{n}}}\left[1+\frac{1}{n-1}\frac{\left(z-\bar{y}\right)^2}{s^2\left(n+1\right)/n}\right]^{-n/2} \\ &= \frac{\Gamma\left(\frac{(n-1)+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\,\pi}s\sqrt{1+\frac{1}{n}}}\left[1+\frac{1}{n-1}\frac{\left(z-\bar{y}\right)^2}{s^2\left(1+\frac{1}{n}\right)}\right]^{-n/2} \\ &= \frac{\Gamma\left(\frac{(n-1)+1}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{(n-1)\,\pi}s\sqrt{1+\frac{1}{n}}}\left[1+\frac{1}{n-1}\left(\frac{z-\bar{y}}{s\sqrt{1+\frac{1}{n}}}\right)^2\right]^{-[(n-1)+1]/2}, \end{aligned}$$

which we recognize as the density of a t-distribution with $\nu = n-1$ degrees of freedom, location parameter $\bar{y}$, and scale parameter $\sigma^2 = s^2\left(1+1/n\right)$, i.e., the posterior predictive distribution of z is

$$z|\mathbf{y} \sim t\left(n-1, \bar{y}, \left(1+\frac{1}{n}\right)s^2\right).$$

The discussion above shows that the conjugate prior density must have the product form $p\left(\mu|\sigma^2\right)p\left(\sigma^2\right)$. For normal data, we have

$$\mu|\sigma^2 \sim \mathcal{N}\left(\mu_0, \sigma^2/\kappa_0\right) \quad \text{and} \quad \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\nu_0\sigma_0^2}{2}\right),$$

or equivalently $1/\sigma^2 \sim \text{Gamma}\left(\nu_0/2, \nu_0\sigma_0^2/2\right)$, which corresponds to the joint prior density

$$\begin{aligned} p\left(\mu, \sigma^2\right) &\propto \left(\sigma^2/\kappa_0\right)^{-1/2}\exp\left\{-\frac{1}{2\sigma^2/\kappa_0}\left(\mu-\mu_0\right)^2\right\}\cdot\left(\sigma^2\right)^{-(\nu_0/2+1)}\exp\left\{-\frac{\sigma_0^2}{2\sigma^2}\right\} \\ &= \left(\sigma^2\right)^{-[(\nu_0+1)/2+1]}\exp\left\{-\frac{\kappa_0}{2\sigma^2}\left(\frac{\sigma_0^2}{\kappa_0}+\left(\mu-\mu_0\right)^2\right)\right\}. \end{aligned}$$

We refer to this distribution as the Normal-Inverse Gamma$\left(\mu_0, \sigma_0^2/\kappa_0; \nu_0/2, \sigma_0^2/2\right)$ density. Note that κ_0 controls how informative the prior is relative to the data. The joint posterior distribution of $\left(\mu, \sigma^2\right)$ is

$$\begin{aligned} p\left(\mu, \sigma^2|\mathbf{y}\right) &\propto \left(\sigma^2\right)^{-[(\nu_0+1)/2+1]}\exp\left\{-\frac{1}{2\sigma^2}\left(\sigma_0^2+\kappa_0\left(\mu-\mu_0\right)^2\right)\right\}\left(\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i-\mu\right)^2\right\} \\ &\propto \left(\sigma^2\right)^{-[(\nu_n+1)/2+1]}\exp\left\{-\frac{\kappa_n}{2\sigma^2}\left(\frac{\sigma_n^2}{\kappa_n}+\left(\mu-\mu_n\right)^2\right)\right\}. \end{aligned}$$

Thus, $\mu, \sigma^2|\mathbf{y} \sim$ Normal-Inverse Gamma $\left(\mu_n, \sigma_n^2/\kappa_n; \nu_n/2, \sigma_n^2/2\right)$, where

$$\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}$$

$$\begin{aligned}\kappa_n &= \kappa_0 + n \\ \nu_n &= \nu_0 + n \\ \nu_n\sigma_n^2 &= \nu_0\sigma_0^2 + (n-1)\,s^2 + \frac{\kappa_0 n}{\kappa_0 + n}\,(\bar{y} - \mu_0)^2\,.\end{aligned}$$

Note that μ_n is a weighted average of the prior and sample means; ν_n combines the prior degrees of freedom and the sample size; and $\nu_n\sigma_n^2$ combines prior variation, observed variation, and variation between the sample mean and prior mean. After some algebra, it can be shown that the conditional posterior distribution of μ is

$$\mu|\sigma^2, \mathbf{y} \sim \mathcal{N}\left(\mu_n, \frac{\sigma^2}{\kappa_n}\right),$$

the marginal posterior distribution of σ^2 is

$$\sigma^2|\mathbf{y} \sim \text{Inv-Gamma}\left(\frac{\nu_n}{2}, \frac{\nu_n}{2}\sigma_n^2\right),$$

the marginal posterior distribution of μ is

$$\mu|\mathbf{y} \sim t\left(\nu_n, \mu_n, \frac{\sigma_n^2}{\kappa_n}\right),$$

and the posterior predictive distribution for a future observation is

$$z|\mathbf{y} \sim t\left(\nu_n, \mu_n, \sigma_n^2\left[1 + \frac{1}{\kappa_n}\right]\right).$$

EXAMPLE 13.3. Sunscreens are assigned Sun Protection Factor (SPF) values by the US Food and Drug Administration (FDA). SPF is a number that refers to the suncreen product's ability to block UVB radiation. A sunscreen product with a SPF of 15 will protect your skin 15 times longer from UVB than if you did not have sunscreen applied. For instance, an individual who can tolerate Y minutes of sunlight without any sunscreen can tolerate $15Y$ minutes with an SPF 15 sunscreen. The exact amount of time will of course vary from person to person. Data on 13 individuals' sun tolerance with and without a particular suncreen have been collected (see data `spf.txt`). The goal is to determine the SPF value for this sunscreen.

```
spf <- read.table("data/spf.txt", header = T, sep = "\t")
par(mgp = c(1.5,0.5,0), mar = c(2.5,3,0.5,2))
boxplot(spf, ylab = "Tolerance (min)", cex.lab = 0.75, cex.axis = 0.75)
```

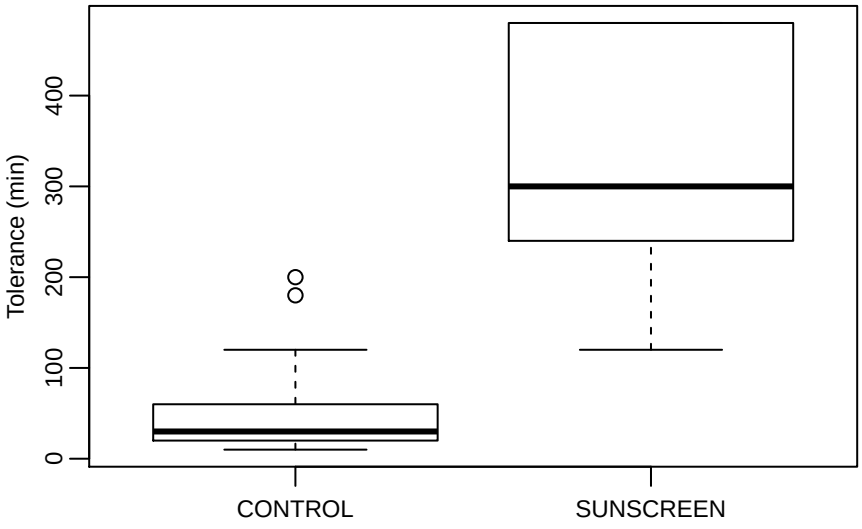


FIGURE 13.1.1. Tolerance in minutes by experimental group

This is matched pairs data and the analysis should take into account the pairing, which induces dependence between observations. We can use pairwise differences or ratios to account for the pairing. In this case, evaluating ratios makes sense since the goal of the analysis is to determine how much longer a person can be exposed to the sun relative to his/her baseline. It is reasonable to assume that the log-ratios are normally distributed. So, we are modeling

$$y = \log\left(\frac{\text{sunscreen}}{\text{control}}\right) = \log(\text{sunscreen}) - \log(\text{control}) \sim \mathcal{N}\left(\mu, \sigma^2\right).$$

We are interested in the SPF, which corresponds to $\exp(\mu)$. Let us take conjugate priors

$$\mu|\sigma^2 \sim \mathcal{N}\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right) \quad \text{and} \quad \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\sigma_0^2}{2}\right)$$

with $\mu_0 = 0$, $\kappa_0 = 0.1$, $\nu_0 = 10$, and $\sigma_0^2 = 4$. Then, the marginal posterior distribution of μ is

$$\mu|\mathbf{y} \sim t\left(\nu_n, \mu_n, \frac{\sigma_n^2}{\kappa_n}\right).$$

To draw samples of SPF from the posterior distribution, we will

(1) draw μ from its posterior distribution:
   (a) draw σ^2 from $p\left(\sigma^2|\mathbf{y}\right)$, Inv-Gamma $\left(\nu_n/2, \nu_n\sigma_n^2/2\right)$, then
   (b) draw μ from $p\left(\mu|\sigma^2, \mathbf{y}\right)$, $\mathcal{N}\left(\mu_n, \sigma^2/\kappa_n\right)$
   (c) (alternatively) draw μ directly from its marginal posterior density $p\left(\mu|\mathbf{y}\right)$, $t\left(\nu_n, \mu_n, \sigma_n^2/\kappa_n\right)$;
(2) use the transformation $\exp(\mu)$ to draw samples of SPF values.

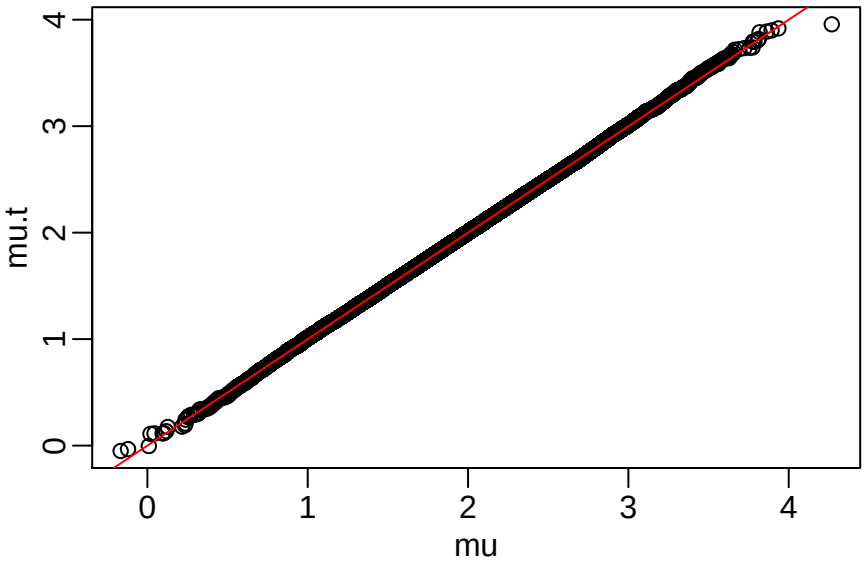
```
y <- log(spf$SUNSCREEN / spf$CONTROL)
ybar <- mean(y)
n <- length(y)

# prior hyperparameters
m0 <- 0
k0 <- 0.1
v0 <- 10
s0 <- 4

# posterior hyperparameters
kn <- k0 + n
mn <- (k0 * m0 + n * ybar) / kn
vn <- v0 + n
sn <- (v0 * s0 + (n - 1) * var(y) + (k0 * n / kn) * (ybar - m0)^2) / vn

# draw samples
nsamp <- 1e5
phi <- rgamma(n = nsamp, shape = vn / 2, rate = sn * vn / 2)
sigma2 <- 1 / phi
mu <- rnorm(nsamp, mn, sqrt(sigma2 / kn))
mu.t <- rt(nsamp, vn) * sqrt(sn / kn) + mn

# compare the distributions of the posterior samples
par(mgp = c(1.5,0.5,0), mar = c(2.5,3,0.5,2))
qqplot(mu, mu.t)
abline(a = 0, b = 1, col = "red")
```



We can obtain the 95% credible interval for the SPF:

```
# code goes here
```

or the 95% HPD interval.

We can obtain samples from the posterior predictive distribution.

13.1.4. Normal data with semi-conjugate prior distribution. The prior distributions for μ and σ^2 may be specified independently, i.e.,

$$\mu|\sigma^2 \sim \mathcal{N}\left(\mu_0, \tau_0^2\right) \quad \text{and} \quad \sigma^2 \sim \text{Inv-Gamma}\left(\frac{\nu_0}{2}, \frac{\sigma_0^2}{2}\right),$$

where we see that μ is independent of σ^2. The joint prior distribution is not a conjugate family for the normal likelihood. The posterior density (which can be shown to be proper) does not follow any standard parametric form, but we can obtain posterior samples by considering $p\left(\mu|\sigma^2, \mathbf{y}\right)$ and $p\left(\sigma^2|\mathbf{y}\right)$.

13.1.5. Multivariate normal model. Let $\mathbf{Y} = (Y_1, \ldots, Y_p)$, let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$, and let $\boldsymbol{\Sigma}$ be a $p \times p$ symmetric and positive definite matrix, where

$$\mathbf{Y}|\boldsymbol{\mu}, \boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right).$$

The sampling model for a single observation $\mathbf{y}$ is given by

$$p\left(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = (2\pi)^{-p/2} \left|\boldsymbol{\Sigma}\right|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{y} - \boldsymbol{\mu}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y} - \boldsymbol{\mu}\right)\right\}.$$

For a sample of n iid observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$,

$$\begin{aligned} p\left(\mathbf{y}_1, \ldots, \mathbf{y}_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) &= (2\pi)^{-np/2} \left|\boldsymbol{\Sigma}\right|^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left(\mathbf{y}_i - \boldsymbol{\mu}\right)^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}_i - \boldsymbol{\mu}\right)\right\} \\ &= (2\pi)^{-np/2} \left|\boldsymbol{\Sigma}\right|^{-n/2} \exp\left\{-\frac{1}{2}\,\text{tr}\left(\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\left(\mathbf{y}_i - \boldsymbol{\mu}\right)\left(\mathbf{y}_i - \boldsymbol{\mu}\right)^{\mathsf{T}}\right)\right\}, \end{aligned}$$

where $\text{tr}\left(\mathbf{A}\right)$ is the trace of a matrix $\mathbf{A}$ and given by the sum of its diagonal terms. Note that $\boldsymbol{\Sigma}$ is the covariance matrix with entries given by

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{Var}\left(y_{i1}\right) & \cdots & \text{Cov}\left(y_{i1}, y_{ik}\right) \\ \ddots & \ddots & \ddots \\ \text{Cov}\left(y_{ik}, y_{ki}\right) & & \text{Var}\left(y_{ip}\right) \end{bmatrix}.$$

The inverse-Wishart, a multivariate generalization of the inverse-Gamma, can be used to describe the prior distribution of the matrix $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{\mu}|\boldsymbol{\Sigma} \sim \mathcal{N}\left(\boldsymbol{\mu}_0, \kappa_0^{-1}\boldsymbol{\Sigma}\right) \quad \text{and} \quad \boldsymbol{\Sigma} \sim \text{Inv-Wishart}\left(\boldsymbol{\Omega}_0^{-1}, \nu_0\right).$$

The inverse-Wishart density is given by

$$p\left(\boldsymbol{\Sigma}\right) = \left[2^{p\nu_0/2}\pi^{p(p-1)/4}\prod_{i=1}^{p}\Gamma\left(\frac{\nu_0+1-i}{2}\right)\right]^{-1}\left|\boldsymbol{\Omega}_0\right|^{\nu_0/2}\left|\boldsymbol{\Sigma}\right|^{-(\nu_0+p+1)/2}\exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}_0\boldsymbol{\Sigma}^{-1}\right)\right\}.$$

The joint prior density is

$$p\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-[(\nu_0+p)/2+1]}\exp\left\{-\frac{1}{2}\text{tr}\left(\boldsymbol{\Omega}_0\boldsymbol{\Sigma}^{-1}\right) - \frac{\kappa_0}{2}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}-\boldsymbol{\mu}_0\right)\right\}.$$

The results from the univariate normal distribution generalize to the multivariate case. Multiplying the joint prior density by the normal likelihood results in a posterior density of the same family with parameters

$$\begin{aligned}
\boldsymbol{\mu}_n &= \frac{\kappa_0}{\kappa_0+n}\boldsymbol{\mu}_0 + \frac{n}{\kappa_0+n}\bar{\mathbf{y}} \\
\kappa_n &= \kappa_0 + n \\
\nu_n &= \nu_0 + n \\
\boldsymbol{\Omega}_n &= \boldsymbol{\Omega}_0 + \sum_{i=1}^{n}\left(\mathbf{y}_i-\bar{\mathbf{y}}\right)\left(\mathbf{y}_i-\bar{\mathbf{y}}\right)^{\mathsf{T}} + \frac{\kappa_0}{\kappa_0+n}\left(\bar{\mathbf{y}}-\boldsymbol{\mu}_0\right)\left(\bar{\mathbf{y}}-\boldsymbol{\mu}_0\right)^{\mathsf{T}}.
\end{aligned}$$

The marginal posterior distribution of $\boldsymbol{\mu}$ is multivariate $t\left(\nu_n-p+1, \boldsymbol{\mu}_n, \left[\kappa_n\left(\nu_n-p+1\right)\right]^{-1}\boldsymbol{\Omega}_n\right)$. Samples from the joint posterior distribution $(\boldsymbol{\mu},\boldsymbol{\Sigma})$ are easily obtained by

(1) drawing $\boldsymbol{\Sigma}|\mathbf{y} \sim \text{Inv-Wishart}\left(\boldsymbol{\Omega}_n^{-1}, \nu_n\right)$,
(2) then drawing $\boldsymbol{\mu}|\boldsymbol{\Sigma},\mathbf{y} \sim \mathcal{N}\left(\boldsymbol{\mu}_n, \kappa_n^{-1}\boldsymbol{\Sigma}\right)$.

A commonly proposed noninformative prior distribution is the multivariate Jeffreys prior density

$$p\left(\boldsymbol{\mu},\boldsymbol{\Sigma}\right) \propto \left|\boldsymbol{\Sigma}\right|^{-(p+1)/2},$$

which is the limit of the conjugate prior density as $\kappa_0 \to 0$, $\nu_0 \to -1$, and $|\boldsymbol{\Omega}_0| \to 0$. Results for the posterior distributions follow from the univariate discussions, assuming that the posterior distribution is proper. It is especially important to check that the posterior distribution is proper when using noninformative prior distributions in high dimensions.

13.1.6. Multinomial model. The binomial distribution can be generalized to allow more than two possible outcomes. The multinomial sampling distribution is used to describe data for which each observation can take one of k possible values. If $\mathbf{y} = (y_1,\ldots,y_k)$ is the vector of counts of the number of observations in each of the k categories for a sample of size n, then $\mathbf{y}|\boldsymbol{\theta} \sim \text{Multinomial}\left(n;\theta_1,\ldots,\theta_k\right)$ with density given by

$$p\left(\mathbf{y}|\boldsymbol{\theta}\right) = \frac{n!}{\prod_{j=1}^{k} y_j!}\prod_{j=1}^{k}\theta_j^{y_j} = \frac{n!}{y_1!y_2!\cdots y_k!}\theta_1^{y_1}\theta_2^{y_2}\cdots\theta_k^{y_k}$$

for $y_j = 0,\ldots,n$ and $\sum_{j=1}^{k} y_j = n$; $0 \le \theta_j \le 1$ and $\sum_j \theta_j = 1$. Let Z_i be the ith outcome, so that $Z_i \in \{1,2,\ldots,k\}$. Then, we can express Y_k as

$$Y_k = \sum_{i=1}^{n} I_{\{Z_i=k\}}.$$

The conjugate prior distribution is a multivariate generalization of the beta distribution known as the Dirichlet, which has density given by

$$p\left(\boldsymbol{\theta}\right) = \frac{\Gamma\left(\alpha_1+\cdots+\alpha_k\right)}{\Gamma\left(\alpha_1\right)\cdots\Gamma\left(\alpha_k\right)}\prod_{j=1}^{k}\theta_j^{\alpha_j-1}.$$

The resulting posterior distribution for the θ_j's is Dirichlet with parameters $\alpha_j + y_j$, i.e.,

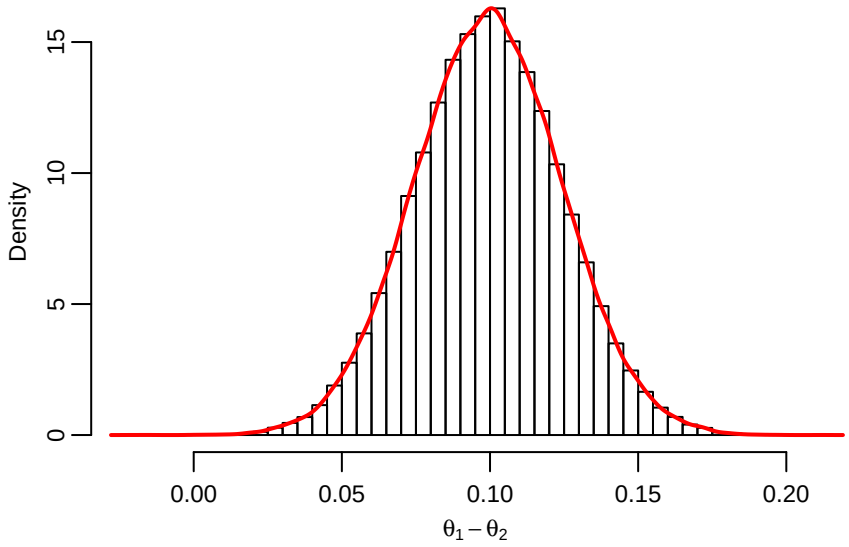
$$\boldsymbol{\theta}|\mathbf{y} \sim \text{Dirichlet}\left(\alpha_1 + y_1, \ldots, \alpha_k + y_k\right),$$

whose density we can derive as

$$\begin{aligned} p\left(\boldsymbol{\theta}|\mathbf{y}\right) &\propto p\left(\mathbf{y}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}\right) \\ &= \left[\frac{n!}{\prod_{j=1}^{k} y_j!} \prod_{j=1}^{k} \theta_j^{y_j}\right] \left[\frac{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}{\prod_{j=1}^{k} \Gamma\left(\alpha_j\right)} \prod_{j=1}^{k} \theta_j^{\alpha_j - 1}\right] \\ &\propto \left(\prod_{j=1}^{k} \theta_j^{y_j}\right) \left(\prod_{j=1}^{k} \theta_j^{\alpha_j - 1}\right) \\ &= \prod_{j=1}^{k} \theta_j^{\alpha_j + y_j - 1}. \end{aligned}$$

EXAMPLE 13.4. A survey of 1447 adults is conducted to determine preferences in an election: $y_1 = 727$ supported candidate A, $y_2 = 583$ supported candidate B, and $y_3 = 137$ supported other candidates or expressed no opinion. If we assume random sampling, then the data (y_1, y_2, y_3) follow a multinomial distribution with parameters $(\theta_1, \theta_2, \theta_3)$. A parameter of interest is $\theta_1 - \theta_2$, the population difference in support of the two major candidates. With a noninformative uniform prior distribution on θ, $\alpha_1 = \alpha_2 = \alpha_3 = 1$, the posterior distribution for $(\theta_1, \theta_2, \theta_3)$ is $\boldsymbol{\theta}|\mathbf{y} \sim \text{Dirichlet}\,(728, 584, 138)$. We could compute the posterior distribution of $\theta_1 - \theta_2$. It is simpler to draw $(\theta_1, \theta_2, \theta_3)$ from the posterior Dirichlet distribution and compute $\theta_1 - \theta_2$ for each drawn sample. You can use the function `rdirichlet` in the package `MCMCpack` to sample from a Dirichlet distribution.

```
library(MCMCpack)
set.seed(123)
thetas <- rdirichlet(1e5, c(728,584,138))
thdiff <- thetas[,1] - thetas[,2]
par(mgp = c(1.5,0.5,0), mar = c(2.5,3,0.5,2))
hist(thdiff, breaks = 50, xlab = expression(theta[1] - theta[2]),
  prob = T, main = "", cex.lab = 0.75, cex.axis = 0.75)
lines(density(thdiff), lwd = 2, col = "red")
```



```
mean(thdiff > 0)
## [1] 0.99995
```

The estimated posterior probability that candidate A has more support than candidate B in the survey population is $P(\theta_1 > \theta_2) = P(\theta_1 - \theta_2 > 0) = 99.99\%$.

CHAPTER 14

Hypothesis testing and Bayes factor

The *Bayes factor* is used to test hypotheses and compare models in the Bayesian framework. Suppose we have two candidate models, H_1 and H_2, with respective parameter vectors $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. The Bayes factor is the ratio of the posterior odds of H_1 to the prior odds of H_1, i.e.,

$$\mathrm{BF} = \frac{p\left(H_1|\mathbf{y}\right)/p\left(H_2|\mathbf{y}\right)}{p\left(H_1\right)/p\left(H_2\right)}.$$

The marginal distribution of Y under each model H_i is

$$p\left(y|H_i\right) = \int p\left(y|\boldsymbol{\theta}_i, H_i\right)\pi_i\left(\boldsymbol{\theta}_i\right)\mathrm{d}\,\boldsymbol{\theta}_i, \quad i \in \{1,2\}.$$

The Bayes factor can also be written as the ratio of the observed marginal densities for the two models, i.e.,

$$\mathrm{BF} = \frac{p\left(H_1|\mathbf{y}\right)/p\left(H_2|\mathbf{y}\right)}{p\left(H_1\right)/p\left(H_2\right)} = \frac{\left[\frac{p(\mathbf{y}|H_1)p(H_1)}{p(\mathbf{y})}\right] / \left[\frac{p(\mathbf{y}|H_2)p(H_2)}{p(\mathbf{y})}\right]}{p\left(H_1\right)/p\left(H_2\right)} = \frac{p\left(\mathbf{y}|H_1\right)}{p\left(\mathbf{y}|H_2\right)}.$$

The Bayes factor is only defined when the marginal density of $\mathbf{y}$ under each model is proper. If $\pi_i\left(\boldsymbol{\theta}_i\right)$ is improper, then $p\left(y|H_i\right)$ will necessarily be improper, and the Bayes factor is not defined. The following interpretation of the Bayes factor was proposed by Jeffreys.

$\log_{10}\mathrm{BF}$	Bayes factor	Interpretation
$0-0.5$	$1 \leq \mathrm{BF} \leq 3.2$	weak
$0.5-1.0$	$3.2 < \mathrm{BF} \leq 10$	substantial
$1.0-2.0$	$10 < \mathrm{BF} \leq 100$	strong
> 2	$\mathrm{BF} > 100$	decisive

TABLE 1. Jeffreys' scale of evidence in favor of H_1

The following alternative interpretation was proposed by Kass and Raftery.

$2\ln\mathrm{BF}$	Bayes factor	Interpretation
$0-2$	$1 \leq \mathrm{BF} \leq 3$	weak
$2-6$	$3 < \mathrm{BF} \leq 20$	positive
$6-10$	$20 < \mathrm{BF} \leq 150$	strong
> 10	$\mathrm{BF} > 150$	very strong

TABLE 2. Kass and Raftery scale of evidence in favor of H_1

14.1. Comparison to frequentist hypothesis testing

Recall from chapter 10 that, in classical hypothesis testing, we proceed as follows:

(1) state a null hypothesis H_0 and an alternative hypothesis H_1;
(2) determine an appropriate test statistic $T\left(\mathbf{Y}\right)$;

(3) compute the p-value of the test as

$$p\text{-value} = P\left(T\left(\mathbf{Y}\right) \text{ more "extreme" than } T\left(\mathbf{y}_{\text{obs}}\right)|\boldsymbol{\theta}, H_0\right)$$

where "extremeness" is in the direction of H_1;
(4) if the p-value is less than the pre-specified Type I error rate α, H_0 is rejected.

Classical hypothesis testing is straightforward only when the two hypotheses are nested. In Bayesian hypothesis testing, we proceed as follows:

(1) state the two hypotheses H_1 and H_2;
(2) assign priors to H_1 and H_2, and specific $p\left(\boldsymbol{\theta}|H_1\right)$ and $p\left(\boldsymbol{\theta}|H_2\right)$;
(3) evaluate $p\left(H_1|\mathbf{y}\right)$ and $p\left(H_2|\mathbf{y}\right)$ via Bayes' theorem;
(4) compute the Bayes factor to assess the evidence in favor of H_1:

$$\text{BF} = \frac{p\left(\mathbf{y}|H_1\right)}{p\left(\mathbf{y}|H_2\right)}.$$

The Bayesian framework does not require the two models to be nested.

EXAMPLE 14.1 (Test of proportion). Suppose 16 customers have been recruited by a fast-food chain to compare two types of ground beef patty on the basis of flavor. All of the patties to be evaluated have been kept frozen for eight months. One set of 16 has been stored in a high-quality freezer that maintains a temperature that is consistently within ±1°F. The other set of 16 has been stored in a freezer with temperature that varies anywhere between 0 and 15°F. The food chain executives are interested in whether storage in the higher-quality freezer translates into a substantial improvement in taste, thus justifying the extra effort and cost associated with equipping all of their stores with these freezers. Suppose that to be regarded as "substantial" improvement, more than 60% of consumers must prefer the more expensive option. 13 of the 16 consumers state a preference for the more expensive patty. Let $Y_i = 1$ if consumer i states a preference for the more expensive patty and $Y_i = 0$ otherwise, i.e., $Y_i \sim \text{Bernoulli}\left(\theta\right)$, so that

$$X = \sum_{i=1}^{16} Y_i \sim \text{Binomial}\left(16, \theta\right)$$

is the total number of consumers who state a preference for the more expensive patty. We want to test

$$H_1 : \theta > 0.6 \quad \text{vs} \quad H_2 \leq 0.6.$$

(Observe that this test cannot be done in a frequentist setting, where a single value must be specified for the null hypothesis.) The executives may believe that the choice could go either way, i.e., $p\left(H_1\right) = p\left(H_2\right) = 0.5$. Suppose we consider "minimally informative" priors $\pi\left(\theta\right)$: Jeffreys prior, $\text{Beta}\left(0.5, 0.5\right)$; a prior that we think of as "noninformative," $\text{Beta}\left(1, 1\right)$; and $\text{Beta}\left(2, 2\right)$. Example 11.4 implies that the posterior distribution is

$$\theta|x \sim \text{Beta}\left(\alpha + x, \beta + n - x\right).$$

The Bayes factor in favor of H_1 is

$$\text{BF} = \frac{p\left(H_1|x\right)/p\left(H_2|x\right)}{p\left(H_1\right)/p\left(H_2\right)}.$$

We can calculate the Bayes factors as shown below.

```
n <- 16
x <- 13
# for Beta(0.5, 0.5) prior
a <- 0.5
b <- 0.5
post.ratio <- (1 - pbeta(0.6, a + x, b + n - x)) / pbeta(0.6, a + x, b + n - x)
prior.ratio <- (1 - pbeta(0.6, a, b)) / pbeta(0.6, a, b)
BF <- post.ratio / prior.ratio
BF
```

```
## [1] 34.43177
```

Thus, under the Beta $(0.5, 0.5)$ prior, the Bayes factor is

$$\text{BF} = \frac{P(\theta > 0.6|x)/P(\theta \leq 0.6|x)}{P(\theta > 0.6)/P(\theta \leq 0.6)} \approx 34.4317663.$$

Since $\log_{10} 34.4317663 \approx 1.5369593$ (or $2 \ln 34.4317663 \approx 7.0779592$, there is strong evidence in favor of H_1. The prior distributions are shown with the resulting posterior quantiles and Bayes factors. We see that,

	Posterior quantile				
Prior	0.025	0.5	0.975	$p(\theta > 0.6\|x)$	Bayes factor
Beta $(0.5, 0.5)$	0.579	0.806	0.944	0.964	34.432
Beta $(1, 1)$	0.566	0.788	0.932	0.954	30.812
Beta $(2, 2)$	0.544	0.758	0.909	0.930	24.604

TABLE 3. Posterior quantiles and Bayes factors for selected prior distributions

under either the Jeffreys or Kass and Raftery interpretation, there is strong evidence in favor of $H_1 : \theta > 0.6$. Alternatively, we can calculate the Bayes factor as

$$\text{BF} = \frac{p(x|H_1)}{p(x|H_2)}, \quad \text{where} \quad p(x|H_j) = \int_\theta p(x|\theta, H_j)\, p(\theta|H_j)\, \mathrm{d}\theta.$$

We have

$$\begin{aligned} p(\theta|H_1) &= \frac{p(\theta, H_1)}{p(H_1)} \\ &= \frac{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_{0.6}^1 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta} \\ &= \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_{0.6}^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta}, \quad \theta > 0.6, \end{aligned}$$

so that

$$\begin{aligned} p(x|H_1) &= \int_\theta \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_{0.6}^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta}\, \mathrm{d}\theta \\ &= \binom{n}{x} \frac{\int_{0.6}^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, \mathrm{d}\theta}{\int_{0.6}^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta}. \end{aligned}$$

Similarly, we have

$$p(\theta|H_2) = \frac{\theta^{\alpha-1}(1-\beta)^{\beta-1}}{\int_0^{0.6} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta}, \quad \theta \leq 0.6,$$

so that

$$p(x|H_2) = \binom{n}{x} \frac{\int_0^{0.6} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, \mathrm{d}\theta}{\int_0^{0.6} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta}.$$

Then, the Bayes factor is

$$\text{BF} = \frac{p(x|H_1)}{p(x|H_2)} = \frac{\int_{0.6}^1 \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, \mathrm{d}\theta}{\int_0^{0.6} \theta^{\alpha+x-1}(1-\theta)^{\beta+n-x-1}\, \mathrm{d}\theta} \times \left[\frac{\int_{0.6}^1 \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta}{\int_0^{0.6} \theta^{\alpha-1}(1-\theta)^{\beta-1}\, \mathrm{d}\theta}\right]^{-1},$$

i.e., the Bayes factor is the product of the ratio of posteriors and the ratio of priors. Let us analyze these data using frequentist approaches.

```
prop.test(13, 16, p = 0.6, alternative = "greater")
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  13 out of 16, null probability 0.6
## X-squared = 2.1901, df = 1, p-value = 0.06945
## alternative hypothesis: true p is greater than 0.6
## 95 percent confidence interval:
##  0.5781717 1.0000000
## sample estimates:
##      p
## 0.8125
```

```
binom.test(13, 16, p = 0.6, alternative = "greater")
```

```
##
##  Exact binomial test
##
## data:  13 and 16
## number of successes = 13, number of trials = 16, p-value = 0.06515
## alternative hypothesis: true probability of success is greater than 0.6
## 95 percent confidence interval:
##  0.5834277 1.0000000
## sample estimates:
## probability of success
##                 0.8125
```

We see that we would not reject the null hypothesis at $\alpha = 0.05$.

EXAMPLE 14.2 (Two-sided test of normal mean). John weighed 170 pounds last year and he is wondering if he still weighs the same. For simplicity, assume he knows the accuracy of the scale and $\sigma = 3$ pounds. We wish to test

$$H_1 : \mu \neq 170 \quad \text{vs} \quad H_2 : \mu = 170.$$

He weighs himself 10 times and obtains the following measurements $y_i \sim \mathcal{N}\left(\mu_0, \sigma^2\right)$.

$$182 \quad 172 \quad 173 \quad 176 \quad 176 \quad 180 \quad 173 \quad 174 \quad 179 \quad 175$$

John seems to believe his weight could have remained the same or could have changed, i.e., $p\left(H_1\right) = p\left(H_2\right) = 0.5$. Under H_1, he may think that it's more likely that μ is close to 170 than far from it and can take as a prior a normal distribution with mean 170 and standard deviation τ, i.e., $\pi\left(\mu\right) \sim \mathcal{N}\left(\mu_0, \tau^2\right)$, so that

$$\begin{aligned}
p\left(\mathbf{y}|H_1\right) &= \int_{\mu} p\left(\mathbf{y}|\mu, H_1\right) p\left(\mu|H_1\right) \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \left[\left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2\right\}\right] \left[\left(2\pi\tau^2\right)^{-1/2} \exp\left\{-\frac{\left(\mu - \mu_0\right)^2}{2\tau^2}\right\}\right] \mathrm{d}\,\mu \\
&= \left(2\pi\sigma^2\right)^{-n/2} \left(2\pi\tau^2\right)^{-1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - \mu\right)^2 - \frac{1}{2\tau^2}\left(\mu - \mu_0\right)^2\right\} \mathrm{d}\,\mu \\
&= \left(2\pi\sigma^2\right)^{-n/2} \left(2\pi\tau^2\right)^{-1/2} \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n}\left(y_i^2 - 2\mu y_i + \mu^2\right)}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\tau^2}\right]\right\} \mathrm{d}\,\mu \\
&= \left(2\pi\sigma^2\right)^{-n/2} \left(2\pi\tau^2\right)^{-1/2} \\
&\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} 2\mu y_i + \sum_{i=1}^{n}\mu^2}{\sigma^2} + \frac{\mu^2 - 2\mu\mu_0 + \mu_0^2}{\tau^2}\right]\right\} \mathrm{d}\,\mu \\
&= \left(2\pi\sigma^2\right)^{-n/2} \left(2\pi\tau^2\right)^{-1/2}
\end{aligned}$$

$$
\begin{aligned}
&\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^n y_i^2}{\sigma^2} - \frac{2\mu\sum_{i=1}^n y_i}{\sigma^2} + \frac{n\mu^2}{\sigma^2} + \frac{\mu^2}{\tau^2} - \frac{2\mu\mu_0}{\tau^2} + \frac{\mu_0^2}{\tau^2}\right]\right\} \mathrm{d}\,\mu \\
&= \left(2\pi\sigma^2\right)^{-n/2}\left(2\pi\tau^2\right)^{-1/2} \\
&\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) - 2\mu\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right) + \frac{\sum_{i=1}^n y_i^2}{\sigma^2} + \frac{\mu_0^2}{\tau^2}\right]\right\} \mathrm{d}\,\mu \\
&= \left(2\pi\sigma^2\right)^{-n/2}\left(2\pi\tau^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^n y_i^2}{\sigma^2} + \frac{\mu_0^2}{\tau^2}\right]\right\} \\
&\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) - 2\mu\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\} \mathrm{d}\,\mu \\
&\propto \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) - 2\mu\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) - 2\mu\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right.\right. \\
&\quad \left.\left. + \frac{\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} - \frac{\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right]\right\} \mathrm{d}\,\mu \\
&= \exp\left\{-\frac{1}{2}\left[-\frac{\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right]\right\} \\
&\quad \times \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left[\mu^2\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right) - 2\mu\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right) + \frac{\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right]\right\} \mathrm{d}\,\mu \\
&\propto \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left[\mu^2 - 2\mu\frac{\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} + \frac{\left(\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)^2}{\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^2}\right]\right\} \mathrm{d}\,\mu \\
&= \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left(\mu - \frac{\frac{\sum_{i=1}^n y_i}{\sigma^2}+\frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}}\right)^2\right\} \mathrm{d}\,\mu.
\end{aligned}
$$

Let

$$\mu_n = \frac{\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}}{\frac{n}{\sigma^2}+\frac{1}{\tau^2}} \quad \text{and} \quad \frac{1}{\tau_n^2} = \frac{n}{\sigma^2}+\frac{1}{\tau^2},$$

so that

$$
\begin{aligned}
p\left(\mathbf{y}|H_1\right) &\propto \int_{-\infty}^{\infty} \exp\left\{-\frac{1}{2\tau_n^2}\left(\mu-\mu_n\right)^2\right\} \mathrm{d}\,\mu \\
&= \left(2\pi\tau_n^2\right)^{1/2} \int_{-\infty}^{\infty} \left(2\pi\tau_n^2\right)^{-1/2} \exp\left\{-\frac{1}{2\tau_n^2}\left(\mu-\mu_n\right)^2\right\} \mathrm{d}\,\mu.
\end{aligned}
$$

We recognize the integrand as the density of a $\mathcal{N}\left(\mu_n, \tau_n^2\right)$ random variable, so that

$$p\left(y|H_1\right) \propto \left(2\pi\tau_n^2\right)^{1/2} \cdot 1.$$

Adding back the constants we dropped gives

$$
\begin{aligned}
p\left(\mathbf{y}|H_1\right) = \left(2\pi\sigma^2\right)^{-n/2}\left(2\pi\tau^2\right)^{-1/2} &\exp\left\{-\frac{1}{2}\left[\frac{\sum_{i=1}^n y_i^2}{\sigma^2} + \frac{\mu_0^2}{\tau^2}\right]\right\} \\
&\times \exp\left\{-\frac{1}{2}\left[-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\}\left(2\pi\tau_n^2\right)^{1/2}
\end{aligned}
$$

$$= \left(2\pi\sigma^2\right)^{-n/2}\left[\frac{2\pi\tau_n^2}{2\pi\tau^2}\right]^{1/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{1}{2}\left[\frac{\mu_0^2}{\tau^2}-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\}$$
$$= \left(2\pi\sigma^2\right)^{-n/2}\left[\frac{\tau_n^2}{\tau^2}\right]^{1/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{1}{2}\left[\frac{\mu_0^2}{\tau^2}-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\}.$$

Now, under H_2,

$$\pi(\mu) = I_{\{\mu=\mu_0\}} = \begin{cases} 1, & \text{if } \mu=\mu_0 \\ 0, & \text{otherwise} \end{cases},$$

i.e., $\pi(\mu)$ is a point mass at μ_0, so that

$$\begin{aligned}
p(\mathbf{y}|H_2) &= \int_\mu p(y|\mu,H_2)\,p(\mu|H_2)\,\mathrm{d}\,\mu \\
&= p(y|\mu_0) \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (y_i-\mu_0)^2\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(y_i^2-2\mu_0 y_i+\mu_0^2\right)\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n y_i^2-\sum_{i=1}^n 2\mu_0 y_i+\sum_{i=1}^n \mu_0^2\right]\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n y_i^2-2\mu_0 n\bar{y}+n\mu_0^2\right]\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{1}{2\sigma^2}\left[n\mu_0^2-2\mu_0 n\bar{y}\right]\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{1}{2\sigma^2}\left[n\mu_0^2-2\mu_0 n\bar{y}+n\bar{y}^2-n\bar{y}^2\right]\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{1}{2\sigma^2}\left(n\mu_0^2-2\mu_0 n\bar{y}+n\bar{y}^2\right)\right\}\exp\left\{\frac{n\bar{y}^2}{2\sigma^2}\right\} \\
&= \left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{n}{2\sigma^2}(\mu_0-\bar{y})^2\right\}\exp\left\{\frac{n\bar{y}^2}{2\sigma^2}\right\}.
\end{aligned}$$

Then, the Bayes factor in support of H_1 is

$$\begin{aligned}
\mathrm{BF} &= \frac{\left(2\pi\sigma^2\right)^{-n/2}\left[\frac{\tau_n^2}{\tau^2}\right]^{1/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{1}{2}\left[\frac{\mu_0^2}{\tau^2}-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\}}{\left(2\pi\sigma^2\right)^{-n/2}\exp\left\{-\frac{\sum_{i=1}^n y_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{n}{2\sigma^2}(\mu_0-\bar{y})^2\right\}\exp\left\{\frac{n\bar{y}^2}{2\sigma^2}\right\}} \\
&= \frac{\left(\tau^2/\tau_n^2\right)^{-1/2}\exp\left\{-\frac{1}{2}\left[\frac{\mu_0^2}{\tau^2}-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\}}{\exp\left\{\frac{n\bar{y}^2}{2\sigma^2}\right\}\exp\left\{-\frac{n}{2\sigma^2}(\mu_0-\bar{y})^2\right\}} \\
&= \frac{\left(\tau^2/\tau_n^2\right)^{-1/2}\exp\left\{-\frac{1}{2}\left[\frac{\mu_0^2}{\tau^2}-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]\right\}\exp\left\{-\frac{n\bar{y}^2}{2\sigma^2}\right\}}{\exp\left\{-\frac{n}{2\sigma^2}(\mu_0-\bar{y})^2\right\}} \\
&= \frac{\left(\tau^2/\tau_n^2\right)^{-1/2}\exp\left\{-\frac{1}{2}\left[\frac{\mu_0^2}{\tau^2}-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)\right]-\frac{n\bar{y}^2}{2\sigma^2}\right\}}{\exp\left\{-\frac{n}{2\sigma^2}(\mu_0-\bar{y})^2\right\}} \\
&= \frac{\left(\tau^2/\tau_n^2\right)^{-1/2}\exp\left\{-\frac{1}{2}\left[\frac{\mu_0^2}{\tau^2}-\mu_n\left(\frac{n\bar{y}}{\sigma^2}+\frac{\mu_0}{\tau^2}\right)+\frac{n\bar{y}^2}{\sigma^2}\right]\right\}}{\exp\left\{-\frac{n}{2\sigma^2}(\mu_0-\bar{y})^2\right\}}
\end{aligned}$$

ALGEBRAIC ISSUES, FIX LATER, FINAL RESULT SHOULD BE

$$\text{BF} = \frac{\left(\sigma^2/n + \tau^2\right)^{-1/2} \exp\left\{-\frac{1}{2(\sigma^2/n+\tau^2)}\left(\bar{y}-\mu_0\right)^2\right\}}{\sqrt{n}/\sigma \exp\left\{-\frac{n}{2\sigma^2}\left(\bar{y}-\mu_0\right)^2\right\}}.$$

Let us consider different values of τ.

```
weights <- c(182,172,173,176,176,180,173,174,179,175)
ybar <- mean(weights)
n <- length(weights)
sigma <- 3
mu0 <- 170
tau <- c(0.5,1,2,5,10)
dnorm(ybar, mu0, sqrt(sigma^2 / n + tau^2)) / dnorm(ybar, mu0, sigma / sqrt(n))
```

```
## [1] 6.839262e+01 2.566051e+04 5.278923e+06 4.513721e+07 3.833448e+07
```

We see that there is very strong evidence in favor of H_1; his current weight is substantially different from 170 lbs.

14.1.1. Relationship to model choice criteria. For large sample sizes n, Schwarz (1978) showed that an approximation to $-2\log \text{BF}$ is given by

$$\text{BIC} = -2\log \frac{\sup_{H_1} f\left(y|\hat{\theta}\right)}{\sup_{H_2} f\left(y|\theta\right)} - \left(p_2 - p_1\right)\log n,$$

where

$$-2\log \frac{\sup_{H_1} f\left(y|\hat{\theta}\right)}{\sup_{H_2} f\left(y|\theta\right)}$$

is the usual likelihood ratio rest statistic and p_i is the number of parameters in model H_i. BIC stands for Bayesian Information Criterion (also known as the Schwartz Criterion). The second term in BIC acts as a penalty term which corrects for differences in size between the models.

CHAPTER 15

Monte Carlo methods

In simple models, especially if conjugate prior distributions are assumed, it is often easy to draw from the posterior distribution without difficulty. When the posterior density does not have a recognizable form, it might be possible to factor the distribution analytically and simulate in parts, as we have done in previous chapters. For more complicated problems, it is not possible to directly generate samples from the target distribution and indirect sampling schemes are used.

15.1. Direct sampling

Monte Carlo sampling is often used in two kinds of related problems:

- Sampling from a distribution $f(\theta)$ (this will often be a posterior distribution for this course).
- Computing approximate integrals of the form $\int h(\theta) f(\theta) \, \mathrm{d}\theta$, i.e., computing the expectation of $h(\theta)$ using the density $f(\theta)$.

The above problems are related because if we can sample from $f(\theta)$, then we can also solve the problem of computing integrals. Suppose we can draw samples $\theta^{(1)}, \ldots, \theta^{(K)}$ from $f(\theta|y)$. If we want to evaluate $\mathrm{E}[h(\theta|y)] = \int_\theta h(\theta) f(\theta|y) \, \mathrm{d}\theta$, then theorem 5.8 implies that

$$\frac{1}{K} \sum_{k=1}^{K} h\left(\theta^{(k)}\right) \to \mathrm{E}[h(\theta|y)] \quad \text{as } K \to \infty.$$

This general procedure is called Monte Carlo integration. Other methods for direct sampling include

(1) cumulative ordered values, which approximate the cdf $F(\theta|Y)$;
(2) the empirical distribution of the samples $\theta^{(1)}, \ldots, \theta^{(K)}$, which approximates $f(\theta|Y)$ (construct a histogram or kernel density estimator);
(3) the proportion of samples where the event $h\left(\theta^{(k)}\right) > g$ occurs, which approximates $P(\{h(\theta) > c\})$;
(4) and sample moments/quantiles/functions, which approximate true moments/quantiles/functions.

These approaches extend easily to higher dimensions.

EXAMPLE 15.1. Suppose $f(\theta|Y) \sim \mathrm{Beta}(7, 82)$. The exact posterior mean is $7/(7+82) = 0.0786$. The median and 95% credible intervals are obtained below.

```
qbeta(c(0.025,0.5,0.975), 7, 82)
```

```
## [1] 0.03258001 0.07550366 0.14251617
```

We can also estimate the parameters through simulation.

```
set.seed(123)
theta <- rbeta(1e4, 7, 82)
mean(theta)
```

```
## [1] 0.07876284
```

```
quantile(theta, c(0.025,0.5,0.725))
```

```
##       2.5%        50%      72.5%
## 0.03168530 0.07548103 0.09360718
```

As the number of draws K increases, the estimate will approach the true value. We can also obtain HPD intervals for θ and the odds ratio $\theta/(1-\theta)$.

```
library(coda)
theta.post <- as.mcmc(theta)
HPDinterval(theta.post)
```

```
##           lower     upper
## var1 0.02686526 0.1340657
## attr(,"Probability")
## [1] 0.95
```

```
odds.post <- as.mcmc(theta / (1 - theta))
HPDinterval(odds.post)
```

```
##           lower     upper
## var1 0.02717214 0.1543494
## attr(,"Probability")
## [1] 0.95
```

Observe that HPD intervals are not invariant under transformation.

15.2. Inverse transformation method

Part 4

Numerical methods

CHAPTER 16

Regression

Let y_i be the ith observation of some response variable, and let $\mathbf{x}^{(i)} \in \mathbb{R}^n$ be the corresponding vector of predictors, so that the data are $\left(y_i, \mathbf{x}^{(i)}\right)$ for $i \in \{1, 2, \ldots, N\}$. Then, the regression model is

$$y_i \sim \alpha_0 + \alpha_1 x_1^{(i)} + \cdots + \alpha_n x_n^{(i)}.$$

Suppose that we wish to examine a hypothesized linear relationship between height and weight in some population of interest. Let y_i and x_i be the height and weight, respectively, of the ith person, so that the regression model is $y_i \sim \alpha_0 + \alpha_1 x_i$. There are infinitely many unique choices of $\boldsymbol{\alpha} = (\alpha_0, \alpha_1)$, so that there are infinitely many lines that "fit" the data. The difference between the observed and predicted values for the ith person, i.e., the ith residual, is then $r_i = y_i - (\alpha_0 + \alpha_1 x_i)$. Then, we define the line that "best" fits the data as that whose coefficients $\boldsymbol{\alpha}$ minimize the least-squares loss function

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{N} r_i^2 = \sum_{i=1}^{N} (y_i - \alpha_0 - \alpha_1 x_i)^2,$$

i.e., we must solve the optimization problem

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^2} \sum_{i=1}^{N} (y_i - \alpha_0 - \alpha_1 x_i)^2.$$

Suppose we wish to optimize the function $f(x) = 2x^2 + 3$, where $x \in \mathbb{R}$. We might choose to solve $f'(x) = 0$, or we might choose to lay down a grid over some domain, graph the function, and perform an exhaustive search. Each evaluation of f requires 2 multiplications and one addition. For a grid whose elements are the integers between -10 and 10, the exhaustive search method will thus require $2 \cdot 21 = 42$ multiplications, which is a nearly trivial computational task. Now suppose that $f(\mathbf{x}) = 3x_1^2 + 2x_2^2 - x_1 x_2 + 5$, where $\mathbf{x} \in \mathbb{R}^2$. We might now lay down a (2-dimensional) grid given by

$$\{\mathbf{x} \,|\, x_1 \in \{-10, -9, \ldots, 9, 10\}, x_2 \in \{-10, -9, \ldots, 9, 10\}\}.$$

Each evaluation of f requires 5 multiplications and 3 additions, so that the exhaustive search method requires $5 \cdot 21^2$ multiplications. Now suppose that $\mathbf{x} \in \mathbb{R}^n$, and consider a real-valued function $f(\mathbf{x})$. Suppose further that for the ith component of $\mathbf{x}$, we lay down a grid such that $x_i \in \{-10, -9, \ldots, 9, 10\}$. The number of multiplications required by the exhaustive search method is now on the order of 21^n. If we assume that we can perform 10^6 multiplications per second, then for $n = 5$, we must evaluate f

$$21^5 \approx (2 \cdot 10)^5 = 2^5 10^5 = 32 \cdot 10^5 \approx 3 \cdot 10 \cdot 10^5 = 3 \cdot 10^6$$

total times, so that the optimization will require roughly 3 seconds. If $n = 10$, then we must perform $21^{10} \approx 2^{10} \cdot 10^{10}$ function evaluations, so that the optimization will require roughly $2^{10} \cdot 10^4 = 1024 \cdot 10^4 \approx 10^3 \cdot 10^4 = 10^7$ seconds, or equivalently, 116 days. We see that the exhaustive search method will not in general be feasible, particularly in high dimensions.

We now consider the alternative method, i.e., solving $f'(x) = 0$. For $f(x) = 2x^2 + 3$, we have $f'(x) = 4x \implies x = 0$, so that f is minimized by $x = 0$. But $f'(x) = 0$ cannot usually be solved analytically, e.g., suppose that $f(x) = \mathrm{e}^x + x^2 \implies f'(x) = \mathrm{e}^x + 2x$.

16.1. Optimization of a quadratic function

Suppose that $f(x) = ax^2 + bx + c$, so that $f'(x) = 2ax + b$. Then, a critical point of f occurs at $x = -b/(2a)$. Now suppose that $f(\mathbf{x}) = 2x_1^2 + 3x_2^2 + x_1x_2 - 2x_1 - 5x_2 + 5$, i.e., $\mathbf{x} \in \mathbb{R}^2$. To find the critical point of this function, we will set the gradient of f equal to zero. We have

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 4x_1 + x_2 - 2 \\ 6x_2 + x_1 - 5 \end{bmatrix},$$

so that

$$\nabla f(\mathbf{x}) = \mathbf{0} \implies \begin{bmatrix} 4x_1 + x_2 - 2 \\ x_1 + 6x_2 - 5 \end{bmatrix} = \begin{bmatrix} 4 & 1 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 2 \\ 5 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies \begin{bmatrix} 4 & 1 \\ 1 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \end{bmatrix}.$$

Let

$$\mathbf{A} = \begin{bmatrix} 4 & 1 \\ 1 & 6 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 2 \\ 5 \end{bmatrix},$$

so that $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$. Now suppose that $n > 1$, i.e., f has the form of an n-dimensional paraboloid.

PROPOSITION 16.1. *Any quadratic $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ can be written as $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{M}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{x} + c$, where $\mathbf{M}$ is a symmetric $n \times n$ matrix, $\mathbf{b}$ is an n-dimensional vector, and c is a scalar.*

EXAMPLE 16.2. Let $f(\mathbf{x}) = 5x_1^2 + 3x_1x_2 + 5x_2 + 3$. We can write f as

$$f(\mathbf{x}) = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 5 & 3 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + 3 = \mathbf{x}^\mathsf{T}\mathbf{M}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{x} + c,$$

where

$$\mathbf{M} = \begin{bmatrix} 5 & 3 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 5 \end{bmatrix}, \quad \text{and} \quad c = 3.$$

Observe also that

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 5 & 3/2 \\ 3/2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 5x_1 + 3x_2/2 \\ 3x_1/2 \end{bmatrix} = 5x_1^2 + \frac{3}{2}x_1x_2 + \frac{3}{2}x_1x_2 = 5x_1^2 + 3x_1x_2,$$

i.e., we can take $\mathbf{M}$ as symmetric.

PROPOSITION 16.3. *Let $\mathbf{u} \in \mathbb{R}^n$, let $\mathbf{v} \in \mathbb{R}^m$, and let $\mathbf{M}$ be an $n \times m$ matrix. Then,*

$$\mathbf{u}^\mathsf{T}\mathbf{M}\mathbf{v} = \sum_{j=1}^{m}\sum_{i=1}^{n} u_i M_{ij} v_j.$$

PROOF. Let M_{ij} be the ijth component of $\mathbf{M}$, and let $\mathbf{m}_j$ be the jth column of $\mathbf{M}$, so that

$$\mathbf{M}\mathbf{v} = \begin{bmatrix} \mathbf{m}_1 & \mathbf{m}_2 & \cdots & \mathbf{m}_m \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix} = \mathbf{m}_1 v_1 + \mathbf{m}_2 v_2 + \cdots + \mathbf{m}_m v_m = \sum_{j=1}^{m} \mathbf{m}_j v_j,$$

so that

$$\mathbf{u}^\mathsf{T}\mathbf{M}\mathbf{v} = \mathbf{u} \cdot \mathbf{M}\mathbf{v} = \mathbf{u} \cdot \sum_{j=1}^{m} \mathbf{m}_j v_j = \sum_{j=1}^{m} \mathbf{u} \cdot \mathbf{m}_j v_j = \sum_{j=1}^{m} (u_1 M_{1j} + u_2 M_{2j} + \cdots + u_n M_{nj}) v_j = \sum_{j=1}^{m}\sum_{i=1}^{n} u_i M_{ij} v_j.$$

□

PROPOSITION 16.4. *The gradient of an n-dimensional quadratic $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{M}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{x} + c$ is*

$$\nabla f(\mathbf{x}) = 2\mathbf{M}\mathbf{x} + \mathbf{b},$$

where $\mathbf{M}$ and $\mathbf{b}$ are as in proposition 16.3.

PROOF. We have

$$\nabla f(\mathbf{x}) = \nabla\left(\mathbf{x}^{\mathsf{T}}\mathbf{M}\mathbf{x} + \mathbf{b}^{\mathsf{T}}\mathbf{x} + c\right) = \nabla\left(\mathbf{x}^{\mathsf{T}}\mathbf{M}\mathbf{x}\right) + \nabla\left(\mathbf{b}^{\mathsf{T}}\mathbf{x}\right) + \nabla c = \nabla\left(\mathbf{x}^{\mathsf{T}}\mathbf{M}\mathbf{x}\right) + \nabla\left(\mathbf{b}^{\mathsf{T}}\mathbf{x}\right).$$

Then, proposition 16.3 implies that

$$\nabla\left(\mathbf{x}^{\mathsf{T}}\mathbf{M}\mathbf{x}\right) = \nabla\left(\sum_{j=1}^{n}\sum_{i=1}^{n} x_i M_{ij} x_j\right).$$

The derivative with respect to x_k of any term that does not include x_k is zero. For $j \neq k$, we have

$$\begin{aligned}
\frac{\partial}{\partial x_k}\sum_{\substack{j=1\\ j\neq k}}^{n}\sum_{i=1}^{n} x_i M_{ij} x_j &= \sum_{\substack{j=1\\ j\neq k}}^{n}\frac{\partial}{\partial x_k}\left(x_1 M_{1j} x_j + \cdots + x_k M_{kj} x_j + \cdots + x_n M_{nj} x_j\right)\\
&= \sum_{\substack{j=1\\ j\neq k}}^{n}\left(0 + \cdots + M_{kj} x_j + \cdots + 0\right)\\
&= \sum_{\substack{j=1\\ j\neq k}}^{n} M_{kj} x_j,
\end{aligned}$$

and for $j = k$, we have

$$\begin{aligned}
\frac{\partial}{\partial x_k}\sum_{i=1}^{n} x_i M_{ik} x_k &= \frac{\partial}{\partial x_k}\left(x_1 M_{1k} x_k + \cdots + x_k M_{kk} x_k + \cdots + x_n M_{nk} x_k\right)\\
&= \frac{\partial}{\partial x_k} x_1 M_{1k} x_k + \cdots + \frac{\partial}{\partial x_k} x_k^2 M_{kk} + \cdots + \frac{\partial}{\partial x_k} x_n M_{nk} x_k\\
&= x_1 M_{1k} + \cdots + 2x_k M_{kk} + \cdots + x_n M_{nk}\\
&= 2x_k M_{kk} + \sum_{\substack{i=1\\ i\neq k}}^{n} x_i M_{ik},
\end{aligned}$$

so that

$$\begin{aligned}
\frac{\partial}{\partial x_k}\sum_{j=1}^{n}\sum_{i=1}^{n} x_i M_{ij} x_j &= \sum_{\substack{j=1\\ j\neq k}}^{n} M_{kj} x_j + 2x_k M_{kk} + \sum_{\substack{i=1\\ i\neq k}}^{n} x_i M_{ik}\\
&= 2x_k M_{kk} + \sum_{\substack{i=1\\ i\neq k}}^{n}\left(M_{ki} x_i + x_i M_{ik}\right)\\
&= \left(M_{kk} + M_{kk}\right) x_k + \sum_{\substack{i=1\\ i\neq k}}^{n}\left(M_{ki} + M_{ik}\right) x_i\\
&= \sum_{i=1}^{n}\left(M_{ki} + M_{ik}\right) x_i.
\end{aligned}$$

Thus,

$$\nabla\left(\mathbf{x}^{\mathsf{T}}\mathbf{M}\mathbf{x}\right) = \begin{bmatrix} \sum_{i=1}^{n}\left(M_{1i} + M_{i1}\right) x_i \\ \sum_{i=1}^{n}\left(M_{2i} + M_{i2}\right) x_i \\ \vdots \\ \sum_{i=1}^{n}\left(M_{ni} + M_{in}\right) x_i \end{bmatrix}$$

$$
\begin{aligned}
&= \begin{bmatrix} (M_{11}+M_{11})\,x_1 + (M_{12}+M_{21})\,x_2 + \cdots + (M_{1n}+M_{n1})\,x_n \\ (M_{21}+M_{12})\,x_1 + (M_{22}+M_{22})\,x_2 + \cdots + (M_{2n}+M_{n2})\,x_n \\ \vdots \\ (M_{n1}+M_{1n})\,x_1 + (M_{n2}+M_{2n})\,x_n + \cdots + (M_{nn}+M_{nn})\,x_n \end{bmatrix} \\
&= \begin{bmatrix} M_{11}+M_{11} & M_{12}+M_{21} & \cdots & M_{1n}+M_{n1} \\ M_{21}+M_{12} & M_{22}+M_{22} & \cdots & M_{2n}+M_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1}+M_{1n} & M_{n2}+M_{2n} & \cdots & M_{nn}+M_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \\
&= \left(\begin{bmatrix} M_{11} & M_{12} & \cdots & M_{1n} \\ M_{21} & M_{22} & \cdots & M_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ M_{n1} & M_{n2} & \cdots & M_{nn} \end{bmatrix} + \begin{bmatrix} M_{11} & M_{21} & \cdots & M_{n1} \\ M_{12} & M_{22} & \cdots & M_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ M_{1n} & M_{2n} & \cdots & M_{nn} \end{bmatrix} \right) \mathbf{x} \\
&= \left(\mathbf{M} + \mathbf{M}^{\mathsf{T}}\right) \mathbf{x} \\
&= 2\mathbf{M}\mathbf{x}.
\end{aligned}
$$

(**M** is symmetric)

Noting that

$$\frac{\partial}{\partial x_i}\left(b_1 x_1 + b_2 x_2 + \cdots + b_i x_i + \cdots + b_n x_n\right) = b_i,$$

we have

$$\nabla\left(\mathbf{b}^{\mathsf{T}}\mathbf{x}\right) = \nabla\left(b_1 x_1 + b_2 x_2 + \cdots + b_n x_n\right) = \mathbf{b},$$

so that

$$\nabla f(\mathbf{x}) = \nabla\left(\mathbf{x}^{\mathsf{T}}\mathbf{M}\mathbf{x}\right) + \nabla\left(\mathbf{b}^{\mathsf{T}}\mathbf{x}\right) = 2\mathbf{M}\mathbf{x} + \mathbf{b}.$$

□

The critical point of an n-dimensional quadratic occurs when $\nabla f(\mathbf{x}) = \mathbf{0}$. We now solve for the critical point.

$$\nabla f(\mathbf{x}) = 2\mathbf{M}\mathbf{x} + \mathbf{b} = \mathbf{0} \implies 2\mathbf{M}\mathbf{x} = -\mathbf{b} \implies \mathbf{M}\mathbf{x} = -\frac{1}{2}\mathbf{b} \implies \mathbf{x} = \mathbf{M}^{-1}\left(-\frac{1}{2}\mathbf{b}\right) = -\frac{1}{2}\mathbf{M}^{-1}\mathbf{b}. \tag{16.1.1}$$

REMARK 16.5. The class of quadratic functions is the only class that can be analyzed in any number of dimensions.

Consider again the least-squares loss function, which is easily seen to be quadratic. We will express $L(\boldsymbol{\alpha})$ in the form of proposition 16.1, i.e.,

$$
\begin{aligned}
L(\boldsymbol{\alpha}) &= \sum_{i=1}^{N} (y_i - \alpha_0 - \alpha_1 x_i)^2 \\
&= \sum_{i=1}^{N} \left(y_i^2 - 2y_i(\alpha_0 + \alpha_1 x_i) + (\alpha_0 + \alpha_1 x_i)^2\right) \\
&= \sum_{i=1}^{N} \left(y_i^2 - 2\alpha_0 y_i - 2\alpha_1 x_i y_i + \alpha_0^2 + 2\alpha_0\alpha_1 x_i + \alpha_1^2 x_i^2\right) \\
&= \sum_{i=1}^{N} \alpha_0^2 + \sum_{i=1}^{N} \alpha_1^2 x_i^2 + \sum_{i=1}^{N} 2\alpha_0\alpha_1 x_i - \sum_{i=1}^{N} 2\alpha_0 y_i - \sum_{i=1}^{N} 2\alpha_1 x_i y_i + \sum_{i=1}^{N} y_i^2 \\
&= N\alpha_0^2 + \alpha_1^2 \sum_{i=1}^{N} x_i^2 + \alpha_0\alpha_1\left(2\sum_{i=1}^{N} x_i\right) - \alpha_0\left(2\sum_{i=1}^{N} y_i\right) - \alpha_1\left(2\sum_{i=1}^{N} x_i y_i\right) + \sum_{i=1}^{N} y_i^2 \\
&= \begin{bmatrix}\alpha_0 & \alpha_1\end{bmatrix} \begin{bmatrix} N & \sum_{i=1}^{N} x_i \\ \sum_{i=1}^{N} x_i & \sum_{i=1}^{N} x_i^2 \end{bmatrix} \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix} + \begin{bmatrix} -2\sum_{i=1}^{N} y_i & -2\sum_{i=1}^{N} x_i y_i \end{bmatrix} \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix} + \sum_{i=1}^{N} y_i^2 \\
&= \boldsymbol{\alpha}^{\mathsf{T}}\mathbf{M}\boldsymbol{\alpha} + \mathbf{b}^{\mathsf{T}}\boldsymbol{\alpha} + c,
\end{aligned}
$$

where

$$\mathbf{M} = \begin{bmatrix} N & \sum_{i=1}^N x_i \\ \sum_{i=1}^N x_i & \sum_{i=1}^N x_i^2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} -2\sum_{i=1}^N y_i \\ -2\sum_{i=1}^N x_i y_i \end{bmatrix}, \quad \text{and} \quad c = \sum_{i=1}^N y_i^2.$$

The critical point of $L(\boldsymbol{\alpha})$ therefore occurs at $\boldsymbol{\alpha} = \mathbf{M}^{-1}\mathbf{b}/2$. Now consider $\mathbf{x} \in \mathbb{R}^n$, so that

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^N \left(y_i - \alpha_0 - \alpha_1 x_1^{(i)} - \alpha_2 x_2^{(i)} - \cdots - \alpha_n x_n^{(i)} \right)^2 = \sum_{i=1}^N r_i^2 = \|\mathbf{r}\|^2,$$

where

$$\begin{aligned}
\mathbf{r} &= \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{bmatrix} \\
&= \begin{bmatrix} y_1 - \alpha_0 - \alpha_1 x_1^{(1)} - \cdots - \alpha_n x_n^{(1)} \\ y_2 - \alpha_0 - \alpha_1 x_1^{(2)} - \cdots - \alpha_n x_n^{(2)} \\ \vdots \\ y_N - \alpha_0 - \alpha_1 x_1^{(N)} - \cdots - \alpha_n x_n^{(N)} \end{bmatrix} \\
&= \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} - \begin{bmatrix} \alpha_0 + \alpha_1 x_1^{(1)} + \cdots + \alpha_n x_n^{(1)} \\ \alpha_0 + \alpha_1 x_1^{(2)} + \cdots + \alpha_n x_n^{(2)} \\ \vdots \\ \alpha_0 + \alpha_1 x_1^{(N)} + \cdots + \alpha_n x_n^{(N)} \end{bmatrix} \\
&= \mathbf{y} - \begin{bmatrix} 1 & x_1^{(1)} & \cdots & x_n^{(1)} \\ 1 & x_1^{(2)} & \cdots & x_n^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(N)} & \cdots & x_n^{(N)} \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \\
&= \mathbf{y} - \mathbf{B}\boldsymbol{\alpha},
\end{aligned}$$

and $\mathbf{B}$ is referred to as the *model matrix*. Thus,

$$\begin{aligned}
L(\boldsymbol{\alpha}) &= \|\mathbf{r}\|^2 \\
&= \|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|^2 \\
&= (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \cdot (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \\
&= (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha})^\mathsf{T} (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \\
&= \left[\mathbf{y}^\mathsf{T} - (\mathbf{B}\boldsymbol{\alpha})^\mathsf{T} \right] (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \\
&= \left(\mathbf{y}^\mathsf{T} - \boldsymbol{\alpha}^\mathsf{T}\mathbf{B}^\mathsf{T} \right) (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \\
&= \mathbf{y}^\mathsf{T} (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) - \boldsymbol{\alpha}^\mathsf{T}\mathbf{B}^\mathsf{T} (\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}) \\
&= \mathbf{y}^\mathsf{T}\mathbf{y} - \mathbf{y}^\mathsf{T}\mathbf{B}\boldsymbol{\alpha} - \boldsymbol{\alpha}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{y} + \boldsymbol{\alpha}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{B}\boldsymbol{\alpha}.
\end{aligned}$$

Now, $\boldsymbol{\alpha}^\mathsf{T}$ is $1 \times (n+1)$, $\mathbf{B}^\mathsf{T}$ is $(n+1) \times N$, and $\mathbf{y}$ is $N \times 1$, so that $\boldsymbol{\alpha}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{y}$ is 1×1, i.e., a scalar, hence equal to its transpose. Then, letting $\mathbf{M} = \mathbf{B}^\mathsf{T}\mathbf{B}$, we have

$$\begin{aligned}
L(\boldsymbol{\alpha}) &= \boldsymbol{\alpha}^\mathsf{T}\mathbf{M}\boldsymbol{\alpha} - \mathbf{y}^\mathsf{T}\mathbf{B}\boldsymbol{\alpha} - \left(\boldsymbol{\alpha}^\mathsf{T}\mathbf{B}^\mathsf{T}\mathbf{y} \right)^\mathsf{T} + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= \boldsymbol{\alpha}^\mathsf{T}\mathbf{M}\boldsymbol{\alpha} - \mathbf{y}^\mathsf{T}\mathbf{B}\boldsymbol{\alpha} - \mathbf{y}^\mathsf{T}\mathbf{B}\boldsymbol{\alpha} + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= \boldsymbol{\alpha}^\mathsf{T}\mathbf{M}\boldsymbol{\alpha} - 2\mathbf{y}^\mathsf{T}\mathbf{B}\boldsymbol{\alpha} + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= \boldsymbol{\alpha}^\mathsf{T}\mathbf{M}\boldsymbol{\alpha} + \left(\left(-2\mathbf{y}^\mathsf{T}\mathbf{B} \right)^\mathsf{T} \right)^\mathsf{T} \boldsymbol{\alpha} + \mathbf{y}^\mathsf{T}\mathbf{y}
\end{aligned}$$

$$
\begin{aligned}
&= \boldsymbol{\alpha}^\mathsf{T}\mathbf{M}\boldsymbol{\alpha} + \left(-2\mathbf{B}^\mathsf{T}\mathbf{y}\right)^\mathsf{T}\boldsymbol{\alpha} + \mathbf{y}^\mathsf{T}\mathbf{y} \\
&= \boldsymbol{\alpha}^\mathsf{T}\mathbf{M}\boldsymbol{\alpha} + \mathbf{b}^\mathsf{T}\boldsymbol{\alpha} + c,
\end{aligned}
$$

where $\mathbf{b} = -2\mathbf{B}^\mathsf{T}\mathbf{y}$ and $c = \mathbf{y}^\mathsf{T}\mathbf{y}$. We recognize this expression as a general quadratic, so proposition 16.4 implies that its gradient is given by $\nabla L(\boldsymbol{\alpha}) = 2\mathbf{M}\boldsymbol{\alpha} + \mathbf{b}$, and (16.1.1) implies that the critical point of $L(\boldsymbol{\alpha})$ occurs at

$$
\boldsymbol{\alpha} = -\frac{1}{2}\mathbf{M}^{-1}\mathbf{b} = -\frac{1}{2}\left(\mathbf{B}^\mathsf{T}\mathbf{B}\right)^{-1}\left(-2\mathbf{B}^\mathsf{T}\mathbf{y}\right) = \left(\mathbf{B}^\mathsf{T}\mathbf{B}\right)^{-1}\mathbf{B}^\mathsf{T}\mathbf{y}, \tag{16.1.2}
$$

where (16.1.2) are referred to as the *normal equations*. Hence, least-squares regression is simply quadratic optimization.

16.2. Optimization

Consider minimizing the function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$. In the case that f is quadratic, then $\nabla f(\mathbf{x}) = \mathbf{0}$ can be solved analytically. But in the case that f is not quadratic, or more generally when $\nabla f(\mathbf{x}) = \mathbf{0}$ does not have an analytic solution, then iterative methods must be applied. Recall that the gradient of f at some point $\mathbf{x}$ gives the direction at $\mathbf{x}$ in which f increases most rapidly, or the direction of *steepest ascent*. Equivalently, $-\nabla f(\mathbf{x})$ gives the direction of *steepest descent* at $\mathbf{x}$. Given some initial point $\mathbf{x}^{(0)}$, we can therefore find a critical point of f by taking a step of length s from $\mathbf{x}^{(0)}$ in the direction of $\nabla f\left(\mathbf{x}^{(0)}\right)$, resulting in a new point $\mathbf{x}^{(1)}$. We can again take a step of length s in the direction of $\nabla f\left(\mathbf{x}^{(1)}\right)$, repeating this procedure until the gradient of f is zero, indicating that we have reached a critical point. When we implement this approach, we will normalize the direction, so that the step length is controlled by the step length parameter s. If we wish to minimize f, we will set the direction at $\mathbf{x}^{(i)}$ as

$$
\mathbf{d}^{(i)} = -\frac{\nabla f\left(\mathbf{x}^{(i)}\right)}{\left\|\nabla f\left(\mathbf{x}^{(i)}\right)\right\|}.
$$

Algorithm 1 gives the general optimization iteration.

Algorithm 1 Steepest Descent

Require: $\mathbf{x} \in \mathbb{R}^n, s \in \mathbb{R}$
1: **while** $\|\nabla f(\mathbf{x})\| > \varepsilon$ **do**
2: $\quad \mathbf{d} \leftarrow -\nabla f(\mathbf{x}) / \|\nabla f(\mathbf{x})\|$
3: $\quad \mathbf{x} \leftarrow \mathbf{x} + s\mathbf{d}$
4: **end while**

Now consider the step length parameter s. For constant s, we may take steps that are too long, in which case we may overshoot the minimum, or we may take steps that are too short, in which case the algorithm may converge slowly. Rather, we would like to choose a step size s such that s is the distance from $\mathbf{x}$ to the lowest point on the surface of f in the direction of $\mathbf{d}$. The value of f at that point is $g(s) = f(\mathbf{x} + s\mathbf{d})$,so to select s, we must minimize g. Taking the derivative of g gives

$$
\begin{aligned}
g'(s) &= \frac{\mathrm{d}}{\mathrm{d}\,s} f(\mathbf{x} + s\mathbf{d}) \\
&= \nabla f(\mathbf{x} + s\mathbf{d}) \cdot \frac{\mathrm{d}}{\mathrm{d}\,s}(\mathbf{x} + s\mathbf{d}) && \text{(Chain Rule)} \\
&= \nabla f(\mathbf{x} + s\mathbf{d}) \cdot \left(\frac{\mathrm{d}}{\mathrm{d}\,s}\mathbf{x} + \frac{\mathrm{d}}{\mathrm{d}\,s}s\mathbf{d}\right) \\
&= \nabla f(\mathbf{x} + s\mathbf{d}) \cdot (\mathbf{0} + \mathbf{d}) \\
&= \nabla f(\mathbf{x} + s\mathbf{d}) \cdot \mathbf{d}.
\end{aligned}
$$

The quantity $\nabla f(\mathbf{x} + s\mathbf{d}) \cdot \mathbf{d}$ is a scalar, so that to minimize g, we must solve

$$
g'(s) = \nabla f(\mathbf{x} + s\mathbf{d}) \cdot \mathbf{d} = 0,
$$

which is a one-dimensional root-finding problem. The solution to this equation satisfies the conditions described above for s.

16.2.1. Root-finding. We now consider the problem of root-finding. Let h be a real-valued function of $s \in \mathbb{R}$. We wish to find s^* such that $h(s^*) = 0$.

16.2.1.1. *Bisection method.* Suppose that h has a root at s^*. Assuming that h is continuous, it follows that there exist values s_L and s_R such that $h(s_L)\,h(s_R) < 0$, i.e., $h(s_L)$ and $h(s_R)$ have opposite signs. Then, s_L and s_R *bracket* the root at s^*. We then calculate the midpoint of the interval $[s_L, s_R]$ as $s_M = (s_L + s_R)/2$. Now the root will be bracketed either by $[s_L, s_M]$ or by $[s_M, s_R]$; we can determine which by calculating $h(s_L)\,h(s_M)$ and $h(s_M)\,h(s_R)$. Taking the new bracket, we repeat the procedure until the interval is sufficiently small, i.e., until $|s_L - s_R| < \varepsilon$.

Algorithm 2 Bisection Method

Require: $\{s_L, s_R : h(s_L)\,h(s_R) < 0\}$
while $|s_L - s_R| > \varepsilon$ **do**
  $s_M \leftarrow (s_L + s_R)/2$
  if $h(s_L)\,h(s_M) < 0$ **then**
    $s_R \leftarrow s_M$
  else
    $s_L \leftarrow s_M$
  end if
end while

Observe that the bisection method requires knowledge only of h, and further that once a root is bracketed, the algorithm always converges (to a root). Bracketing a root is not in general straightforward, and convergence of the algorithm is slow.

16.2.1.2. *Newton's method.* If in addition to h, we also have its derivative h', we can overcome the difficulties of the bisection method: the idea is to replace $h(s)$ by its linear approximation. The equation of the line tangent to h at s_0 is given by the point-slope formula, i.e., $h(s) - h(s_0) = h'(s_0)(s - s_0)$. Provided that s_0 is not a critical point of h, i.e., provided the slope of the tangent line at s_0 is non-zero, then the tangent line will intersect the s-axis. Denote this point of intersection as s_1, and observe that

$$
\begin{aligned}
& & h(s_1) - h(s_0) &= h'(s_0)(s_1 - s_0) \\
(h(s_1) = 0) & \implies & -h(s_0) &= s_1 h'(s_0) - s_0 h'(s_0) \\
& \implies & s_1 h'(s_0) &= s_0 h'(s_0) - h(s_0) \\
& \implies & s_1 &= s_0 - \frac{h(s_0)}{h'(s_0)}.
\end{aligned}
$$

Newton's method (also known as the Newton-Raphson method) can also be derived using a Taylor series expansion, as in section 9.1.2.1.

Algorithm 3 Newton's Method (root-finding)

Require: $s \in \mathbb{R}$
while $\|h(s)\| > \varepsilon$ **do**
  $s \leftarrow h(s)/h'(s)$
end while

Observe that Newton's method eliminates the need to first bracket the root, as is required by the bisection method. Newton's method is also fast: if s_0 is sufficiently close to s^*, then $|s_{i+1} - s^*| < c\,|s_i - s^*|^2$, i.e., Newton's method exhibits quadratic convergence. (By comparison, the bisection method exhibits linear convergence.) Newton's method presents certain challenges: it is not guaranteed to converge (it can "chase" roots to infinity); if h has multiple roots, there is no control over which root is found; and the derivative h' is required.

Algorithm 4 Brent's method

Require: $\{s_L, s_R : h(s_L)\, h(s_R) < 0\}$
while $|s_L - s_R| > \varepsilon$ **do**
  $s_M \leftarrow (s_L + s_R)/2$
  $\tilde{s}_M \leftarrow s_M + h(s_M)/h'(s_M)$
  if $\tilde{s}_M \in [s_L, s_R]$ **then**
    if $h(s_L)\, h(\tilde{s}_M) < 0$ **then**
      $s_R \leftarrow \tilde{s}_M$
    else
      $s_L \leftarrow \tilde{s}_M$
    end if
  else
    if $h(s_L)\, h(s_M) < 0$ **then**
      $s_R \leftarrow s_M$
    else
      $s_L \leftarrow s_M$
    end if
  end if
end while

16.2.1.3. *Hybrid methods.* Additional methods have been developed to combine the best attributes of both the bisection method and Newton's method while avoiding their respective difficulties. Given an initial bracket $[s_L, s_R]$ such that $s^* \in [s_L, s_R]$, *Brent's method* attempts to use Newton's method with the starting point taken as the midpoint of the bracket, and applies the bisection method if Newton's method fails.

Because the derivative of h may be difficult to compute, the *secant method* approximates the tangent line at s (whose slope is given by $h'(s)$) by instead constructing a secant line.

16.2.2. Newton's Method for optimization. Consider again the problem of minimizing $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ by Steepest Descent. Suppose that $f(\mathbf{x}) = 10^5 x_1^2 + x_2^2$, whose level curves are shown in figure 16.2.1. The minimum of f is easily seen to occur at $\mathbf{x}^\star = (0, 0)$. Thus, the "best" direction given some starting point $\mathbf{x}^{(0)}$ is one that points toward the origin. Suppose that we begin the optimization with $\mathbf{x}^{(0)} = (0.01, -1)$. The best direction is depicted as a dashed blue arrow. Noting that $\nabla f\left(\mathbf{x}^{(0)}\right) = \left(2 \cdot 10^3, -2\right)$, we see that the (unnormalized) steepest descent direction $\mathbf{d}^{(0)} = \left(-2 \cdot 10^3, 2\right)$, depicted as a red arrow, is quite poor.

```
f <- function(x, y) 10^5 * x^2 + y^2
x <- seq(-1e-2, 1e-2, 1e-4)
y <- seq(-1, 1, 1e-2)
z <- outer(x, y, f)
par(mgp = c(1.5,0.5,0), mar = c(2.5,2,0.5,1))
contour(x = x, y = y, z = z, levels = c(0.01, 0.1, 0.25, 0.5, 1, 2.5, 5, 10),
        drawlabels = F, cex.lab = 0.75, cex.axis = 0.75)
x0 <- c(1e-2, -1)
points(x0[1], x0[2])
grad.f <- function(x, y) c(2 * 10^5 * x, 2 * y)
grad.x0 <- grad.f(x0[1], x0[2])
x1 <- x0 + 5e-6 * -grad.x0
arrows(x0 = x0[1], y0 = x0[2], x1 = 0, y1 = 0, col = "blue", lty = 2,
       length = 0.1)
arrows(x0 = x0[1], y0 = x0[2], x1 = x1[1], y1 = x1[2], col = "red",
       length = 0.1)
```

The Steepest Descent algorithm converges slowly in part because it gives poor descent directions, and in part because finding the "optimal" step length s requires finding the root of $g'(s)$, which is also slow. We would prefer to choose step size more quickly, which we can do via *backtracking*. At each iteration, we will

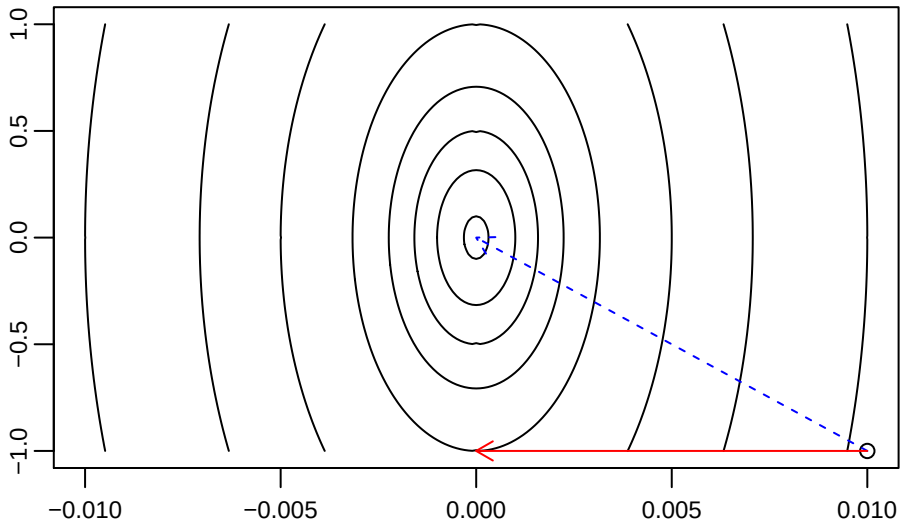


FIGURE 16.2.1. Level curves of $f(\mathbf{x}) = 10^5 x_1^2 + x_2^2$

set the initial step length $\tilde{s}$ to some user-chosen constant k. If $f(\mathbf{x} + \tilde{s}\mathbf{d}) < f(\mathbf{x})$, i.e., if taking a step from $\mathbf{x}$ in the direction of $\mathbf{d}$ and of length $\tilde{s}$ decreases value of f, then set $s = \tilde{s}$. If instead $f(\mathbf{x} + \tilde{s}\mathbf{d}) \geq f(\mathbf{x})$, i.e., we did not descend, then divide $\tilde{s}$ by 2 and check whether this produces a descent step. Repeat this procedure until $\tilde{s}$ is such that a descent step is obtained. Although it is possible to backtrack too little or too much, in general backtracking outperforms the bisection method.

Having improved our choice of step length, we now turn our attention to improving our choice of direction. The general idea is to provide a *replacement function* $r(\mathbf{x})$ that is easy to optimize. A first-order Taylor series expansion for f around the base point $\mathbf{x}^{(0)}$ is $f(\mathbf{x}) \approx f\left(\mathbf{x}^{(0)}\right) + \nabla f\left(\mathbf{x}^{(0)}\right) \cdot \left(\mathbf{x} - \mathbf{x}^{(0)}\right)$. Suppose that $\mathbf{x}^{(0)} = (0, 3)$, and that f is such that $f\left(\mathbf{x}^{(0)}\right) = 6$ and $\nabla f\left(\mathbf{x}^{(0)}\right) = (1, 1)$. Then,

$$f(\mathbf{x}) \approx 6 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 0 \\ 3 \end{bmatrix} \right) = 6 + \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 3 \end{bmatrix} = 6 + x_1 + x_2 - 3,$$

which is the equation of a plane, hence has neither a maximum nor a minimum. To make the problem tractable, we can add a term to produce the second-order Taylor series expansion

$$f(\mathbf{x}) \approx f\left(\mathbf{x}^{(0)}\right) + \nabla f\left(\mathbf{x}^{(0)}\right) \cdot \left(\mathbf{x} - \mathbf{x}^{(0)}\right) + \frac{1}{2}\left(\mathbf{x} - \mathbf{x}^{(0)}\right)^{\mathsf{T}} \mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right),$$

where $\mathbf{H}_f$ is the Hessian of f. Define the replacement function

$$r(\mathbf{x}) = \frac{1}{2}\left(\mathbf{x} - \mathbf{x}^{(0)}\right)^{\mathsf{T}} \mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right) + \nabla f\left(\mathbf{x}^{(0)}\right) \cdot \left(\mathbf{x} - \mathbf{x}^{(0)}\right) + f\left(\mathbf{x}^{(0)}\right)$$

and observe that $\mathbf{H}_f$ is an $n \times n$ matrix, and that ∇f is an n-dimensional vector. Setting

$$\tilde{\mathbf{x}} = \mathbf{x} - \mathbf{x}^{(0)}, \quad \mathbf{M} = \frac{1}{2}\mathbf{H}_f\left(\mathbf{x}^{(0)}\right), \quad \mathbf{b} = \nabla f\left(\mathbf{x}^{(0)}\right), \quad \text{and} \quad c = f\left(\mathbf{x}^{(0)}\right),$$

we have

$$r(\tilde{\mathbf{x}}) = \tilde{\mathbf{x}}^{\mathsf{T}}\mathbf{M}\tilde{\mathbf{x}} + \mathbf{b}^{\mathsf{T}}\tilde{\mathbf{x}} + c,$$

which we recognize as a general quadratic. Assuming that f has continuous second derivatives, i.e.,

$$\frac{\partial^2}{\partial x_i \partial x_j} f(\mathbf{x}) = \frac{\partial^2}{\partial x_j \partial x_i},$$

so that $\mathbf{H}_f$ is symmetric, it follows from proposition 16.4 that the gradient of r is given by $\nabla r(\tilde{\mathbf{x}}) = 2\mathbf{M}\tilde{\mathbf{x}}+\mathbf{b}$. Then, (16.1.1) implies that the critical point of r occurs at

$$\begin{aligned}
\tilde{\mathbf{x}} &= -\frac{1}{2}\mathbf{M}^{-1}\mathbf{b} \\
&= -\frac{1}{2}\left[\frac{1}{2}\mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\right]^{-1}\nabla f\left(\mathbf{x}^{(0)}\right) \\
((c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}) \qquad &= -\frac{1}{2}\left(2^{-1}\right)^{-1}\left[\mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\right]^{-1}\nabla f\left(\mathbf{x}^{(0)}\right) \\
&= -\left[\mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\right]^{-1}\nabla f\left(\mathbf{x}^{(0)}\right) \\
\implies \mathbf{x}-\mathbf{x}^{(0)} &= -\left[\mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\right]^{-1}\nabla f\left(\mathbf{x}^{(0)}\right) \\
\implies \mathbf{x} &= \mathbf{x}^{(0)} - \left[\mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\right]^{-1}\nabla f\left(\mathbf{x}^{(0)}\right).
\end{aligned}$$

Observe that for $x \in \mathbb{R}$, this provides the update

$$x^{(i+1)} = x^{(i)} - \left[f''\left(x^{(i)}\right)\right] f'\left(x^{(i)}\right) = x^{(i)} - \frac{f'\left(x^{(i)}\right)}{f''\left(x^{(i)}\right)},$$

precisely as obtained in (9.1.2). Thus, we see that Newton's method can be applied to optimization.

Algorithm 5 Newton's method (optimization)

Require: $\mathbf{x}$
1: **while** $\|\nabla f(\mathbf{x})\| > \varepsilon$ **do**
2: $\quad \mathbf{x} \leftarrow \mathbf{x} - [\mathbf{H}_f(\mathbf{x})]^{-1}\nabla f(\mathbf{x})$
3: **end while**

We see that there is no need in this case to choose a direction or step length. Although Newton's method may fail to converge altogther (as when used for root-finding), when it does converge, it converges rapidly. It is also easy to implement. In addition to possible non-convergence, Newton's method can converge to the "wrong" critical point (analogous to finding a different root than the one desired). Further, the Hessian may not be invertible, and even when it is, calculating $[\mathbf{H}_f(\mathbf{x})]^{-1}\nabla f(\mathbf{x})$ can be time-consuming, especially in higher dimensions, and may even break down for sufficiently large n (due to the requirement that the Hessian be inverted).

EXAMPLE 16.6. Let $f(\mathbf{x}) = 3x_1^2 + 5x_2^2 + x_1x_2 + x_1 + 10x_2$. Find $\mathbf{x}^*$ such that $\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x})$.
We have

$$\nabla f(\mathbf{x}) = \begin{bmatrix} 6x_1 + x_2 + 1 \\ 10x_2 + x_1 + 10 \end{bmatrix} \quad \text{and} \quad \mathbf{H}_f(\mathbf{x}) = \begin{bmatrix} 6 & 1 \\ 1 & 10 \end{bmatrix}.$$

Choose $\mathbf{x}^{(0)} = (2, 4)$, so that $\nabla f\left(\mathbf{x}^{(0)}\right) = (17, 52)$. Then, the Newton's method update is

$$\begin{aligned}
\mathbf{x}^{(1)} &= \mathbf{x}^{(0)} - \left[\mathbf{H}_f\left(\mathbf{x}^{(0)}\right)\right]^{-1}\nabla f\left(\mathbf{x}^{(0)}\right) \\
&= \begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 6 & 1 \\ 1 & 10 \end{bmatrix}^{-1}\begin{bmatrix} 17 \\ 52 \end{bmatrix} \\
&= \begin{bmatrix} 2 \\ 4 \end{bmatrix} - \frac{1}{60-1}\begin{bmatrix} 10 & -1 \\ -1 & 6 \end{bmatrix}\begin{bmatrix} 17 \\ 52 \end{bmatrix} \\
&= \begin{bmatrix} 2 \\ 4 \end{bmatrix} - \frac{1}{59}\begin{bmatrix} 170 - 52 \\ -17 + 312 \end{bmatrix} \\
&= \begin{bmatrix} 2 \\ 4 \end{bmatrix} - \begin{bmatrix} 118/59 \\ 295/59 \end{bmatrix} \\
&= \begin{bmatrix} 118/59 \\ 236/59 \end{bmatrix} - \begin{bmatrix} 118/59 \\ 295/59 \end{bmatrix}
\end{aligned}$$

$$= \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

Thus, $\mathbf{x}^* = (0, -1)$ minimizes f. Observe that Newton's Method finds the critical point in a single step because f is quadratic.

DEFINITION 16.7. Suppose that $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$. An iterative optimization algorithm is a *descent method* if $f\left(\mathbf{x}^{(0)}\right) \geq f\left(\mathbf{x}^{(1)}\right) \geq f\left(\mathbf{x}^{(2)}\right) \geq \cdots$.

Descent algorithms always converge, i.e., the longer the algorithm runs, the closer it comes to the solution of the optimization problem. Observe that Steepest Descent with backtracking is a descent method, but that Steepest Descent with bisection is not. Because Newton's Method may, for example, find a maximum rather than a minimum, it is also not a descent method. To shape it into one, we must "fix" both the direction and the step length. Supposing for the moment that the direction is acceptable, we will fix the step length by *damping*.

Algorithm 6 Damped Newton's Method

Require: $\mathbf{x}$
1: **while** $\|\nabla f(\mathbf{x})\| > \varepsilon$ **do**
2: $\quad \mathbf{d} \leftarrow -\left[\mathbf{H}_f(\mathbf{x})\right]^{-1} \nabla f(\mathbf{x})$
3: $\quad s \leftarrow 1$
4: $\quad$ **while** $f(\mathbf{x}) < f(\mathbf{x} + s\mathbf{d})$ **do**
5: $\quad\quad s \leftarrow s/2$
6: $\quad$ **end while**
7: $\quad \mathbf{x} \leftarrow \mathbf{x} + s\mathbf{d}$
8: **end while**

Observe that the first attempt of algorithm 6 is simply (undamped) Newton's Method, i.e., we check whether Newton's Method satisfies $f\left(\mathbf{x}^{(i-i)}\right) \geq f\left(\mathbf{x}^{(i)}\right)$, and damp (normalize) the direction only if it does not. Now, the direction might actually be wrong, so Damped Newton's Method is still not a descent algorithm.

DEFINITION 16.8. For a function f at a point $\mathbf{x}$, the direction $\mathbf{d}$ is a *descent direction* if there exists a value $\tilde{s}$ such that for all $s < \tilde{s}$, $f(\mathbf{x} + s\mathbf{d}) < f(\mathbf{x})$.

Noting that the step length s will be "small," consider the Taylor series expansion of $f(\mathbf{x} + s\mathbf{d})$ around the base point $\mathbf{x}$

$$f(\mathbf{x} + s\mathbf{d}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (s\mathbf{d}) + \varepsilon,$$

where the error term ε will be small, on the order of $\mathcal{O}\left(s^2\right)$. Then, $\mathbf{d}$ will satisfy $f(\mathbf{x} + s\mathbf{d}) < f(\mathbf{x})$, i.e., $\mathbf{d}$ will be a descent direction, if

$$f(\mathbf{x}) > f(\mathbf{x} + s\mathbf{d}) \approx f(\mathbf{x}) + \nabla f(\mathbf{x}) \cdot (s\mathbf{d}) = f(\mathbf{x}) + s\nabla f(\mathbf{x}) \cdot \mathbf{d} \implies s\nabla f(\mathbf{x}) \cdot \mathbf{d} < 0.$$

Now, the step length s is strictly positive, so this inequality will hold if and only if $\nabla f(\mathbf{x}) \cdot \mathbf{d} < 0$, i.e., if the dot product of the gradient of f and the direction $\mathbf{d}$ is negative.

16.2.3. Quadratic optimization. Recall that an n-dimensional quadratic has the form $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{x} + c$, where $\mathbf{A}$ can be taken as symmetric. We would like to know under what circumstances f will have a maximum, a minimum, or a saddle point, i.e., the characteristics of the critical point. Suppose that $\mathbf{x} \in \mathbb{R}^2$, suppose for the moment that f has no cross-term, and let $f(\mathbf{x}) = x_1^2 + 2x_2^2$. It is clear that f is a paraboloid opening up, hence the critical point of f is a minimum. If we instead let $f(\mathbf{x}) = -x_1^2 - 2x_2^2$, then f is a paraboloid opening down, hence the critical point of f is a maximum. If we let $f(\mathbf{x}) = x_1^2 - 2x_2^2$, then the critical point of f is a saddle point. Now suppose that $f(\mathbf{x}) = x_1^2 + 6x_1x_2 + 3x_2^2$. f now has a cross-term, and it is not immediately clear how to characterize the critical point. We will see that *all* quadratics have no cross-term, given an appropriate transformation.

THEOREM 16.9 (Spectral Theorem). *Let $\mathbf{A}$ be an $n \times n$ real, symmetric matrix. Then, $\mathbf{A}$ has an orthonormal basis of eigenvectors.*

PROOF. [proof goes here] □

COROLLARY 16.10. *Suppose that* $\mathbf{A}$ *is an* $n \times n$ *real, symmetric matrix, and let* $\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \ldots, \mathbf{q}^{(n)}$ *be the eigenvectors corresponding to the ordered eigenvalues* $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$*. Let* $\mathbf{Q}$ *be the matrix whose* j*th column is given by* $\mathbf{q}^{(j)}$*, i.e.,* $\mathbf{Q} = \begin{bmatrix} \mathbf{q}^{(1)} & \mathbf{q}^{(2)} & \cdots & \mathbf{q}^{(n)} \end{bmatrix}$*, and let* $\mathbf{D}$ *be the diagonal matrix whose* i*th diagonal entry is* λ_i*. Then,* $\mathbf{A}$ *has the decomposition* $\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^\mathsf{T}$*.*

PROOF. [proof goes here] □

Let $\left\{\mathbf{e}^{(i)}\right\}_{i=1}^n$ be the standard basis vectors for $\mathbb{R}^n$, i.e., the ith component of $\mathbf{e}^{(i)}$ is 1 and all other components are zero. Then, any $\mathbf{x} \in \mathbb{R}^n$ can be represented as a linear combination of the $\mathbf{e}^{(i)}$, i.e.,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1\mathbf{e}^{(1)} + x_2\mathbf{e}^{(2)} + \cdots + x_n\mathbf{e}^{(n)} = \sum_{i=1}^n x_i\mathbf{e}^{(i)}.$$

Consider again the general quadratic, and observe that the matrix $\mathbf{A}$ determines the behavior of f. For simplicity, we consider only the initial term, i.e., $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$. Then, we can express f in terms of the $\mathbf{e}^{(i)}$, i.e.,

$$\begin{aligned}
f(\mathbf{x}) &= f\left(\sum_{i=1}^n x_i\mathbf{e}^{(i)}\right) \\
&= \left(\sum_{i=1}^n x_i\mathbf{e}^{(i)}\right)^\mathsf{T} \mathbf{A} \left(\sum_{j=1}^n x_j\mathbf{e}^{(j)}\right) \\
((\mathbf{u}+\mathbf{v})^\mathsf{T} = \mathbf{u}^\mathsf{T} + \mathbf{v}^\mathsf{T}) \quad &= \left(\sum_{i=1}^n x_i \left(\mathbf{e}^{(i)}\right)^\mathsf{T}\right) \mathbf{A} \left(\sum_{j=1}^n x_j\mathbf{e}^{(j)}\right) \\
(\mathbf{M}(\mathbf{u}+\mathbf{v}) = \mathbf{M}\mathbf{u} + \mathbf{M}\mathbf{v}) \quad &= \left(\sum_{i=1}^n x_i \left(\mathbf{e}^{(i)}\right)^\mathsf{T}\right) \left(\sum_{j=1}^n x_j\mathbf{A}\mathbf{e}^{(j)}\right) \\
&= \sum_{i=1}^n x_i \sum_{j=1}^n x_j \left(\mathbf{e}^{(i)}\right)^\mathsf{T} \mathbf{A}\mathbf{e}^{(j)} \\
&= \sum_{i=1}^n \sum_{j=1}^n x_i x_j \left(\mathbf{e}^{(i)}\right)^\mathsf{T} \mathbf{A}\mathbf{e}^{(j)}.
\end{aligned}$$

Now, right-multiplying $\mathbf{A}$ by $\mathbf{e}^{(j)}$ will extract the jth column of $\mathbf{A}$, so that each product $\mathbf{A}\mathbf{e}^{(j)}$ will be jth column of $\mathbf{A}$, i.e.,

$$\mathbf{A}\mathbf{e}^{(j)} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix}.$$

Then, left-multiplying $\mathbf{A}\mathbf{e}^{(j)}$ by $\left(\mathbf{e}^{(i)}\right)^\mathsf{T}$ will extract the ith component of $\mathbf{A}\mathbf{e}^{(j)}$, so that each product $\left(\mathbf{e}^{(i)}\right)^\mathsf{T}\mathbf{A}\mathbf{e}^{(j)}$ will be the (i,j) th component of $\mathbf{A}$, i.e.,

$$\left(\mathbf{e}^{(i)}\right)^\mathsf{T} \mathbf{A}\mathbf{e}^{(j)} = \begin{bmatrix} 0 & \cdots & 1 & \cdots & 0 \end{bmatrix} \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix} = a_{ij}.$$

Then, our expression for $f(\mathbf{x})$ becomes

$$f(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j \left(\mathbf{e}^{(i)}\right)^{\mathsf{T}} \mathbf{A}\mathbf{e}^{(j)} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j a_{ij}.$$

EXAMPLE 16.11. We now consider when cross-terms appear in f. Suppose that $\mathbf{x} \in \mathbb{R}^2$, and let

$$\mathbf{A} = \begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} \implies f(\mathbf{x}) = \sum_{i=1}^{2}\sum_{j=1}^{2} x_i x_j a_{ij} = x_1^2 \cdot 1 + x_1 x_2 \cdot 3 + x_2 x_1 \cdot 3 + x_2^2 \cdot 2 = x_1^2 + 6x_1 x_2 + 2x_2^2.$$

We see that f contains the cross-term $6x_1x_2$. It is clear that f would not have a cross-term if $\mathbf{A}$ were diagonal. Now, $\mathbf{A}$ is an $n \times n$ real, symmetric matrix, so theorem 16.9 implies that $\mathbf{A}$ has an orthonormal basis of eigenvectors, which we denote by $\left\{\mathbf{q}^{(i)}\right\}_{i=1}^{n}$. Because the $\mathbf{q}^{(i)}$ form a basis for $\mathbb{R}^n$, it follows that we can express $\mathbf{x}$ in terms of the $\mathbf{q}^{(i)}$ with corresponding weights $\{w_i\}_{i=1}^{n}$, i.e.,

$$\mathbf{x} = w_1\mathbf{q}^{(1)} + w_2\mathbf{q}^{(2)} + \cdots + w_n\mathbf{q}^{(n)} = \sum_{i=1}^{n} w_i \mathbf{q}^{(i)}.$$

Then, our result above implies that

$$f(\mathbf{x}) = f\left(\sum_{i=1}^{n} w_i \mathbf{q}^{(i)}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left(\mathbf{q}^{(i)}\right)^{\mathsf{T}} \mathbf{A}\mathbf{q}^{(j)}.$$

Now, $\mathbf{q}^{(i)}$ is an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ_i, so that $\mathbf{A}\mathbf{q}^{(j)} = \lambda_j \mathbf{q}^{(j)}$, hence

$$f(\mathbf{x}) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \left(\mathbf{q}^{(i)}\right)^{\mathsf{T}} \lambda_j \mathbf{q}^{(j)} = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \lambda_j \left(\mathbf{q}^{(i)}\right)^{\mathsf{T}} \mathbf{q}^{(j)}.$$

Because the $\mathbf{q}^{(i)}$ are orthonormal, we will have $\left(\mathbf{q}^{(i)}\right)^{\mathsf{T}} \mathbf{q}^{(j)} = 1$ in the case that $i = j$ and zero otherwise. Thus,

$$f(\mathbf{x}) = w_i \cdot w_i \lambda_i \cdot 1 = w_i^2 \lambda_i.$$

The characteristic polynomial of $\mathbf{A}$ is

$$\begin{aligned} \det(\mathbf{A} - \lambda \mathbf{I}_2) &= \det\left(\begin{bmatrix} 1 & 3 \\ 3 & 2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} 1-\lambda & 3 \\ 3 & 2-\lambda \end{bmatrix}\right) \\ &= (1-\lambda)(2-\lambda) - 9 \\ &= 2 - \lambda - 2\lambda + \lambda^2 - 9 \\ &= \lambda^2 - 3\lambda - 7. \end{aligned}$$

Solving for λ gives

$$\lambda = \frac{3 \pm \sqrt{9 - 4(-7)}}{2} = \frac{3 \pm \sqrt{37}}{2} \implies \lambda_1 = \frac{3+\sqrt{37}}{2} \approx 4.5413813,\ \lambda_2 = \frac{3-\sqrt{37}}{2} \approx -1.5413813,$$

so that expressing $\mathbf{x}$ in the $\mathbf{q}$-basis gives

$$f(\mathbf{x}) = x_1^2 + 6x_1x_2 + 2x_2^2 \iff f(\mathbf{w}) = 4.5w_1^2 - 1.5w_2^2.$$

We see that when we express $\mathbf{x}$ in the $\mathbf{q}$-basis, our expression for $f(\mathbf{w})$ does not contain a cross-term. Figure 16.2.2 shows the standard and $\mathbf{q}$ bases for $f(\mathbf{x})$.

```
lambda <- eigen(matrix(c(1,3,3,2), nr = 2))$vectors
par(mgp = c(1.5,0.5,0), mar = c(0.5,1,0.5,1))
plot.new()
plot.window(xlim = c(-1, 1.1), ylim = c(0, 1.2), asp = 1)
arrows(x0 = 0, y0 = 0, x1 = 1, y1 = 0, length = 0.1, col = "black")
arrows(x0 = 0, y0 = 0, x1 = 0, y1 = 1, length = 0.1, col = "black")
arrows(x0 = 0, y0 = 0, x1 = lambda[1,1], y1 = lambda[2,1], length = 0.1,
```

```
       lty = 2, col = "blue")
arrows(x0 = 0, y0 = 0, x1 = lambda[1,2], y1 = lambda[2,2], length = 0.1,
       lty = 2, col = "blue")
text(x = 1.05, y = 0.05, cex = 0.75, labels = expression(e^(1)))
text(x = 0.05, y = 1.05, cex = 0.75, labels = expression(e^(2)))
text(x = lambda[1,1] + 0.05, y = lambda[2,1] + 0.05, cex = 0.75,
     labels = expression(q^(1)), col = "blue")
text(x = lambda[1,2] - 0.05, y = lambda[2,2] + 0.05, cex = 0.75,
     labels = expression(q^(2)), col = "blue")
```

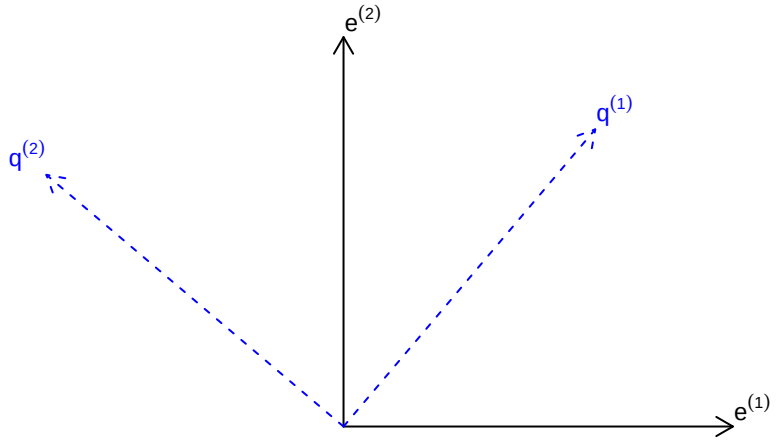


FIGURE 16.2.2. Standard and $\mathbf{q}$ bases for $f(\mathbf{x})$

We can apply the same idea to other quadratics, e.g., the regression least-squares loss function, where $\mathbf{A} = \mathbf{B}^\mathsf{T}\mathbf{B}$ for some design matrix $\mathbf{B}$. Computing the eigenvalues of $\mathbf{A}$ also provides information about the nature of the critical point, which we state in the following theorem.

THEOREM 16.12. *Let f be an n-dimensional quadratic, i.e., $f(\mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} + \mathbf{b}^\mathsf{T}\mathbf{x} + c$, and let $\{\lambda_i\}_{i=1}^n$ be the eigenvalues of $\mathbf{A}$. Then, the critical point of f is*

1. *a maximum if all the eigenvalues of $\mathbf{A}$ are negative, i.e., if $\{\lambda_i : \lambda_i < 0\ \forall i\}$;*
2. *a minimum if all the eigenvalues of $\mathbf{A}$ are positive, i.e., if $\{\lambda_i : \lambda_i > 0\ \forall i\}$;*
3. *a saddle point if the eigenvalues of $\mathbf{A}$ are of mixed sign.*

PROOF. [proof goes here] □

DEFINITION 16.13. An $n \times n$ matrix $\mathbf{A}$ is *positive definite* if $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} > 0$ for all $\{\mathbf{x} : \mathbf{x} \in \mathbb{R}^n, \mathbf{x} \neq \mathbf{0}\}$.

If $\mathbf{A}$ is positive definite, then for every nonzero $\mathbf{x}$, the quantity $\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x}$ will be positive, which implies that the critical point of $\mathbf{A}$ is a minimum, hence all the eigenvalues of $\mathbf{A}$ are positive.

PROPOSITION 16.14. *An $n \times n$ matrix $\mathbf{A}$ is positive definite if and only if $\mathbf{A}$ has all positive eigenvalues.*

PROOF. [proof goes here] □

Now consider the action of $\mathbf{A}$ on a nonzero vector $\mathbf{x}$. If $\mathbf{A}$ is positive definite, then $0 < \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = \mathbf{x} \cdot \mathbf{A}\mathbf{x}$. $\mathbf{x}$ is nonzero, so the dot product of $\mathbf{x}$ and $\mathbf{A}\mathbf{x}$ will be positive only in the case that $\mathbf{A}$ does not "flip" the direction of $\mathbf{x}$.

Recall that the inner while loop in algorithm 6 will exit only when $f(\mathbf{x}+s\mathbf{d})$ is less than or equal to $f(\mathbf{x})$, i.e., when the step length s results in "downhill" (or at least "flat") movement. This will occur only in the case that $\mathbf{d}$ is a descent direction, so it becomes natural to ask which choices of $\mathbf{d}$ will provide descent directions. Recall that the dot product of two vectors $\mathbf{u}$ and $\mathbf{v}$ can be written as $\mathbf{u}\cdot\mathbf{v} = \|\mathbf{u}\|\,\|\mathbf{v}\|\cos\theta$, where θ is the angle formed by $\mathbf{u}$ and $\mathbf{v}$. Noting that the gradient of f evaluated at some point $\tilde{\mathbf{x}}$ points in the direction of steepest increase of f at $\tilde{\mathbf{x}}$, it follows that a descent direction will be a direction "opposite" to $\nabla f(\tilde{\mathbf{x}})$, i.e., a direction for which $\mathbf{d}\cdot\nabla f(\tilde{\mathbf{x}}) < 0$, which will occur when $\theta \in (\pi/2, 3\pi/2)$.

Now, Damped Newton's Method defines the direction at a point $\tilde{\mathbf{x}}$ to be $\mathbf{d} = -\left[\mathbf{H}_f(\tilde{\mathbf{x}})\right]^{-1}\nabla f(\tilde{\mathbf{x}})$. The Hessian of a general quadratic f can be shown to be equal (possibly up to a constant, depending on the exact form of f) to the matrix $\mathbf{A}$, where $\mathbf{A}$ does not depend on $\mathbf{x}$. Thus, $\mathbf{d} = -\mathbf{A}^{-1}\nabla f(\tilde{\mathbf{x}})$, which implies that $\mathbf{d}$ will be a descent direction if

$$\begin{aligned}
& 0 > \left(-\mathbf{A}^{-1}\nabla f(\tilde{\mathbf{x}})\right)\cdot\nabla f(\tilde{\mathbf{x}}) \\
\implies\ & 0 < \mathbf{A}^{-1}\nabla f(\tilde{\mathbf{x}})\cdot\nabla f(\tilde{\mathbf{x}}) \\
& = \left[\mathbf{A}^{-1}\nabla f(\tilde{\mathbf{x}})\right]^{\mathsf{T}}\nabla f(\tilde{\mathbf{x}}) \\
& = \left[\nabla f(\tilde{\mathbf{x}})\right]^{\mathsf{T}}\left(\mathbf{A}^{-1}\right)^{\mathsf{T}}\nabla f(\tilde{\mathbf{x}}) \\
\text{($\mathbf{A}$ is symmetric)}\quad & = \left[\nabla f(\tilde{\mathbf{x}})\right]^{\mathsf{T}}\left(\mathbf{A}^{\mathsf{T}}\right)^{\mathsf{T}}\nabla f(\tilde{\mathbf{x}}) \\
& = \left[\nabla f(\tilde{\mathbf{x}})\right]^{\mathsf{T}}\mathbf{A}\nabla f(\tilde{\mathbf{x}}).
\end{aligned}$$

The quantity $\left[\nabla f(\tilde{\mathbf{x}})\right]^{\mathsf{T}}\mathbf{A}\nabla f(\tilde{\mathbf{x}})$ will be positive for all $\nabla f(\tilde{\mathbf{x}}) \neq \mathbf{0}$ in the case that $\mathbf{A}$ is positive definite, and in this case $\mathbf{d}$ will be a descent direction.

PROPOSITION 16.15. *Let f be a twice-differentiable function of $\mathbf{x}\in\mathbb{R}^n$ with Hessian $\mathbf{H}_f(\mathbf{x})$, and let $\mathbf{d}$ be the Damped Newton's Method direction at a point $\tilde{\mathbf{x}}$. Then,*

1. *if $\left[\mathbf{H}_f(\tilde{\mathbf{x}})\right]^{-1}$ is positive definite, then $\mathbf{d}$ is a descent direction.*
2. *if $\mathbf{H}_f(\tilde{\mathbf{x}})$ is positive definite, then $\mathbf{d}$ is a descent direction.*

We are now ready to state a result that extends Newton's Method in the case that the Hessian of the function f is positive definite.

THEOREM 16.16. *Let f be a twice-differentiable function of $\mathbf{x}\in\mathbb{R}^n$ with Hessian $\mathbf{H}_f(\mathbf{x})$. If $\mathbf{H}_f(\mathbf{x})$ is positive definite for all $\mathbf{x}\neq\mathbf{0}$, then Newton's Method always gives descent directions.*

PROOF. [proof goes here] □

Thus, if the Hessian of f is positive definite, then Damped Newton's Method is a descent algorithm.

16.2.4. Convexity.

DEFINITION 16.17. Suppose that $f:\mathbb{R}^n\to\mathbb{R}$, and let $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$ and $\lambda\in[0,1]$. Then,

1. f is *convex* if $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$;
2. f is *concave* if $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \geq \lambda f(\mathbf{x}) + (1-\lambda)f(\mathbf{y})$.

Geometrically, if the function f is plotted, and a line is drawn between the points $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$, f will be convex if the line is always "above" the graph of f.

EXAMPLE 16.18. Concretely, suppose that $x\in\mathbb{R}$, and let $f(x) = x^2$. Suppose that $x_1 = -0.5$ and $x_2 = 1$, so that $f(x_1) = 0.25$ and $f(x_2) = 1$. Figure 16.2.3 shows f and the points $(x_1, f(x_1))$ and $(x_2, f(x_2))$. Observe that the (dashed) line connecting these points lies above the graph of f. It is easy to see that this will be true for any choices of x_1 and x_2.

```
par(mgp = c(1.5,0.5,0), mar = c(2.5,2.5,0.5,1))
curve(x^2, xlim = c(-2, 2), ylab = "f(x)", cex.lab = 0.75,
      xaxt = "n", yaxt = "n")
x1 <- c(-0.5, 0.25)
```

```
x2 <- c(1, 1)
axis(side = 1, at = c(x1[1], x2[1]), cex.axis = 0.75,
     labels = c(expression(x[1]), expression(x[2])))
points(x = c(-0.5, 1), y = c(0.25, 1))
segments(x0 = -0.5, y0 = 0.25, x1 = 1, y1 = 1, lty = 2)
```

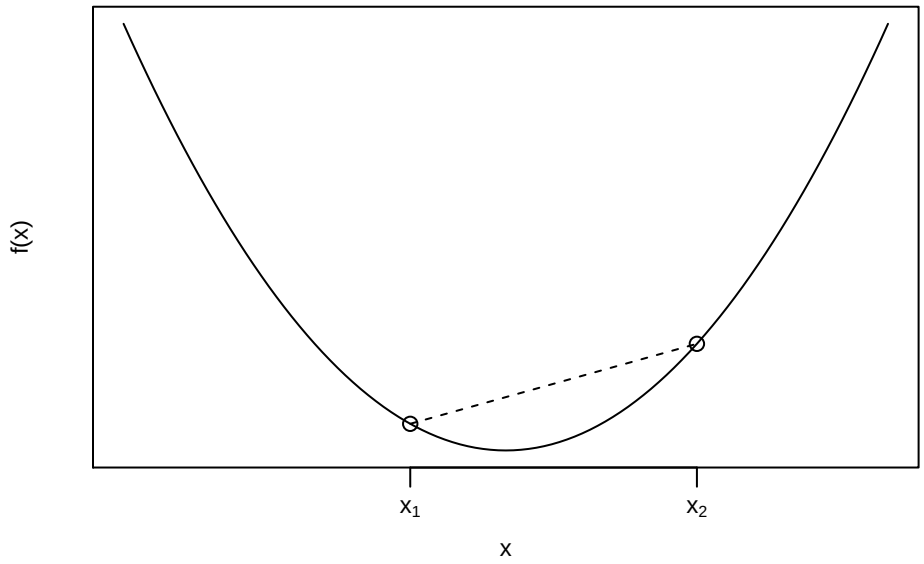


FIGURE 16.2.3. $f(x)$ is convex

We also have

$$\begin{aligned} f(\lambda x_1 + (1-\lambda) x_2) &= (\lambda x_1 + (1-\lambda) x_2)^2 \\ &= \lambda^2 x_1^2 + 2(\lambda x_1)(1-\lambda) x_2 + (1-\lambda)^2 x_2^2 \\ &= \lambda^2 x_1^2 + 2\lambda x_1 (x_2 - \lambda x_2) + \left(1 - 2\lambda + \lambda^2\right) x_2^2 \\ &= \lambda^2 x_1^2 + 2\lambda x_1 x_2 - 2\lambda^2 x_1 x_2 + x_2^2 - 2\lambda x_2^2 + \lambda^2 x_2^2 \end{aligned}$$

and

$$\lambda f(x_1) + (1-\lambda) f(x_2) = \lambda x_1^2 + (1-\lambda) x_2^2 = \lambda x_1^2 + x_2^2 - \lambda x_2^2.$$

Then, f will be convex if

$$\lambda^2 x_1^2 + 2\lambda x_1 x_2 - 2\lambda^2 x_1 x_2 + x_2^2 - 2\lambda x_2^2 + \lambda^2 x_2^2 \leq \lambda x_1^2 + x_2^2 - \lambda x_2^2$$

for all $x_1, x_2 \in \mathbb{R}$ and all $\lambda \in [0,1]$. Suppose that the inequality holds, so that

$$\begin{aligned} 0 &\leq \lambda x_1^2 + x_2^2 - \lambda x_2^2 - \lambda^2 x_1^2 - 2\lambda x_1 x_2 + 2\lambda^2 x_1 x_2 - x_2^2 + 2\lambda x_2^2 - \lambda^2 x_2^2 \\ &= (1-\lambda)\lambda x_1^2 + \lambda x_2^2 - 2(1-\lambda)\lambda x_1 x_2 - \lambda^2 x_2^2 \\ &= (1-\lambda)\lambda x_1^2 + (1-\lambda)\lambda x_2^2 - 2(1-\lambda)\lambda x_1 x_2 \\ &= \lambda(1-\lambda)\left(x_1^2 + x_2^2 - 2x_1 x_2\right) \\ &= \lambda(1-\lambda)(x_1 - x_2)^2. \end{aligned}$$

We have $\lambda \in [0,1]$, so that both λ and $1-\lambda$ will be nonnegative. Then, for any real numbers x_1 and x_2, the quantity $(x_1 - x_2)^2$ will be nonnegative. It follows that the product $\lambda(1-\lambda)(x_1 - x_2)^2$ will be nonnegative for all $x_1, x_2 \in \mathbb{R}$ and $\lambda \in [0,1]$, so that the inequality holds, and it follows that f is convex.

THEOREM 16.19. *Let f be a twice-differentiable function of $\mathbf{x} \in \mathbb{R}^n$ with Hessian $\mathbf{H}_f(\mathbf{x})$. Then, f is convex if and only if $\mathbf{H}_f(\mathbf{x})$ is positive definite.*

PROOF. [proof goes here] □

Observe that if $x \in \mathbb{R}$, the equivalent condition for convexity is that $f''(x) > 0$.

THEOREM 16.20.

(1) *Damped Newton's Method applied to a convex function is a descent algorithm.*
(2) *If a function $f : \mathbb{R}^n \to \mathbb{R}$ has a minimum, then Damped Newton's Method with converge to the minimum.*

PROOF. [proof goes here] □

We now state some important facts.

PROPOSITION 16.21. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$.*

(1) *If f is convex, then $-f$ is concave.*
(2) *If f and g are convex, then $c_1 f(\mathbf{x}) + c_2 g(\mathbf{x})$ is convex if $c_1, c_2 > 0$.*
(3) *A linear function is both concave and convex.*

PROOF. [proof goes here] □

In the absence of constraints, convex optimization is a "solved" problem, whereas non-convex optimization remains a "hard" problem.

16.3. Logistic regression

We will see that *logistic regression* is not quadratic, but is nevertheless a convex problem. As in example 7.1, suppose that we collect data (Y_i, x_i) for $i \in \{1, 2, \ldots, n\}$, where $x_i \in \mathbb{R}$ and $Y_i \in \{0, 1\}$. We would like to fit a *sigmoid curve* to the data. Logistic regression assumes the model

$$P(Y = 1|x, \alpha_0, \alpha_1) = \frac{1}{1 + \exp(\alpha_0 + \alpha_1 x)}.$$

We now show that fitting the logistic regression corresponds to a convex optimization problem. Either $Y_i = 0$ or $Y_i = 1$, so that the pmf of Y_i can be written as

$$P(Y_i = y_i|x_i, \boldsymbol{\alpha}) = [P(Y_i = 1|x_i, \boldsymbol{\alpha})]^{y_i} [P(Y_i = 0|x_i, \boldsymbol{\alpha})]^{1-y_i}.$$

We assume the Y_i's are a random sample, so that they are iid. Then, the log-likelihood is given by

$$\begin{aligned}
\ell(\boldsymbol{\alpha}|\mathbf{x}, \mathbf{y}) &= \log \prod_{i=1}^{N} P(Y_i = y_i|x_i, \boldsymbol{\alpha}) \\
&= \sum_{i=1}^{N} \log [P(Y_i = 1|x_i, \boldsymbol{\alpha})]^{y_i} [P(Y_i = 0|x_i, \boldsymbol{\alpha})]^{1-y_i} \\
&= \sum_{i=1}^{N} \left[\log [P(Y_i = 1|x_i, \boldsymbol{\alpha})]^{y_i} + \log [P(Y_i = 0|x_i, \boldsymbol{\alpha})]^{1-y_i} \right] \\
&= \sum_{i=1}^{N} [y_i \log P(Y_i = 1|x_i, \boldsymbol{\alpha}) + (1 - y_i) \log P(Y_i = 0|x_i, \boldsymbol{\alpha})] \\
&= \sum_{i=1}^{N} \left[y_i \log \frac{1}{1 + \mathrm{e}^{\alpha_0 + \alpha_1 x_i}} + (1 - y_i) \log [1 - P(Y_i = 1|x_i \boldsymbol{\alpha})] \right] \\
&= \sum_{i=1}^{N} \left[y_i \left(\log 1 - \log \left(1 + \mathrm{e}^{\alpha_0 + \alpha_1 x_i} \right) \right) + (1 - y_i) \log \left(1 - \frac{1}{1 + \mathrm{e}^{\alpha_0 + \alpha_1 x_i}} \right) \right] \\
&= \sum_{i=1}^{N} \left[y_i \left(0 - \log \left(1 + \mathrm{e}^{\alpha_0 + \alpha_1 x_i} \right) \right) + (1 - y_i) \log \left(\frac{1 + \mathrm{e}^{\alpha_0 + \alpha_1 x_i} - 1}{1 + \mathrm{e}^{\alpha_0 + \alpha_1 x_i}} \right) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \left[-y_i \log\left(1 + \mathrm{e}^{\alpha_0+\alpha_1 x_i}\right) + (1-y_i)\left[\log \mathrm{e}^{\alpha_0+\alpha_1 x_i} - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right]\right] \\
&= \sum_{i=1}^{N} \left[-y_i \log\left(1 + \mathrm{e}^{\alpha_0+\alpha_1 x_i}\right) + (1-y_i)\left[\alpha_0+\alpha_1 x_i - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right]\right] \\
&= \sum_{i=1}^{N} \left[-y_i \log\left(1 + \mathrm{e}^{\alpha_0+\alpha_1 x_i}\right) + \alpha_0+\alpha_1 x_i - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right. \\
&\qquad \left. -y_i\left(\alpha_0+\alpha_1 x_i\right) + y_i \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right] \\
&= \sum_{i=1}^{N} \left[(1-y_i)\left(\alpha_0+\alpha_1 x_i\right) - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right].
\end{aligned}
$$

Taking the derivative with respect to α_0 gives

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_0}\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) &= \frac{\partial}{\partial \alpha_0}\sum_{i=1}^{N}\left[(1-y_i)\left(\alpha_0+\alpha_1 x_i\right) - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right] \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial \alpha_0}\left[(1-y_i)\left(\alpha_0+\alpha_1 x_i\right) - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right] \\
&= \sum_{i=1}^{N}\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right).
\end{aligned}
$$

Taking the derivative with respect to α_1 gives

$$
\begin{aligned}
\frac{\partial}{\partial \alpha_1}\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) &= \frac{\partial}{\partial \alpha_1}\sum_{i=1}^{N}\left[(1-y_i)\left(\alpha_0+\alpha_1 x_i\right) - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right] \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial \alpha_1}\left[(1-y_i)\left(\alpha_0+\alpha_1 x_i\right) - \log\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\right] \\
&= \sum_{i=1}^{N}\left[(1-y_i)\,x_i - \frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right] \\
&= \sum_{i=1}^{N} x_i\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right).
\end{aligned}
$$

Then,

$$
\begin{aligned}
\nabla\ell\left(\boldsymbol{\alpha}\right) &= \begin{bmatrix} \sum_{i=1}^{N}\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\ \sum_{i=1}^{N} x_i\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right)\end{bmatrix} \\
&= \sum_{i=1}^{N}\begin{bmatrix} 1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}} \\ x_i\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right)\end{bmatrix} \\
&= \sum_{i=1}^{N}\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right)\begin{bmatrix}1 \\ x_i\end{bmatrix}.
\end{aligned}
$$

We now evaluate the second partial derivatives of the log-likelihood.

$$
\begin{aligned}
\frac{\partial^2}{\partial \alpha_0^2}\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) &= \frac{\partial}{\partial \alpha_0}\sum_{i=1}^{N}\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N}\frac{\partial}{\partial \alpha_0}\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \left[0 - 0 - \frac{\partial}{\partial \alpha_0} \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}} \right] \\
&= \sum_{i=1}^{N} - \frac{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\mathrm{e}^{\alpha_0+\alpha_1 x_i} \cdot 1 - \mathrm{e}^{\alpha_0+\alpha_1 x_i}\left(\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\cdot 1}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\
&= \sum_{i=1}^{N} - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i} + \left(\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2 - \left(\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\
&= \sum_{i=1}^{N} - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\
\frac{\partial^2}{\partial \alpha_1^2}\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) &= \frac{\partial}{\partial \alpha_1} \sum_{i=1}^{N} x_i \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \alpha_1} x_i \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N} x_i \frac{\partial}{\partial \alpha_1} \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N} x_i \left[0 - 0 - \frac{\partial}{\partial \alpha_1} \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}} \right] \\
&= \sum_{i=1}^{N} - x_i \frac{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\mathrm{e}^{\alpha_0+\alpha_1 x_i} \cdot x_i - \mathrm{e}^{\alpha_0+\alpha_1 x_i}\left(\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)\cdot x_i}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\
&= \sum_{i=1}^{N} - x_i \frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i} - \mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\
&= \sum_{i=1}^{N} - \frac{x_i^2 \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\
\frac{\partial^2}{\partial \alpha_1 \partial \alpha_0}\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) &= \frac{\partial}{\partial \alpha_1} \sum_{i=1}^{N} \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \alpha_1} \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N} \left[0 - 0 - \frac{\partial}{\partial \alpha_1} \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}} \right] \\
&= \sum_{i=1}^{N} - \frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\
\frac{\partial^2}{\partial \alpha_0 \partial \alpha_1}\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) &= \frac{\partial}{\partial \alpha_0} \sum_{i=1}^{N} x_i \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial \alpha_0} x_i \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right) \\
&= \sum_{i=1}^{N} x_i \frac{\partial}{\partial \alpha_0} \left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right)
\end{aligned}
$$

$$= \sum_{i=1}^{N} -\frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2}$$

Then, the Hessian is

$$\begin{aligned}
\mathbf{H}_\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) &= \begin{bmatrix} -\sum_{i=1}^{N} \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} & -\sum_{i=1}^{N} \frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\ -\sum_{i=1}^{N} \frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} & -\sum_{i=1}^{N} \frac{x_i^2 \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \end{bmatrix} \\
&= -\sum_{i=1}^{N} \begin{bmatrix} \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} & \frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \\ \frac{x_i \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} & \frac{x_i^2 \mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \end{bmatrix} \\
&= -\sum_{i=1}^{N} \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} \\
&= -\sum_{i=1}^{N} \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix}.
\end{aligned}$$

We have $\mathrm{e}^z > 0$ for $z \in \mathbb{R}$, so it follows that the quantity

$$k_i = \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2}, \quad i = 1, \ldots, N$$

is positive. Now, suppose that $\mathbf{z} \in \mathbb{R}^2$ and $\mathbf{z} \neq \mathbf{0}$. Noting that we can write the Hessian as

$$\mathbf{H}_\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) = -\sum_{i=1}^{N} k_i \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} = \begin{bmatrix} -k_1 & -k_1 x_1 \\ -k_1 x_1 & -k_1 x_1^2 \end{bmatrix} + \cdots + \begin{bmatrix} -k_N & -k_N x_N \\ -k_N x_N & -k_N x_N^2 \end{bmatrix} = \mathbf{A}_1 + \cdots + \mathbf{A}_N,$$

we have

$$\begin{aligned}
\mathbf{z}^\mathsf{T} \mathbf{H}_\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) \mathbf{z} &= \mathbf{z}^\mathsf{T} \left(\mathbf{A}_1 + \cdots + \mathbf{A}_N\right) \mathbf{z} \\
\text{(distributive property)} \qquad &= \mathbf{z}^\mathsf{T} \left(\mathbf{A}_1 \mathbf{z} + \cdots + \mathbf{A}_N \mathbf{z}\right) \\
\text{(distributive property)} \qquad &= \mathbf{z}^\mathsf{T} \mathbf{A}_1 \mathbf{z} + \cdots + \mathbf{z}^\mathsf{T} \mathbf{A}_N \mathbf{z}.
\end{aligned}$$

For any $\mathbf{A}_i$, we have

$$\begin{aligned}
\mathbf{z}^\mathsf{T} \mathbf{A}_i \mathbf{z} &= \mathbf{z}^\mathsf{T} \begin{bmatrix} -k_i & -k_i x_i \\ -k_i x_i & -k_i x_i^2 \end{bmatrix} \mathbf{z} \\
\text{(k_i is a scalar)} \qquad &= \mathbf{z}^\mathsf{T} - k_i \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix} \mathbf{z} \\
&= -k_i \mathbf{z}^\mathsf{T} \begin{bmatrix} 1 \\ x_i \end{bmatrix} \begin{bmatrix} 1 & x_i \end{bmatrix} \mathbf{z} \\
(\mathbf{u}^\mathsf{T}\mathbf{v} = \mathbf{u} \cdot \mathbf{v}) \qquad &= -k_i \left(\mathbf{z} \cdot \begin{bmatrix} 1 \\ x_i \end{bmatrix}\right) \left(\begin{bmatrix} 1 \\ x_i \end{bmatrix} \cdot \mathbf{z}\right) \\
(\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}) \qquad &= -k_i \left(\mathbf{z} \cdot \begin{bmatrix} 1 \\ x_i \end{bmatrix}\right)^2.
\end{aligned}$$

k_i is positive, and the square of the dot product of $\mathbf{z}$ and $\begin{bmatrix} 1 \\ x_i \end{bmatrix}$ will be nonnegative (we may have $z_1 + z_2 x_i = 0$), so that

$$\mathbf{z}^\mathsf{T} \mathbf{A}_i \mathbf{z} = -k_i \left(\mathbf{z} \cdot \begin{bmatrix} 1 \\ x_i \end{bmatrix}\right)^2 \leq 0,$$

i.e., $\mathbf{A}_i$ is negative semidefinite. Now, $\mathbf{z}^\mathsf{T} \mathbf{H}_\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) \mathbf{z}$ is equal to the sum of the $\mathbf{z}^\mathsf{T} \mathbf{A}_i \mathbf{z}$ terms, all of which are less than or equal to zero, i.e., each $\mathbf{A}_i$ is negative semidefinite, so it follows that $\mathbf{z}^\mathsf{T} \mathbf{H}_\ell\left(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}\right) \mathbf{z}$ will be be less than or equal to zero. (The inequality will be strict in the case that at least two of the x_i's are

non-zero and unequal, which is likely for a reasonable data set.) The log-likelihood is thus concave, and the maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$ corresponds to a global maximum. Now consider the negative log-likelihood, i.e., $-\ell(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y})$. We may equivalently minimize this function to find the maximum likelihood estimate $\hat{\boldsymbol{\alpha}}$. In this case, we will have

$$\nabla\left[-\ell(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y})\right] = -\sum_{i=1}^{N}\left(1 - y_i - \frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}}\right)\begin{bmatrix}1\\ x_i\end{bmatrix}$$

and thus

$$\mathbf{H}_{-\ell}(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y}) = -\left[-\sum_{i=1}^{N}\frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2}\begin{bmatrix}1\\ x_i\end{bmatrix}\begin{bmatrix}1 & x_i\end{bmatrix}\right] = \sum_{i=1}^{N}\frac{\mathrm{e}^{\alpha_0+\alpha_1 x_i}}{\left(1+\mathrm{e}^{\alpha_0+\alpha_1 x_i}\right)^2}\begin{bmatrix}1\\ x_i\end{bmatrix}\begin{bmatrix}1 & x_i\end{bmatrix}.$$

We have shown that $\mathbf{z}^{\mathsf{T}}\mathbf{H}_{\ell}(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y})\,\mathbf{z} \leq 0$ for all $\mathbf{z} \neq \mathbf{0}$, so it follows that

$$\mathbf{z}^{\mathsf{T}}\mathbf{H}_{\ell}(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y})\,\mathbf{z} \leq 0 \implies -\mathbf{z}^{\mathsf{T}}\mathbf{H}_{\ell}(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y})\,\mathbf{z} \geq 0 \implies \mathbf{z}^{\mathsf{T}}\mathbf{H}_{-\ell}(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y})\,\mathbf{z} \geq 0 \quad \forall\, \mathbf{z} \neq \mathbf{0},$$

i.e., the Hessian of the negative log-likelihood is positive semidefinite. In the case that the inequality is strict (which will be true for a reasonable data set), then theorem 16.19 implies that $-\ell(\boldsymbol{\alpha}|\mathbf{x},\mathbf{y})$ is convex, so that finding the maximum likelihood estimate, i.e., minimizing the negative log-likelihood, is a convex optimization problem.

We now consider the problem in multiple dimensions, i.e., when $\mathbf{x} \in \mathbb{R}^n$ for $n > 1$. In this case, logistic regression assumes the model

$$P(Y = 1|x,\boldsymbol{\alpha}) = \frac{1}{1+\exp\left(\alpha_0 + \sum_{i=1}^{n}\alpha_i x_i\right)}.$$

In the multidimensional case, we can apply similar ideas as in the 1-dimensional case to conclude that logistic regression corresponds to a convex optimization.

16.4. Non-convex optimization

We have previously considered optimizing functions known to be convex. We saw that we can optimize a quadratic analytically by solving the normal equations, and that we can optimize any convex function by applying Damped Newton's Method. We will see that these approaches will fail for non-convex functions.

EXAMPLE 16.22. Suppose that we wish to find the global minimum of $f(x) = x^4 - 5x^3 - 2x^2$, shown in figure 16.4.1.

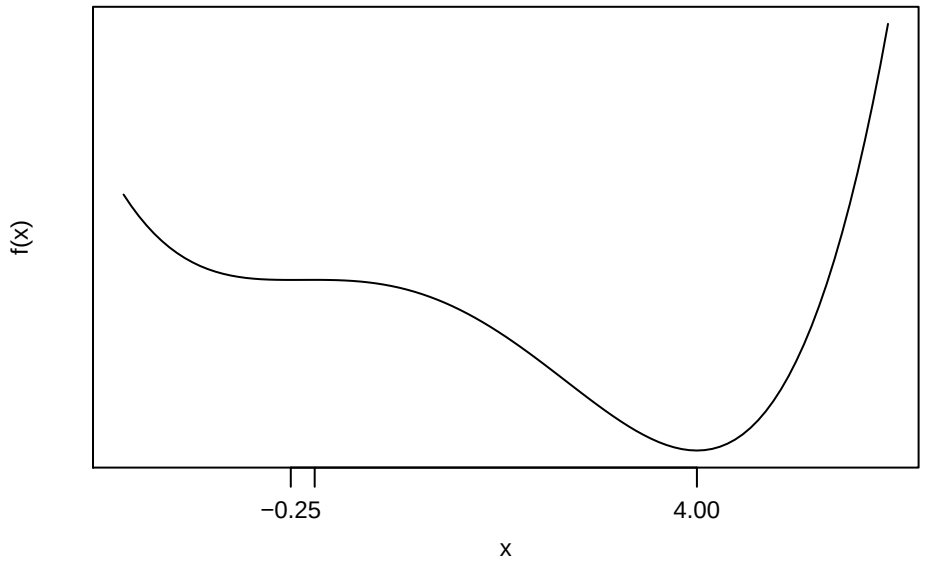


FIGURE 16.4.1. $f(x) = x^4 - 5x^3 - 2x^2$ is non-convex

The first derivative of f is $f'(x) = 4x^3 - 15x^2 - 4x$. Setting this equal to zero and solving gives

$$0 = 4x^3 - 15x^2 - 4x = x\left(4x^2 - 15x - 4\right) = x(4x+1)(x-4)$$

so f' will equal zero in the case that any of these terms is zero, i.e., when $x = 0$, when $x = -1/4$, and when $x = 4$. To characterize each of the critical points of f, we examine its second derivative, which is given by $f''(x) = 12x^2 - 30x - 4$. We have

$$f''(0) = -4, \quad f''\left(-\frac{1}{4}\right) = \frac{12}{16} + \frac{30}{4} - 4 = \frac{17}{4}, \quad f''(4) = 192 - 120 - 4 = 68,$$

so that f has a maximum at $x = 0$ and minima at $x = -1/4$ and $x = 4$. We have

$$f\left(-\frac{1}{4}\right) = \frac{1}{256} + \frac{5}{64} - \frac{2}{16} = -\frac{11}{256} \quad \text{and} \quad f(4) = 256 - 320 - 32 = -96,$$

so it is clear that f attains a global minimum at $x = 4$. Although we were able to find the global minimum analytically in this case, in higher dimensions the analysis becomes much more complicated, and in addition closed-form solutions may not exist, so that we must optimize such functions numerically. Had we applied Newton's Method to this problem, we would have had no guarantee of finding the global minimum, i.e., depending on our starting point, we may have instead converged to the local minimum at $x = -1/4$.

We saw in the previous example that Newton's Method is not guaranteed to find the global minimum of a non-convex function, in particular because it may not produce descent directions. We might instead applied Steepest Descent with backtracking, but this will be slow. We will consider three alternatives: Newton's Method, modified to produce descent directions, is suitable for any non-convex function, and both the Gauss-Newton and Levenberg-Marquardt algorithms are suitable for fitting a non-linear least-squares regression.

16.4.1. Newton's Method with Hessian modification. Suppose that we wish to minimize some non-convex function $f(\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^n$. Recall that Damped Newton's Method will give a descent direction at $\mathbf{x}$ if the Hessian of f evaluated at $\mathbf{x}$ $\mathbf{H}_f(\mathbf{x})$ is positive definite. If $\mathbf{H}_f$ is not positive definite, then we can create a new matrix $\mathbf{A}$ that is both positive definite and "close" to the Hessian. Now, under suitable regularity conditions, Clairaut's Theorem implies that $\mathbf{H}_f$ is symmetric, and it follows from corollary 16.10 that $\mathbf{H}_f$ may be decomposed as $\mathbf{H}_f = \mathbf{Q}\mathbf{D}\mathbf{Q}^\mathsf{T}$, where $\mathbf{Q}$ is an orthonormal matrix and $\mathbf{D}$ is a diagonal matrix whose diagonal entries λ_i are the eigenvalues of $\mathbf{H}_f$. If $\mathbf{H}_f$ is not positive definite, then we will have $\min(\lambda_1, \ldots, \lambda_n) < 0$. Let

$$\lambda_{\min} = |\min(\lambda_1, \ldots, \lambda_n)|$$

and consider the sum $\mathbf{D}^* = (\lambda_{\min} + k)\mathbf{I}_n + \mathbf{D}$, i.e.,

$$\begin{bmatrix} \lambda_{\min}+k & 0 & \cdots & 0 \\ 0 & \lambda_{\min}+k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{\min}+k \end{bmatrix} + \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

where $k \in \mathbb{R}$ is some positive constant. By adding $\lambda_{\min} + k$ to each entry of $\mathbf{D}$, we ensure that the entries of $\mathbf{D}^*$ will all be positive. We will compute $\mathbf{A}$ as

$$\begin{aligned} \mathbf{A} &= \mathbf{Q}\mathbf{D}^*\mathbf{Q}^\mathsf{T} \\ &= \mathbf{Q}\left(\mathbf{D} + (\lambda_{\min}+k)\mathbf{I}\right)\mathbf{Q}^\mathsf{T} \\ &= \left(\mathbf{Q}\mathbf{D} + \mathbf{Q}(\lambda_{\min}+k)\mathbf{I}\right)\mathbf{Q}^\mathsf{T} \\ &= \mathbf{Q}\mathbf{D}\mathbf{Q}^\mathsf{T} + \mathbf{Q}(\lambda_{\min}+k)\mathbf{I}\mathbf{Q}^\mathsf{T} \\ &= \mathbf{H}_f(\mathbf{x}) + (\lambda_{\min}+k)\mathbf{Q}\mathbf{I}\mathbf{Q}^\mathsf{T} \\ &= \mathbf{H}_f(\mathbf{x}) + (\lambda_{\min}+k)\mathbf{Q}\mathbf{Q}^\mathsf{T} \\ (\mathbf{Q}^\mathsf{T} = \mathbf{Q}^{-1}) \quad &= \mathbf{H}_f(\mathbf{x}) + (\lambda_{\min}+k)\mathbf{I}, \end{aligned}$$

i.e., at the ith iteration, we will find $\lambda_{\min}$ of $\mathbf{H}_f$ and form $\mathbf{A}$ as the sum of $(\lambda_{\min}+k)\mathbf{I}$ and $\mathbf{H}_f$. Now, consider the tuning parameter k. If $\lambda_{\min} + k = 0$, then $\mathbf{A} = \mathbf{H}_f$, i.e., we have the generic Newton's Method direction at $\boldsymbol{\alpha}$. If $\lambda_{\min} + k$ is very large, the eigenvalues of $\mathbf{A}$ will be very large, so that the eigenvalues of

$\mathbf{A}^{-1}$ will be very small. In this case, multiplication by $\mathbf{A}^{-1}$ will have a negligible effect on $-\nabla f(\mathbf{x})$, so that the resulting direction will be roughly that given by Steepest Descent. We will therefore choose k through trial-and-error, attempting to find a value that works well for a given starting point.

Algorithm 7 Newton's Method with Hessian Modification

Require: $\mathbf{x} \in \mathbb{R}^n$, $k \in \mathbb{R}$
while $\|\nabla f(\mathbf{x})\| > \epsilon$ **do**
  $\lambda_{\min} \leftarrow \min(\lambda_1, \ldots, \lambda_n)$ ▷ λ_j is the jth eigenvalue of $\mathbf{H}_f$
  if $\lambda_{\min} > 0$ **then**
    $\mathbf{A} \leftarrow \mathbf{H}_f(\mathbf{x})$
  else
    $\mathbf{A} \leftarrow (|\lambda_{\min}| + k)\mathbf{I} + \mathbf{H}_f(\mathbf{x})$
  end if
  solve $\mathbf{A}\mathbf{d} = -\nabla f(\mathbf{x})$ for $\mathbf{d}$
  $s \leftarrow 1$
  while $f(\mathbf{x} + s\mathbf{d}) > f(\mathbf{x})$ **do**
    $s \leftarrow s/2$
  end while
  $\mathbf{x} \leftarrow \mathbf{x} + s\mathbf{d}$
end while

We also may be able to use Hessian modification to improve the condition number of the matrix we invert to find the Newton's Method direction. If the condition number of $\mathbf{H}_f$ (which is $\kappa(\mathbf{H}_f) = |\lambda_1/\lambda_n|$) is large, then the direction obtained by solving $\mathbf{d} = -[\mathbf{H}_f(\mathbf{x})]^{-1}\nabla f(\mathbf{x})$ will be inaccurate. Informally, if f is "flat" around the value $\tilde{\mathbf{x}}$, then the Hessian of f will be close to zero, so that vectors will go to zero, hence $\lambda_n \approx 0$, and $\kappa(\mathbf{H}_f(\mathbf{x}))$ will be large. When we modify the Hessian, the ith eigenvalue of the resulting matrix $\mathbf{A}$ will be on the order of $\lambda_i + \lambda_{\min} + k$, where λ_i is the ith eigenvalue of $\mathbf{H}_f$. The condition number of $\mathbf{A}$ is thus on the order of

$$\kappa(\mathbf{A}) = \frac{\lambda_1 + \lambda_{\min} + k}{\lambda_n + \lambda_{\min} + k}.$$

Suppose that $\mathbf{x} \in \mathbb{R}^2$, and that the eigenvalues of $\mathbf{H}_f$ are $\lambda_1 = 10$ and $\lambda_2 = 10^{-5}$, so that $\kappa(\mathbf{H}_f) = 10/10^{-5} = 10^6$. Modifying the Hessian with $k = 1$, the eigenvalues of $\mathbf{A}$ are on the order of

$$\lambda_1 + \lambda_{\min} + 1 = 10 + 10^{-5} + 1 \quad \text{and} \quad \lambda_2 + \lambda_{\min} + 1 = 10^{-5} + 10^{-5} + 1,$$

so that

$$\kappa(\mathbf{A}) \approx \frac{11 + 10^{-5}}{1 + 2 \cdot 10^{-5}} = 10.99979,$$

which we see is greatly improved relative to $\kappa(\mathbf{H}_f)$.

16.4.2. Gauss-Newton method. Suppose that we collect data that we believe is well represented by a mixture of two Gaussians. Let y_i be the ith observation and x_i be the ith covariate, so that our model is

$$y_i \sim a_1 \exp\left\{-\frac{(x_i - \mu_1)^2}{\sigma_1^2}\right\} + a_2 \exp\left\{-\frac{(x_i - \mu_2)^2}{\sigma_2^2}\right\},$$

where (μ_1, σ_1^2) and (μ_2, σ_2^2) are the mean and variance of the respective Gaussians and a_1 and a_2 are the mixture weights, and where we have omitted the normalizing constants for clarity. Figure 16.4.2 depicts the model and observations.

Suppose that we draw n observations and wish to infer the parameters $\boldsymbol{\alpha} = (a_1, u_1, \sigma_1^2, a_2, \mu_2, \sigma_2^2)$. Define the least-squares loss function

$$L(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\left(y_i - a_1 \exp\left\{-\frac{(x_i - \mu_1)^2}{\sigma_1^2}\right\} - a_2 \exp\left\{-\frac{(x_i - \mu_2)^2}{\sigma_2^2}\right\}\right)^2 = \sum_{i=1}^{n}\left(r_i(\boldsymbol{\alpha})\right)^2,$$

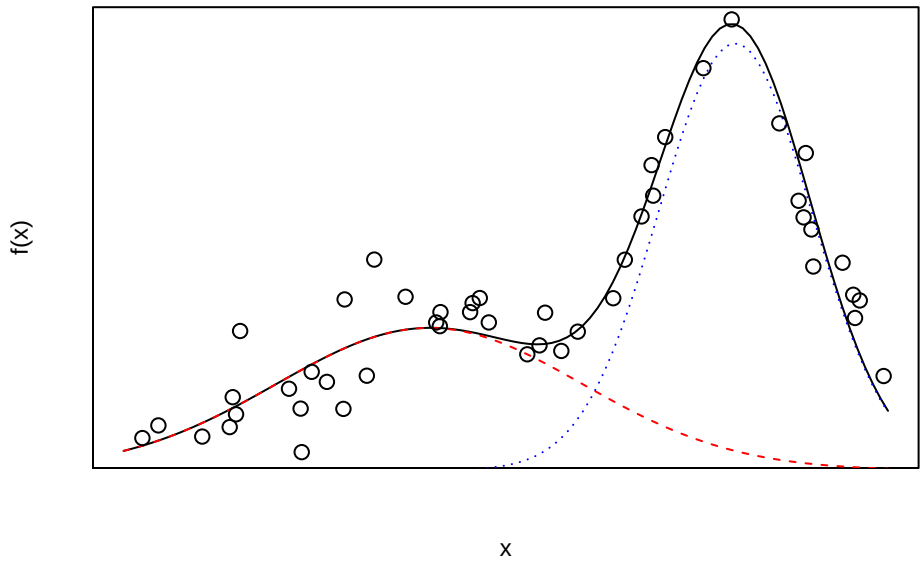


FIGURE 16.4.2. Observations drawn from Gaussian mixture model

and consider finding the $\hat{\boldsymbol{\alpha}}$ that minimizes the sum of the squared residuals, i.e.,

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \left(r_i(\boldsymbol{\alpha})\right)^2.$$

Now, the gradient of $L(\boldsymbol{\alpha})$ can be expressed as

$$\nabla L(\boldsymbol{\alpha}) = \nabla\left(\sum_{i=1}^{n} \left(r_i(\boldsymbol{\alpha})\right)^2\right) = \sum_{i=1}^{n} \nabla\left(r_i(\boldsymbol{\alpha})\right)^2 = \sum_{i=1}^{n} 2 r_i(\boldsymbol{\alpha}) \nabla r_i(\boldsymbol{\alpha}),$$

so that the Hessian of $L(\boldsymbol{\alpha})$ is a 6×6 matrix whose kjth entry is

$$\begin{aligned}
\frac{\partial^2}{\partial\alpha_k \partial\alpha_j} L(\boldsymbol{\alpha}) &= \frac{\partial}{\partial\alpha_k}\left[\frac{\partial}{\partial\alpha_j} L(\boldsymbol{\alpha})\right] \\
&= \frac{\partial}{\partial\alpha_k} \sum_{i=1}^{n} 2 r_i(\boldsymbol{\alpha}) r_i^{(j)}(\boldsymbol{\alpha}) \\
&= \sum_{i=1}^{n} 2 \frac{\partial}{\partial\alpha_k} r_i(\boldsymbol{\alpha}) r_i^{(j)}(\boldsymbol{\alpha}) \\
&= 2 \sum_{i=1}^{n} \left[r_i^{(k)}(\boldsymbol{\alpha}) r_i^{(j)}(\boldsymbol{\alpha}) + r_i(\boldsymbol{\alpha}) \frac{\partial}{\partial\alpha_k} r_i^{(j)}(\boldsymbol{\alpha})\right],
\end{aligned}$$

where $r_i^{(j)}(\boldsymbol{\alpha})$ denotes the derivative of $r_i(\boldsymbol{\alpha})$ with respect to the jth component of $\boldsymbol{\alpha}$, and where the final equality follows from the product rule. Then, the Hessian is

$$\mathbf{H}_L(\boldsymbol{\alpha}) = 2 \sum_{i=1}^{n} \begin{bmatrix}
r_i^{(1)}(\boldsymbol{\alpha}) r_i^{(1)}(\boldsymbol{\alpha}) + r_i(\boldsymbol{\alpha}) \dfrac{\partial}{\partial\alpha_1} r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & r_i^{(1)}(\boldsymbol{\alpha}) r_i^{(6)}(\boldsymbol{\alpha}) + r_i(\boldsymbol{\alpha}) \dfrac{\partial}{\partial\alpha_1} r_i^{(6)}(\boldsymbol{\alpha}) \\
\vdots & \ddots & \vdots \\
r_i^{(6)}(\boldsymbol{\alpha}) r_i^{(1)}(\boldsymbol{\alpha}) + r_i(\boldsymbol{\alpha}) \dfrac{\partial}{\partial\alpha_6} r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & r_i^{(6)}(\boldsymbol{\alpha}) r_i^{(6)}(\boldsymbol{\alpha}) + r_i(\boldsymbol{\alpha}) \dfrac{\partial}{\partial\alpha_6} r_i^{(6)}(\boldsymbol{\alpha})
\end{bmatrix}$$

$$= 2\sum_{i=1}^{n}\left(\begin{bmatrix} r_i^{(1)}(\boldsymbol{\alpha})\, r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & r_i^{(1)}(\boldsymbol{\alpha})\, r_i^{(6)}(\boldsymbol{\alpha}) \\ \vdots & \ddots & \vdots \\ r_i^{(6)}(\boldsymbol{\alpha})\, r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & r_i^{(6)}(\boldsymbol{\alpha})\, r_i^{(6)}(\boldsymbol{\alpha}) \end{bmatrix} + \begin{bmatrix} r_i(\boldsymbol{\alpha})\dfrac{\partial}{\partial\alpha_1} r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & r_i(\boldsymbol{\alpha})\dfrac{\partial}{\partial\alpha_1} r_i^{(6)}(\boldsymbol{\alpha}) \\ \vdots & \ddots & \vdots \\ r_i(\boldsymbol{\alpha})\dfrac{\partial}{\partial\alpha_6} r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & r_i(\boldsymbol{\alpha})\dfrac{\partial}{\partial\alpha_6} r_i^{(6)}(\boldsymbol{\alpha}) \end{bmatrix}\right)$$

$$= 2\sum_{i=1}^{n}\left(\begin{bmatrix} r_i^{(1)}(\boldsymbol{\alpha}) \\ \vdots \\ r_i^{(6)}(\boldsymbol{\alpha}) \end{bmatrix}\begin{bmatrix} r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & r_i^{(6)}(\boldsymbol{\alpha}) \end{bmatrix} + r_i(\boldsymbol{\alpha})\begin{bmatrix} \dfrac{\partial}{\partial\alpha_1} r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & \dfrac{\partial}{\partial\alpha_1} r_i^{(6)}(\boldsymbol{\alpha}) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial}{\partial\alpha_6} r_i^{(1)}(\boldsymbol{\alpha}) & \cdots & \dfrac{\partial}{\partial\alpha_6} r_i^{(6)}(\boldsymbol{\alpha}) \end{bmatrix}\right),$$

where we observe that $r_i(\boldsymbol{\alpha})$ is a scalar to obtain the final equality. Now, the entires of the matrix in the second term of the final equality are the second-order partial derivatives of $r_i(\boldsymbol{\alpha})$, which is precisely the Hessian of r_i. Thus, the Hessian of $L(\boldsymbol{\alpha})$ finally becomes

$$\mathbf{H}_L(\boldsymbol{\alpha}) = 2\sum_{i=1}^{n}\left(\nabla r_i(\boldsymbol{\alpha})\left(r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}} + r_i(\boldsymbol{\alpha})\,\mathbf{H}_{r_i}(\boldsymbol{\alpha})\right).$$

Let $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$, and let θ be the angle formed by $\mathbf{u}$ and $\mathbf{w}$. Then, another definition of the dot product of $\mathbf{u}$ and $\mathbf{w}$ is $\mathbf{u}\cdot\mathbf{w} = \|\mathbf{u}\|_2\|\mathbf{w}\|_2\cos\theta$, which implies

$$\mathbf{u}\cdot\mathbf{u} = \|\mathbf{u}\|_2\|\mathbf{u}\|_2\cos\theta = \|\mathbf{u}\|_2^2\cos 0 = \|\mathbf{u}\|_2^2\cdot 1 = \|\mathbf{u}\|_2^2 = \left(\sqrt{u_1^2+\cdots+u_n^2}\right)^2 = \sum_{i=1}^{n} u_i^2,$$

which is easily seen to be nonnegative. Now, suppose that $\mathbf{z}\in\mathbb{R}^6$ and $\mathbf{z}\neq\mathbf{0}$. Then,

$$\begin{aligned}
&\text{(associativity)} && \mathbf{z}^{\mathsf{T}}\left[\nabla r_i(\boldsymbol{\alpha})\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\right]\mathbf{z} = \left(\mathbf{z}^{\mathsf{T}}\nabla r_i(\boldsymbol{\alpha})\right)\left[\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\mathbf{z}\right] \\
&((\mathbf{uv})^{\mathsf{T}} = \mathbf{v}^{\mathsf{T}}\mathbf{u}^{\mathsf{T}}) && = \left[\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\mathbf{z}\right]^{\mathsf{T}}\left[\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\mathbf{z}\right] \\
&(\mathbf{u}^{\mathsf{T}}\mathbf{v} = \mathbf{u}\cdot\mathbf{v}) && = \left[\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\mathbf{z}\right]\cdot\left[\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\mathbf{z}\right] \\
&&& = \left\|\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\mathbf{z}\right\|_2^2.
\end{aligned}$$

This quantity will be nonnegative for all $\mathbf{z}\neq\mathbf{0}$, so it follows that the matrix $\nabla r_i(\boldsymbol{\alpha})\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}$ is positive semidefinite (and for any reasonable data set, it will be positive definite). Thus, while $\mathbf{H}_{r_i}(\boldsymbol{\alpha})$ may be positive definite, we are assured that $\nabla r_i(\boldsymbol{\alpha})\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}$ will be (at least) positive semidefinite. We wish to minimize the sum of the squared residuals, so as we approach the minimum, i.e., as the model becomes a better fit to the data, the residuals will go to zero, so that the term $r_i(\boldsymbol{\alpha})\,\mathbf{H}_{r_i}(\boldsymbol{\alpha})$ will go to zero. The *Gauss-Newton Method* drops this term and replaces $\mathbf{H}_L(\boldsymbol{\alpha})$ by

$$\mathbf{H}_L^{\star}(\boldsymbol{\alpha}) = 2\sum_{i=1}^{n}\nabla r_i(\boldsymbol{\alpha})\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}},$$

which will be positive semidefinite because it is the sum of n positive semidefinite matrices. Algorithm 8 provides a formal statement of the procedure.

16.4.3. Levenberg-Marquardt method. The Gauss-Newton Method replaces the Hessian of the least-square loss function by the sum of positive semidefinite matrices (positive definite for any reasonable data set), which will reliably give descent directions. While the ith matrix in this sum is positive semidefinite, it may be ill-conditioned, i.e., $\kappa\left(\nabla r_i(\boldsymbol{\alpha})\left(\nabla r_i(\boldsymbol{\alpha})\right)^{\mathsf{T}}\right)$ may be large, in which case the direction obtained (by inverting this matrix) will be inaccurate. The *Levenberg-Marquardt Method* introduces an additional term

Algorithm 8 Gauss-Newton Method with backtracking

Require: $\boldsymbol{\alpha}$
while $\|\nabla L(\boldsymbol{\alpha})\| > \epsilon$ **do**
  $\mathbf{H}_L^\star(\boldsymbol{\alpha}) \leftarrow 2\sum_{i=1}^n \nabla r_i(\boldsymbol{\alpha})(\nabla r_i(\boldsymbol{\alpha}))^\mathsf{T}$
  solve $\mathbf{H}_L^\star(\boldsymbol{\alpha})\mathbf{d} = -\nabla L(\boldsymbol{\alpha})$ for $\mathbf{d}$
  $s \leftarrow 1$
  while $L(\boldsymbol{\alpha} + s\mathbf{d}) > L(\boldsymbol{\alpha})$ **do**
    $s \leftarrow s/2$
  end while
  $\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} + s\mathbf{d}$
end while

$c\mathbf{I}_n$ for some $c > 0$ to improve the condition number, so that the replacement Hessian is

$$\mathbf{H}_L^\star(\boldsymbol{\alpha}) = 2\sum_{i=1}^n \left[c\mathbf{I}_n + \nabla r_i(\boldsymbol{\alpha})(\nabla r_i(\boldsymbol{\alpha}))^\mathsf{T}\right].$$

CHAPTER 17

Numerical linear algebra

17.1. Machine representation

Suppose that we encode an integer using 32 *bits* (equivalently, 4 *bytes*), each of which may be 0 or 1. We allocate a single bit s to hold the sign, and we denote the kth bit by i_k. Then, an integer x can be written as

$$x = (-1)^s \cdot \left(i_1 2^0 + i_2 2^1 + i_3 2^2 + \cdots + i_{31} 2^{30}\right) = (-1)^s \sum_{k=1}^{31} i_k 2^{k-1},$$

and we can store this representation as

s	i_1	i_2	i_3	$\cdots$	i_{30}	i_{31}

where $i_k = 1$ if the kth term is included in the sum and 0 if it is not. Now, this encoding is inefficient because it represents zero in two ways (zero is unsigned, so we may have $s = 0$ or $s = 1$), hence we are losing a bit. Suppose instead that we allocate all 32 bits to terms in the above summation, and by convention subtract 2^{31} from the sum, so that negative integers can be represented. Then, under this scheme,

$$x = \left(i_1 2^0 + i_2 2^1 + i_3 2^2 + \cdots + i_{31} 2^{30} + i_{32} 2^{31}\right) - 2^{31} = \sum_{k=1}^{32} i_k 2^{k-1} - 2^{31}.$$

To represent 1, we will have $i_1 = 1$, $i_{32} = 1$, and $i_k = 0$ for $k \in \{2, 3, \ldots, 31\}$. The next consecutive integer is 2, which we represent with $i_2 = 1$, $i_{32} = 1$, and $i_k = 0$ for $k \in \{1\} \cup \{3, 4, \ldots, 31\}$. We represent 3 as $i_1 = 1$, $i_2 = 1$, $i_{32} = 1$, and $i_k = 0$ for $k \in \{3, 4, \ldots, 31\}$. We proceed in this fashion until $i_k = 1$ for $k \in \{1, 2, \ldots, 32\}$, i.e.,

$$x_{\max} = \left[2^0 + 2^1 + 2^2 + \cdots + 2^{30} + 2^{31}\right] - 2^{31} = 2^0 + 2^1 + 2^2 + \cdots + 2^{30} = 2147483647 = 2^{31} - 1,$$

which is the largest integer that can be stored under this encoding scheme. The largest negative integer that can be stored occurs when each i_k is 0, and this is -2^{31}. Zero is represented by setting $i_{32} = 1$ and $i_k = 0$ for $k \in \{1, 2, \ldots, 31\}$.

17.1.1. Floating-point numbers. Under the IEEE standard, numbers are encoded as 64-bit words of the form

s	e_1	e_2	$\cdots$	e_{11}	b_1	b_2	$\cdots$	b_{52}

where s is the sign bit and the 52 b_j bits represent the mantissa. The 11 e_k bits represent the positive binary integer that is the sum of the exponent and the *bias* $2^{10} - 1 = 1023$ (for exponents between -1022 and 1023). For example, an exponent of 0 is represented as $0 + 1023 = (11\,1111\,1111)_2$, so that $\boxed{e_1 e_2 \ldots e_{11}} =$ `011 1111 1111`. An exponent of 1 is represented as $1+1023 = 1024 = (100\,0000\,0000)_2$, so that $\boxed{e_1 e_2 \ldots e_{11}} =$ `100 0000 0000`. To find the actual exponent, we must subtract the bias. Then, under this scheme, the representation of $x \in \mathbb{R}$ is

$$x = (-1)^s \cdot 1.\boxed{b_1 b_2 \ldots b_{52}} \cdot 2^{\boxed{e_1 e_2 \ldots e_{11}} - 1023},$$

where $\boxed{b_1 b_2 \ldots b_{52}}$ is the concatenation of the b_j bits, so that the term 2^{-j} should be included in the decimal representation of x if the jth mantissa bit is 1, and $\boxed{e_1 e_2 \ldots e_{11}}$ is the concatenation of the e_k bits.

Now, the exponent value $2047_{10} = (111\,1111\,1111)_2$ is reserved to represent infinity if every b_j is zero, i.e., if the mantissa bits are all zero, and to represent NaN (Not a Number) otherwise. The exponent value 0 is used to represent *subnormal* floating point numbers, or those numbers where the left-most bit is not

assumed to be 1. Thus, the smallest non-reserved value the exponent bits can take is `000 0000 001`, which corresponds to an exponent of $2^0 - 1023 = -1022$. The largest non-reserved value the exponent bits can take is `111 1111 1110`, which corresponds to an exponent of

$$\sum_{k=1}^{10} 2^k - 1023 = 1023.$$

Thus, the range of the exponent is $\left(-2^{10}+2, 2^{10}-1\right)$, so that the largest number that can be represented using the *double precision* floating-point encoding is $2^{2^{10}-1}$. In this encoding, the *radix point* (generalization of a decimal point) "floats." We also observe that we cannot represent every real number exactly. For example, $\sqrt{2}$ does not have a finite decimal expansion, and extremely large or small numbers cannot be represented exactly due to the defined (and finite) number of bits available for representation.

DEFINITION 17.1. For some $x \in \mathbb{R}$, denote by $\mathrm{fl}(x)$ the closest double precision floating-point number to x.

We now consider the accuracy of floating-point representation. We have 52 stored mantissa bits plus 1 implicit leading bit to store the precision of x, which is equivalent to 15-17 decimal bits. The *relative roundoff error* in representing x is $|\mathrm{fl}(x) - x| / |x|$.

EXAMPLE 17.2. Suppose that $x = 12345678901234567890$. Assuming that 16 decimal bits of precision are available to represent x, we have $\mathrm{fl}(x) = 12345678901234560000$, so that the relative roundoff error is

$$\frac{|\mathrm{fl}(x) - x|}{|x|} = \frac{|7890|}{|x|} \approx \frac{10^4}{10^{20}} = 10^{-16}.$$

DEFINITION 17.3. The number *machine epsilon*, denoted ϵ_{mach}, is the distance between 1 and the smallest floating point number greater than 1.

Alternatively, ϵ_{mach} is the smallest positive number for which $\mathrm{fl}(1 + \epsilon_{\text{mach}}) \neq 1$. Under the IEEE double precision floating-point standard, machine epsilon is $2^{-52} \approx 10^{-16}$. The relative roundoff error in representing some $x \neq 0$ will be at most ϵ_{mach}, i.e.,

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \leq \epsilon_{\text{mach}}.$$

In the worst case, i.e., equality, we will have $|\mathrm{fl}(x) - x| = |x|\,\epsilon_{\text{mach}}$. If $\mathrm{fl}(x) \geq x$, then $|\mathrm{fl}(x) - x| \geq 0$, so that this expression becomes

$$|\mathrm{fl}(x) - x| = |x|\,\epsilon_{\text{mach}} \implies \mathrm{fl}(x) - x = |x|\,\epsilon_{\text{mach}} \implies \mathrm{fl}(x) = x + |x|\,\epsilon_{\text{mach}}.$$

If $\mathrm{fl}(x) < x$, then $|\mathrm{fl}(x) - x| < 0$, so that the expression becomes

$$|\mathrm{fl}(x) - x| = |x|\,\epsilon_{\text{mach}} \implies -(\mathrm{fl}(x) - x) = |x|\,\epsilon_{\text{mach}} \implies \mathrm{fl}(x) = x - |x|\,\epsilon_{\text{mach}},$$

which we can express compactly as $\mathrm{fl}(x) = x \pm |x|\,\epsilon_{\text{mach}}$. If $x \geq 0$, then the right side of this expression becomes $x \pm x\epsilon_{\text{mach}} = x(1 \pm \epsilon_{\text{mach}})$. If $x < 0$, the right side of the expression becomes

$$x \pm (-x)\,\epsilon_{\text{mach}} = x \pm x\epsilon_{\text{mach}} = x(1 \pm \epsilon_{\text{mach}}),$$

hence $\mathrm{fl}(x) = x(1 \pm \epsilon_{\text{mach}})$ for all real x. Thus, the relative error of floating-point representation is bounded by $x(1 \pm \epsilon_{\text{mach}})$.

EXAMPLE 17.4. Suppose $x, y \in \mathbb{R}$, and consider the floating-point representation of $(x+y)^2$. We have

$$\begin{aligned}
\mathrm{fl}\left((x+y)^2\right) &= \mathrm{fl}\left((\mathrm{fl}(x) + \mathrm{fl}(y))^2\right) \\
&= \mathrm{fl}\left((\mathrm{fl}(x) + \mathrm{fl}(y)) \cdot (\mathrm{fl}(x) + \mathrm{fl}(y))\right) \\
&= \mathrm{fl}\left((x(1+\epsilon_1) + y(1+\epsilon_2)) \cdot (x(1+\epsilon_1) + y(1+\epsilon_2))\right) \\
&= \mathrm{fl}\left(x^2(1+\epsilon_1)^2(1+\epsilon_3) + y^2(1+\epsilon_2)^2(1+\epsilon_4) + 2xy(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_5)\right) \\
&= \left[x^2(1+\epsilon_1)^2(1+\epsilon_3) + y^2(1+\epsilon_2)^2(1+\epsilon_4) + 2xy(1+\epsilon_1)(1+\epsilon_2)(1+\epsilon_5)\right](1+\epsilon_6),
\end{aligned}$$

where the ϵ_i terms are specific to the respective floating-point representations of each quantity and are on the order of ϵ_{mach}. This expression can be simplified as

$$\text{fl}\left((x+y)^2\right) = (x+y)^2\left(1+c\epsilon_{\text{mach}}\right),$$

where $c \approx 4$ (and in particular, $c > 1$).

We see from this example that floating-point errors can add up. Let

$$\mathbf{A} = \begin{bmatrix} 10^7 & 0 & 0 \\ 0 & 10^{20} & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 5 \\ 3 \\ 2 \end{bmatrix},$$

and consider solving $\mathbf{Ax} = \mathbf{b}$. When we invert $\mathbf{A}$ to find the solution $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, we will mix quantities, e.g., 1 and 10^{20}, on very different scales, and the 1 will be wiped out by error.

17.2. Gaussian elimination

Consider the problem of finding $\mathbf{x} \in \mathbb{R}^n$ that satisfies $\mathbf{Ax} = \mathbf{b}$ for some $n \times n$ matrix $\mathbf{A}$ and n-dimensional vector $\mathbf{b}$. This problem arises in varied contexts, including solving the normal equations $\left(\mathbf{B}^\mathsf{T}\mathbf{B}\right)\boldsymbol{\alpha} = \mathbf{B}^\mathsf{T}\mathbf{y}$ and using Newton's method for optimization, where we must solve $\left[\mathbf{H}_f(\mathbf{x})\right]\mathbf{z} = \nabla f(\mathbf{x})$.

We can numerically solve the system $\mathbf{Ax} = \mathbf{b}$ by using *Gaussian elimination*, which reduces the augmented matrix $\left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \end{array}\right]$ to upper triangular form, then back-substitutes to solve for $\mathbf{x}$. Recall that there are two permitted operations: multiplying an equation by a constant and subtracting one equation from another.

Example 17.5. Solve the system of equations

$$\begin{aligned} 3x_1 + 5x_2 + x_3 &= 1 \\ 2x_1 + 2x_2 &= 2 \\ 3x_1 + 6x_2 + 10x_3 &= 3. \end{aligned}$$

We can write this system as

$$\begin{bmatrix} 3 & 5 & 1 \\ 2 & 2 & 0 \\ 3 & 6 & 10 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix},$$

which can be expressed the system in row-echelon form as

$$\left[\begin{array}{ccc|c} 3 & 5 & 1 & 1 \\ 2 & 2 & 0 & 2 \\ 3 & 6 & 10 & 3 \end{array}\right].$$

We wish to make this matrix upper triangular, i.e.,

$$\left[\begin{array}{ccc|c} 3 & 5 & 1 & 1 \\ 2 & 2 & 0 & 2 \\ 3 & 6 & 10 & 3 \end{array}\right] \begin{array}{l} \\ -\frac{2}{3}R_1 \\ \\ \end{array} \sim \left[\begin{array}{ccc|c} 3 & 5 & 1 & 1 \\ 0 & -\frac{4}{3} & -\frac{2}{3} & \frac{4}{3} \\ 3 & 6 & 10 & 3 \end{array}\right] \begin{array}{l} \\ \\ -R_1 \end{array} \sim \left[\begin{array}{ccc|c} 3 & 5 & 1 & 1 \\ 0 & -\frac{4}{3} & -\frac{2}{3} & \frac{4}{3} \\ 0 & 1 & 9 & 2 \end{array}\right] \begin{array}{l} \\ \\ +\frac{3}{4}R_2 \end{array} \sim \left[\begin{array}{ccc|c} 3 & 5 & 1 & 1 \\ 0 & -\frac{4}{3} & -\frac{2}{3} & \frac{4}{3} \\ 0 & 0 & \frac{17}{2} & 3 \end{array}\right].$$

We can now back-solve, so that

$$\begin{aligned} \frac{17}{2}x_3 = 3 \implies x_3 &= \frac{6}{17}, \\ -\frac{4}{3}x_2 - \frac{2}{3}\left(\frac{6}{17}\right) = \frac{4}{3} \implies -\frac{4}{3}x_2 = \frac{4}{3} + \frac{4}{17} \implies -\frac{4}{3}x_2 = \frac{80}{51} \implies x_2 &= -\frac{20}{17} \\ 3x_1 - 5\left(\frac{20}{17}\right) + \frac{6}{17} = 1 \implies 3x_1 = \frac{111}{17} \implies x_1 &= \frac{111}{51}. \end{aligned}$$

Thus,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \frac{111}{51} \\ -\frac{20}{17} \\ \frac{6}{17} \end{bmatrix}$$

is the solution to the system.

We now consider the computational complexity of Gaussian elimination for the general augmented matrix

$$\left[\begin{array}{c|c} \mathbf{A} & \mathbf{b} \end{array}\right] = \left[\begin{array}{cccc|c} A_{11} & A_{12} & \cdots & A_{1n} & b_1 \\ A_{21} & A_{22} & \cdots & A_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} & b_n \end{array}\right].$$

The first step is to produce zeros in the first column of rows 2 through n. Not including the first column, which will become zero, this requires one division (to find the factor c that satisfies $A_{i1} - cA_{11} = 0$), $n-1+1 = n$ multiplications, and n additions per row. Similarly, to produce zeros in the second column of rows 3 through n requires one division, $n-2+1 = n-1$ multiplications, and $n-1$ additions. Producing a zero in the $(n-1)$ th column of row n requires one division, $n-(n-1)+1 = 2$ multiplications, and 2 additions. The elimination step thus requires

$$n\,(n-1) + (n-1)\,(n-2) + (n-2)\,(n-3)\cdots + (n-(n-1))\,(n-(n-2)) = \sum_{i=1}^{n-1} i\,(i+1)$$

multiplications (and the same number of additions). Observe that

$$\sum_{i=1}^{n-1} i\,(i+1) = \sum_{i=1}^{n-1} \left(i^2+i\right) = \sum_{i=1}^{n-1} i^2 + \sum_{i=1}^{n-1} i = \sum_{i=1}^{n-1} i^2 + \frac{(n-1)\,(n-1+1)}{2},$$

where the final equality follows from (9.1.1). Now, for any positive integer n,

$$1^2 + 2^2 + \cdots + n^2 = \sum_{i=1}^{n} i^2 = \frac{n\,(n+1)\,(2n+1)}{6},$$

so it follows that

$$\begin{aligned} \sum_{i=1}^{n-1} i\,(i+1) &= \frac{(n-1)\,(n-1+1)\,[2\,(n-1)+1]}{6} + \frac{(n-1)\,(n-1+1)}{2} \\ &= \frac{n\,(n-1)}{2}\left(\frac{2n-2+1}{3}+1\right) \\ &= \frac{n\,(n-1)}{2}\left(\frac{2n+2}{3}\right) \\ &= n\,(n-1)\,\frac{n+1}{3} \\ &= \frac{1}{3}n\left(n^2-1\right) \\ &= \frac{1}{3}n^3 - \frac{1}{3}n \end{aligned}$$

total multiplications are required. Now, reducing the first column requires $1\,(n-1) = n-1$ divisions, reducing the second column requires $n-2$ divisions, and in general reducing the jth column requires $n-j$ divisions. It follows that

$$\begin{aligned} (n-1) + (n-2) + \cdots + (n-(n-1)) &= \sum_{j=1}^{n-1} (n-j) \\ &= \sum_{j=1}^{n-1} n - \sum_{j=1}^{n-1} j \\ &= n\,(n-1) - \frac{(n-1)\,(n-1+1)}{2} \\ &= n\,(n-1)\left(1-\frac{1}{2}\right) \\ &= \frac{1}{2}n^2 - \frac{1}{2}n \end{aligned}$$

total divisions are required, so that the total operation count for the elimination step is

$$2\left(\frac{1}{3}n^3-\frac{1}{3}n\right)+\left(\frac{1}{2}n^2-\frac{1}{2}n\right)=\frac{2}{3}n^3-\frac{2}{3}n+\frac{1}{2}n^2-\frac{1}{2}n=\frac{2}{3}n^3+\frac{1}{2}n^2-\frac{4}{6}n-\frac{3}{6}n=\frac{2}{3}n^3+\frac{1}{2}n^2-\frac{7}{6}n.$$

The resulting matrix is lower triangular, so we can solve the system by back-substituting, working from the bottom up. The nth row requires one division to solve for x_n. Let a_{ij} be the ijth entry of $\mathbf{A}$ after the elimination step, so that the $(n-1)$ th row is solved by

$$a_{n-1,n-1}x_{n-1}+a_{n-1,n}x_n=b_{n-1}\implies x_{n-1}=\frac{b_{n-1}-a_{n-1,n}x_n}{a_{n-1,n-1}},$$

which requires one multiplication, one addition, and one division. Similarly, the $(n-2)$ th row is solved by

$$x_{n-2}=\frac{b_{n-2}-a_{n-2,n-1}x_{n-1}-a_{n-2,n}x_n}{a_{n-2,n-2}},$$

which requires two multiplications, two additions, and one division, i.e., 5 total operations. Finally, the first row is solved by

$$x_{11}=\frac{b_1-a_{12}x_2-a_{13}x_3+\cdots+a_{1n}x_n}{a_{11}},$$

which requires $n-1$ multiplications, $n-1$ additions, and one division, i.e., $(n-1)+(n-1)+1=2n-1$ total operations. Then, the total operation count for the back-substitution step is

$$1+3+5+\cdots+(2n-1)=\sum_{i=1}^{n}(2i-1)=2\sum_{i=1}^{n}i-\sum_{i=1}^{n}1=2\left(\frac{n\,(n+1)}{2}\right)-n=n^2+n-n=n^2.$$

Noting that this total operation count consists of

$$1+2+3+\cdots+(n-1)=\sum_{i=1}^{n-1}i=\frac{(n-1)\,(n-1+1)}{2}=\frac{1}{2}n\,(n-1)=\frac{1}{2}n^2-\frac{1}{2}n$$

total multiplications, the same number of additions, and n divisions (one for each of the n rows), we see that

$$\left(\frac{1}{2}n^2-\frac{1}{2}n\right)+\left(\frac{1}{2}n^2-\frac{1}{2}n\right)+n=n^2-n+n=n^2.$$

Thus, solving $\mathbf{x}=\mathbf{A}^{-1}\mathbf{b}$ by Gaussian elimination requires

$$\left(\frac{1}{3}n^3-\frac{1}{3}n\right)+\left(\frac{1}{2}n^2-\frac{1}{2}n\right)=\frac{1}{3}n^3-\frac{2}{6}n+\frac{1}{2}n^2-\frac{3}{6}n=\frac{1}{3}n^3+\frac{1}{2}n^2-\frac{5}{6}n$$

total multiplications, the same number of additions, and

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

total divisions, so that the total operation count for the entire procedure is

$$2\left(\frac{1}{3}n^3+\frac{1}{2}n^2-\frac{5}{6}n\right)+\left(\frac{1}{2}n^2+\frac{1}{2}n\right)=\frac{2}{3}n^3+\frac{3}{2}n^2-\frac{7}{6}n,$$

which is of $\mathcal{O}\left(n^3\right)$ complexity. Increasing n by 10, for example, thus increases the number of operations required to perform Gaussian elimination by roughly $10^3=1000$.

17.3. Condition number

Consider again the problem of finding $\mathbf{x}\in\mathbb{R}^n$ that satisfies $\mathbf{A}\mathbf{x}=\mathbf{b}$ for some $n\times n$ matrix $\mathbf{A}$ and n-dimensional vector $\mathbf{b}$, and suppose that $\mathbf{A}\mathbf{x}^\star=\mathbf{b}+\Delta\mathbf{b}$. Let $f:\mathbb{R}^n\to\mathbb{R}^n$ be the function defined by $f(\mathbf{b})=\mathbf{A}^{-1}\mathbf{b}$, i.e., $\mathbf{x}=f(\mathbf{b})$ and $\mathbf{x}^\star=f(\mathbf{b}+\Delta\mathbf{b})$, where $\Delta\mathbf{b}\in\mathbb{R}^n$. We wish to calculate the "derivative" of f,

$$\frac{\|f(\mathbf{b}+\Delta\mathbf{b})-f(\mathbf{b})\|}{\|\Delta\mathbf{b}\|}=\frac{\|\mathbf{x}^\star-\mathbf{x}\|}{\|\Delta\mathbf{b}\|}=\frac{\left\|\mathbf{A}^{-1}(\mathbf{b}+\Delta\mathbf{b})-\mathbf{A}^{-1}\mathbf{b}\right\|}{\|\Delta\mathbf{b}\|}=\frac{\left\|\mathbf{A}^{-1}\Delta\mathbf{b}\right\|}{\|\Delta\mathbf{b}\|}.$$

Now, $\|\Delta\mathbf{b}\|$ is a constant, so that for some $\Delta\mathbf{b} \neq \mathbf{0}$, we have

$$\frac{\left\|\mathbf{A}^{-1}\Delta\mathbf{b}\right\|}{\|\Delta\mathbf{b}\|} = \left\|\frac{\mathbf{A}^{-1}\Delta\mathbf{b}}{\|\Delta\mathbf{b}\|}\right\| = \left\|\mathbf{A}^{-1}\left(\frac{\Delta\mathbf{b}}{\|\Delta\mathbf{b}\|}\right)\right\| \leq \max_{\mathbf{z}\in\mathbb{R}^n, \|\mathbf{z}\|=1} \left\|\mathbf{A}^{-1}\mathbf{z}\right\|,$$

where we have replaced the unit vector $\Delta\mathbf{b}/\|\Delta\mathbf{b}\|$ by $\mathbf{z}$. Thus, the "derivative" of f is the maximum value of the norm of the vector $\mathbf{A}^{-1}\mathbf{z}$ over all n-dimensional unit vectors $\mathbf{z}$, or informally, how far $\mathbf{A}^{-1}$ "stretches" the unit ball defined by $\mathbf{z} \in \mathbb{R}^n$.

DEFINITION 17.6. Suppose that $\mathbf{A} \in \mathbf{M}_{m,n}(\mathbb{R})$. Then, the *norm* of $\mathbf{A}$, written $\|\mathbf{A}\|$, is given by

$$\|\mathbf{A}\| = \max_{\mathbf{z}\in\mathbb{R}^n, \|\mathbf{z}\|=1} \|\mathbf{A}\mathbf{z}\|.$$

We can thus express the "derivative" of f as

$$\frac{\|\mathbf{x}^\star - \mathbf{x}\|}{\|\Delta\mathbf{b}\|} \leq \left\|\mathbf{A}^{-1}\right\|.$$

Now suppose that $\mathbf{A}$ is symmetric, and consider calculating $\|\mathbf{A}\|$. Theorem 16.9 implies that $\mathbf{A}$ has an orthonormal basis of eigenvectors, which we denote by $\left\{\mathbf{q}^{(i)}\right\}_{i=1}^n$, with corresponding eigenvalues $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. Noting that we can express any $\{\mathbf{z} : \mathbf{z} \in \mathbb{R}^n, \|\mathbf{z}\| = 1\}$ in the $\mathbf{q}$-basis, i.e.,

$$\mathbf{z} = c_1\mathbf{q}^{(1)} + c_2\mathbf{q}^{(2)} + \cdots + c_n\mathbf{q}^{(n)} = \sum_{i=1}^n c_i\mathbf{q}^{(i)},$$

it follows that

$$\|\mathbf{z}\|^2 = \mathbf{z}\cdot\mathbf{z} = \left(\sum_{i=1}^n c_i\mathbf{q}^{(i)}\right)\cdot\left(\sum_{j=1}^n c_j\mathbf{q}^{(j)}\right) = \left(\sum_{i=1}^n c_i\mathbf{q}^{(i)}\right)\cdot c_1\mathbf{q}^{(1)} + \cdots + \left(\sum_{i=1}^n c_i\mathbf{q}^{(i)}\right)\cdot c_n\mathbf{q}^{(n)}.$$

The expansion of the kth term in this sum is another sum whose summands are of the form $c_ic_k\left(\mathbf{q}^{(i)}\cdot\mathbf{q}^{(k)}\right)$. Because the $\mathbf{q}^{(i)}$ are orthonormal, these summands will be equal to one in the case that $i = k$ and zero otherwise, i.e.,

$$c_ic_k\left(\mathbf{q}^{(i)}\cdot\mathbf{q}^{(k)}\right) = \begin{cases} 1, & \text{if } i = k \\ 0, & \text{otherwise} \end{cases},$$

and it follows that

$$\|\mathbf{z}\|^2 = c_1\cdot c_1\cdot 1 + c_2\cdot c_2\cdot 1 + \cdots + c_n\cdot c_n\cdot 1 = \sum_{i=1}^n c_i^2.$$

We have $\|\mathbf{z}\| = 1$, which implies that $\|\mathbf{z}\|^2 = 1$, hence that $\sum_{i=1}^n c_i^2 = 1$. Then,

$$\mathbf{A}\mathbf{z} = \mathbf{A}\left(\sum_{i=1}^n c_i\mathbf{q}^{(i)}\right) = \sum_{i=1}^n c_i\mathbf{A}\mathbf{q}^{(i)} = \sum_{i=1}^n c_i\lambda_i\mathbf{q}^{(i)}.$$

Now, if a unit-norm vector $\mathbf{z}^\star \in \mathbb{R}^n$ maximizes $\|\mathbf{A}\mathbf{z}\|$, i.e., if

$$\mathbf{z}^\star = \operatorname*{arg\,max}_{\mathbf{z}\in\mathbb{R}^n, \|\mathbf{z}\|=1} \|\mathbf{A}\mathbf{z}\|,$$

then $\mathbf{z}^\star$ will also maximize $\|\mathbf{A}\mathbf{z}\|^2$. We can then consider maximizing

$$\|\mathbf{A}\mathbf{z}\|^2 = (\mathbf{A}\mathbf{z})\cdot(\mathbf{A}\mathbf{z}) = \left(\sum_{i=1}^n c_i\lambda_i\mathbf{q}^{(i)}\right)\cdot\left(\sum_{j=1}^n c_j\lambda_j\mathbf{q}^{(j)}\right).$$

Reasoning as above, i.e., exploiting the orthonormality of the $\mathbf{q}^{(i)}$, this expression becomes

$$\|\mathbf{A}\mathbf{z}\|^2 = \sum_{i=1}^n (c_i\lambda_i)^2 = \sum_{i=1}^n c_i^2\lambda_i^2.$$

We can thus recast calculating the norm of $\mathbf{A}$ as a constrained optimization, i.e., as finding the $\mathbf{z}$ that maximizes $\|\mathbf{A}\mathbf{z}\|^2$ subject to $\sum_{i=1}^n c_i^2 = 1$. Recalling that λ_k is eigenvalue of $\mathbf{A}$ with the kth largest absolute value, it follows that to maximize $\|\mathbf{A}\mathbf{z}\|^2$, we should put all the "weight" into c_1, i.e., set $c_1 = 1$. Thus, $\|\mathbf{A}\mathbf{z}\|^2$

is maximized when $c_1 = 1$, and in this case $\|\mathbf{Az}\|^2 = \lambda_1^2$. When we set $c_1 = 1$, we set all the remaining weights c_i to zero, so that

$$\mathbf{Az} = \sum_{i=1}^{n} c_i \lambda_i \mathbf{q}^{(i)} = 1 \cdot \lambda_1 \mathbf{q}^{(1)} + 0 + \cdots + 0 = \lambda_1 \mathbf{q}^{(1)} \implies \|\mathbf{Az}\|^2 = \left\|\lambda_1 \mathbf{q}^{(1)}\right\|^2 = \left\|\mathbf{A}\mathbf{q}^{(1)}\right\|^2 \implies \mathbf{z} = \mathbf{q}^{(1)},$$

where we have used the fact that λ_1 is an eigenvalue of $\mathbf{A}$ with corresponding eigenvector $\mathbf{q}^{(1)}$. Thus, the norm of $\mathbf{A}$ is

$$\|\mathbf{A}\| = \max_{\mathbf{z}\in\mathbb{R}^n, \|\mathbf{z}\|=1} \|\mathbf{Az}\| = \max_{\mathbf{z}\in\mathbb{R}^n, \|\mathbf{z}\|=1} \left\|\lambda_1 \mathbf{q}^{(1)}\right\| = \left\|\lambda_1 \mathbf{q}^{(1)}\right\| = \lambda_1 \left\|\mathbf{q}^{(1)}\right\| = \lambda_1 \cdot 1 = \lambda_1.$$

PROPOSITION 17.7. *The norm of a real $n \times n$ matrix $\mathbf{A}$ with eigenvalues $|\lambda_1| > |\lambda_2| > \cdots |\lambda_n|$ is $\|\mathbf{A}\| = \lambda_1$.*

PROOF. [proof goes here] □

In practice, our interest will typically be in the *relative* error. We will examine the ratio of the *relative forward error* $\|\mathbf{x}^\star - \mathbf{x}\| / \|\mathbf{x}\|$ to the *relative backward error* $\|\Delta\mathbf{b}\| / \|\mathbf{b}\|$, where "forward" refers to solving for $\mathbf{x}$, i.e., the error in the solution, and "backward" refers to the fact that $\mathbf{b}$ is an input. We have

$$\begin{aligned}
\frac{\|\mathbf{x}^\star - \mathbf{x}\| / \|\mathbf{x}\|}{\|\Delta\mathbf{b}\| / \|\mathbf{b}\|} &= \frac{\|\mathbf{x}^\star - \mathbf{x}\| / \|\Delta\mathbf{b}\|}{\|\mathbf{x}\| / \|\mathbf{b}\|} \\
&= \frac{\left\|\mathbf{A}^{-1}\Delta\mathbf{b}\right\| / \|\Delta\mathbf{b}\|}{\|\mathbf{x}\| / \|\mathbf{Ax}\|} \\
&= \frac{\left\|\mathbf{A}^{-1}\Delta\mathbf{b}\right\|}{\|\Delta\mathbf{b}\|} \cdot \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \\
&= \left\|\mathbf{A}^{-1}\frac{\Delta\mathbf{b}}{\|\Delta\mathbf{b}\|}\right\| \cdot \left\|\mathbf{A}\frac{\mathbf{x}}{\|\mathbf{x}\|}\right\| \\
&\leq \left(\max_{\mathbf{z}\in\mathbb{R}^n, \|\mathbf{z}\|=1} \left\|\mathbf{A}^{-1}\mathbf{z}\right\|\right)\left(\max_{\mathbf{w}\in\mathbb{R}^n, \|\mathbf{w}\|=1} \|\mathbf{Aw}\|\right) \\
&= \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\|,
\end{aligned}$$

where the penultimate inequality follows from two applications of our result above. Thus, the ratio of the relative forward error to the relative backward error is bounded by the product of the norms of $\mathbf{A}$ and $\mathbf{A}^{-1}$.

PROPOSITION 17.8. *Let $\mathbf{M}$ be a real, invertible $n \times n$ matrix, and let λ be an eigenvalue of $\mathbf{M}$ with corresponding eigenvector $\mathbf{v}$. Then, $1/\lambda$ is an eigenvalue of $\mathbf{M}^{-1}$ with corresponding eigenvector $\mathbf{v}$.*

PROOF. We have

$$\mathbf{Mv} = \lambda\mathbf{v} \implies \mathbf{M}^{-1}\mathbf{Mv} = \mathbf{M}^{-1}\lambda\mathbf{v} \implies \mathbf{I}_n\mathbf{v} = \lambda\mathbf{M}^{-1}\mathbf{v} \implies \mathbf{M}^{-1}\mathbf{v} = \frac{1}{\lambda}\mathbf{v}.$$

□

Suppose $\mathbf{A}$ is symmetric with eigenvalues $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. Then, proposition 17.8 implies that $\mathbf{A}^{-1}$ has eigenvalues $|1/\lambda_n| > |1/\lambda_{n-1}| > \cdots > |1/\lambda_1|$, so that

$$\frac{\|\mathbf{x}^\star - \mathbf{x}\| / \|\mathbf{x}\|}{\|\Delta\mathbf{b}\| / \|\mathbf{b}\|} \leq \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| = \frac{1}{\lambda_n} \cdot \lambda_1 = \frac{\lambda_1}{\lambda_n},$$

i.e., the ratio of the relative errors is bounded by the ratio of the largest eigenvalue of $\mathbf{A}$ to the smallest.

DEFINITION 17.9. Let $\mathbf{A}$ be a real, invertible $n \times n$ matrix $\mathbf{A}$ with eigenvalues $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$. The *condition number* of $\mathbf{A}$, written $\kappa(\mathbf{A})$, is given by

$$\kappa(\mathbf{A}) = \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| = \frac{\lambda_1}{\lambda_n}.$$

We can now express our previous result for the ratio of the relative errors in terms of the condition number of $\mathbf{A}$, i.e.,

$$\frac{\|\mathbf{x}^\star - \mathbf{x}\| / \|\mathbf{x}\|}{\|\Delta\mathbf{b}\| / \|\mathbf{b}\|} \leq \left\|\mathbf{A}^{-1}\right\| \|\mathbf{A}\| = \kappa(\mathbf{A}) \implies \frac{\|\mathbf{x}^\star - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \cdot \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

The condition number quantifies the sensitivity of the solution $\mathbf{x}$ to changes in $\mathbf{b}$, or how much "damage" $\mathbf{A}$ can do given a small change in $\mathbf{b}$. The condition number is thus, in a sense, a derivative.

Intuitively, $\Delta\mathbf{b}$ "smears" $\mathbf{b}$, and the question arises of the optimal direction in which to smear $\mathbf{b}$ so that the solutions $\mathbf{x}$ and $\mathbf{x}^\star$ of $\mathbf{A}\mathbf{x} = \mathbf{b}$ and $\mathbf{A}\mathbf{x}^\star = \mathbf{b} + \Delta\mathbf{b}$, respectively, are maximally different. Take $\mathbf{b} = \mathbf{q}^{(1)}$, so that

$$\mathbf{A}\mathbf{x} = \mathbf{q}^{(1)} \implies \mathbf{x} = \mathbf{A}^{-1}\mathbf{q}^{(1)} = \frac{1}{\lambda_1}\mathbf{q}^{(1)}.$$

We see that, having taken $\mathbf{b} = \mathbf{q}^{(1)}$, which has unit length, we divide by the largest eigenvalue λ_1, so we have "shrunk" $\mathbf{x}$ as much as possible. Now let $\Delta\mathbf{b} = \varepsilon\mathbf{q}^{(n)}$, so that

$$\mathbf{A}\mathbf{x}^\star = \mathbf{q}^{(1)} + \varepsilon\mathbf{q}^{(n)} \implies \mathbf{x}^\star = \mathbf{A}^{-1}\left(\mathbf{q}^{(1)} + \varepsilon\mathbf{q}^{(n)}\right) = \mathbf{A}^{-1}\mathbf{q}^{(1)} + \varepsilon\mathbf{A}^{-1}\mathbf{q}^{(n)} = \frac{1}{\lambda_1}\mathbf{q}^{(1)} + \frac{\varepsilon}{\lambda_n}\mathbf{q}^{(n)}.$$

Thus, we have taken $\mathbf{b}$, which "lived" along $\mathbf{q}^{(1)}$, and smeared it such that it now has a component along $\mathbf{q}^{(n)}$. We now examine the ratio of the relative errors

$$\begin{aligned}
\frac{\|\mathbf{x}^\star - \mathbf{x}\| / \|\mathbf{x}\|}{\|\Delta\mathbf{b}\| / \|\mathbf{b}\|} &= \frac{\left\|\left(\mathbf{q}^{(1)}/\lambda_1 + \varepsilon\mathbf{q}^{(n)}/\lambda_n\right) - \mathbf{q}^{(1)}/\lambda_1\right\| / \left\|\mathbf{q}^{(1)}/\lambda_1\right\|}{\left\|\varepsilon\mathbf{q}^{(n)}\right\| / \left\|\mathbf{q}^{(1)}\right\|} \\
&= \frac{(\varepsilon/\lambda_n)\left\|\mathbf{q}^{(n)}\right\| / (1/\lambda_1)\left\|\mathbf{q}^{(1)}\right\|}{\varepsilon\left\|\mathbf{q}^{(n)}\right\| / \left\|\mathbf{q}^{(1)}\right\|} \\
&= \frac{(\varepsilon/\lambda_n)\cdot 1/(1/\lambda_1)\cdot 1}{\varepsilon\cdot 1/1} \\
&= \frac{\lambda_1}{\lambda_n}.
\end{aligned}$$

We see that our result is precisely the condition number of $\mathbf{A}$, i.e., the maximum "damage" is indeed inflicted by taking $\Delta\mathbf{b} = \varepsilon\mathbf{q}^{(n)}$. If $\mathbf{A}$ has eigenvectors on very different scales, then $\kappa(\mathbf{A})$ will be very large. We now consider the implications of roundoff error. On a computer, we are solving not $\mathbf{A}\mathbf{x} = \mathbf{b}$, but $\mathbf{A}\mathbf{x}^\star = \mathrm{fl}(\mathbf{b})$ (and leaving aside for the moment the fact that we have not $\mathbf{A}$, but its floating-point representation as well). Now, $\mathrm{fl}(\mathbf{b})$ will be equal to $\mathbf{b}$ plus a term on the order of $\epsilon_{\text{mach}} \approx 10^{-16}$ times $\Delta\mathbf{b}$, i.e., $\mathrm{fl}(\mathbf{b}) = \mathbf{b} + \epsilon_{\text{mach}}\Delta\mathbf{b}$. Accouting for floating-point error, our result above becomes

$$\frac{\|\mathbf{x}^\star - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \cdot \frac{\|\epsilon_{\text{mach}}\Delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Now, the quantity $\|\Delta\mathbf{b}\| / \|\mathbf{b}\|$ will be some constant, which for the moment we ignore, leaving only machine epsilon, so that the relative error due to roundoff is

$$\frac{\|\mathbf{x}^\star - \mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \cdot \epsilon_{\text{mach}}.$$

It is clear that if $\kappa(\mathbf{A}) = 10^{20}$, for example, then $\kappa(\mathbf{A}) \cdot \epsilon_{\text{mach}} = 10^{20} \cdot 10^{-16} = 10^4$, so that the relative error will be large, and the results of such a computation compromised by roundoff error.

Now consider a one-dimensional example, i.e., $x^\star, x \in \mathbb{R}$. Then, in a "worst-case" scenario, the relative error will be

$$|x^\star - x| = \kappa(\mathbf{A}) \cdot \epsilon_{\text{mach}} |x| \implies x^\star = x\left(1 \pm \kappa(\mathbf{A}) \cdot \epsilon_{\text{mach}}\right).$$

If $\kappa(\mathbf{A}) = 10^d$ and taking $\epsilon_{\text{mach}} = 10^{-16}$, we have $x^\star = x\left(1 \pm 10^d \cdot 10^{-16}\right) = x\left(1 \pm 10^{d-16}\right)$, so that x will be "shifted" by $d - 16$ digits. For example, if $x = 1234567$ and $\kappa(\mathbf{A}) = 10^{14}$, then we will have

$$x^\star = x\left(1 + 10^{-2}\right) = 1234567 + 12345.67 = 1246912.67,$$

i.e., we retain only 2 significant digits of x (the remaining digits are incorrect). (We obtain the same result when we consider $x^\star = x\left(1 - 10^{-2}\right)$, i.e., we retain only two correct digits.) In general, we will retain $\min(|d - 16|, 0)$ significant digits, so that in the case that $d \geq 16$, we will not retain *any* correct digits. Consider the following implications:

(1) *Solving systems of the form* $\mathbf{Ax} = \mathbf{b}$. For example, when we compute a direction in Newton's Method, we are solving $[\mathbf{H}_f(\mathbf{x})]\,\mathbf{d} = -\nabla f(\mathbf{x})$. If the Hessian $\mathbf{H}_f(\mathbf{x})$ has high condition number, this computation may fail.
(2) *Fitting a linear regression by solving the normal equations.* We fit a linear regression by solving the normal equations $\mathbf{B}^\mathsf{T}\mathbf{B}\boldsymbol{\alpha} = \mathbf{B}^\mathsf{T}\mathbf{y}$ for $\boldsymbol{\alpha}$. The matrix $\mathbf{B}^\mathsf{T}\mathbf{B}$ tends to have the square of the condition number of $\mathbf{B}$, which may be high.
(3) *Fitting a linear regression given observations* $\mathbf{y}$. When we solve the normal equations, our observations $\mathbf{y}$ are actually $\mathbf{y}$ plus some noise term $\Delta\mathbf{y}$, which is on the order of 10^{-2} for the NBA data set.

The normal equations tend to be badly conditioned because $\kappa\left(\mathbf{B}^\mathsf{T}\mathbf{B}\right)$ is often high.

17.4. Projections

We now consider fitting a linear regression in the case that solution via the normal equations is impractical, e.g., due to the high condition number of $\mathbf{B}^\mathsf{T}\mathbf{B}$.

DEFINITION 17.10. Let Ω be a set in $\mathbb{R}^n$, and suppose that $\mathbf{x} \in \mathbb{R}^n$. Then, the *projection* of $\mathbf{x}$ onto Ω, written $\text{proj}_\Omega(\mathbf{x})$, is defined by

$$\text{proj}_\Omega(\mathbf{x}) = \underset{\mathbf{z}\in\Omega}{\arg\min} \|\mathbf{z} - \mathbf{x}\|.$$

Informally, the projection of $\mathbf{x}$ onto Ω is the vector $\mathbf{z} \in \Omega$ closest to $\mathbf{x}$. Let $\left\{\mathbf{v}^{(i)}\right\}_{i=1}^k \in \mathbb{R}^n$, and let Ω be the span of the $\mathbf{v}^{(i)}$. We wish to find $\text{proj}_\Omega(\mathbf{x})$. Because the $\mathbf{v}^{(i)}$ span Ω, we can write any $\mathbf{z} \in \Omega$ as a linear combination of the $\mathbf{v}^{(i)}$, i.e.,

$$\mathbf{z} = \alpha_1\mathbf{v}^{(1)} + \alpha_2\mathbf{v}^{(2)} + \cdots + \alpha_k\mathbf{v}^{(k)} = \sum_{i=1}^k \alpha_i\mathbf{v}^{(i)},$$

so that we can rewrite the projection as the equivalent problem of finding the $\boldsymbol{\alpha}$ that solves

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\arg\min} \left\|\mathbf{x} - \sum_{i=1}^k \alpha_i\mathbf{v}^{(i)}\right\|.$$

Now, any $\boldsymbol{\alpha}$ that minimizes this quantity will also minimize its square, so that

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\arg\min} \left\|\mathbf{x} - \sum_{i=1}^k \alpha_i\mathbf{v}^{(i)}\right\|^2.$$

Let $\mathbf{V}$ be the matrix whose jth column is given by $\mathbf{v}^{(j)}$, so that

$$\sum_{i=1}^k \alpha_i\mathbf{v}^{(i)} = \begin{bmatrix}\mathbf{v}^{(1)} & \mathbf{v}^{(2)} & \cdots & \mathbf{v}^{(k)}\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\\ \vdots\\ \alpha_k\end{bmatrix} = \mathbf{V}\boldsymbol{\alpha} \implies \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}\in\mathbb{R}^k}{\arg\min} \|\mathbf{x} - \mathbf{V}\boldsymbol{\alpha}\|^2.$$

We recognize this expression as having the same form as the least-squares loss function, and it follows that this quantity is minimized by $\boldsymbol{\alpha} = \left(\mathbf{V}^\mathsf{T}\mathbf{V}\right)^{-1}\mathbf{V}^\mathsf{T}\mathbf{x}$. Thus, linear regression is simply a projection of the observations $\mathbf{y}$ onto the basis (the span of the columns) of the model matrix $\mathbf{B}$ (and observe that we can replace the particular vectors that form $\mathbf{B}$ with a collection of vectors having the same span without affecting the projection). The normal equations are then seen to be badly conditioned because $\mathbf{B}$ is in general a "bad" basis, i.e., its columns are "close" to being linearly dependent. Suppose that we have just two columns, i.e., $\left\{\mathbf{v}^{(i)}\right\}_{i=1}^2$. If $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ are close, then $\kappa(\mathbf{V})$ will be high. If $\mathbf{v}^{(1)}$ and $\mathbf{v}^{(2)}$ are orthogonal, then $\kappa(\mathbf{V})$ will be low, so that the best possible basis consists of mutually orthogonal vectors. In particular, suppose that the model matrix admits the decomposition $\mathbf{B} = \mathbf{QR}$.

DEFINITION 17.11. Let $\mathbf{A} \in \mathbf{M}_{m,n}(\mathbb{R})$. Then, there exists an orthonormal matrix $\mathbf{Q}$ and an invertible, upper triangular matrix $\mathbf{R}$ such that $\mathbf{A} = \mathbf{QR}$, and the span of the columns of $\mathbf{A}$ is equal to the span of the columns of $\mathbf{Q}$. This decomposition is known as the $\mathbf{QR}$ *decomposition.*

If we can compute the $\mathbf{QR}$ decomposition of the model matrix $\mathbf{B}$, then we can replace a possibly badly conditioned matrix with a matrix having much better condition number (because the $\mathbf{Q}$ matrix consists of orthogonal columns), i.e.,

$$\begin{aligned}
\boldsymbol{\alpha} &= \left(\mathbf{B}^\mathsf{T}\mathbf{B}\right)^{-1}\mathbf{B}^\mathsf{T}\mathbf{y} \\
&= \left(\left(\mathbf{QR}\right)^\mathsf{T}\mathbf{QR}\right)^{-1}\left(\mathbf{QR}\right)^\mathsf{T}\mathbf{y} \\
&= \left(\mathbf{R}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{QR}\right)^{-1}\mathbf{R}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{y} \\
\text{($\mathbf{Q}$ is orthogonal)} \qquad &= \left(\mathbf{R}^\mathsf{T}\mathbf{Q}^{-1}\mathbf{QR}\right)^{-1}\mathbf{R}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{y} \\
&= \left(\mathbf{R}^\mathsf{T}\mathbf{R}\right)^{-1}\mathbf{R}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{y} \\
&= \mathbf{R}^{-1}\left(\mathbf{R}^\mathsf{T}\right)^{-1}\mathbf{R}^\mathsf{T}\mathbf{Q}^\mathsf{T}\mathbf{y} \\
&= \mathbf{R}^{-1}\mathbf{Q}^\mathsf{T}\mathbf{y}.
\end{aligned}$$

We can interpret the product $\mathbf{Q}^\mathsf{T}\mathbf{y}$ as projecting $\mathbf{y}$ onto the span of the columns of $\mathbf{Q}$, and the action of $\mathbf{R}^{-1}$ as returning this product to the $\mathbf{B}$-basis. In general, $\mathbf{R}^{-1}$ is usually much better conditioned than $\mathbf{B}^\mathsf{T}\mathbf{B}$.

We continue to consider the problem of fitting a linear regression, i.e., solving the optimization

$$\hat{\boldsymbol{\alpha}} = \arg\min_{\boldsymbol{\alpha}} \left\|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\right\|_2,$$

where $\mathbf{y}$ is a vector of observations and $\mathbf{B}$ is the model matrix. We can view the product $\mathbf{B}\boldsymbol{\alpha}$ as linearly combining a basis, i.e., the columns of $\mathbf{B}$. Recall that $\mathbf{B}$ is often a "bad" basis in the sense that it may have columns that are close to being (or actually are) linearly dependent, which occurs often when collecting data. The $\mathbf{QR}$ decomposition may allow us to fit the regression by finding a matrix $\mathbf{Q}$ such that the span of the columns of $\mathbf{Q}$ is equal to the span of the columns of $\mathbf{B}$. We then project $\mathbf{y}$ onto $\mathbf{Q}$ by solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\|\mathbf{y} - \mathbf{Q}\boldsymbol{\beta}\right\|_2^2,$$

which is solved by $\boldsymbol{\beta} = \left(\mathbf{Q}^\mathsf{T}\mathbf{Q}\right)^{-1}\mathbf{Q}^\mathsf{T}\mathbf{y} = \left(\mathbf{Q}^{-1}\mathbf{Q}\right)^{-1}\mathbf{Q}^\mathsf{T}\mathbf{y} = \mathbf{Q}^\mathsf{T}\mathbf{y}$. Observe that the product $\mathbf{Q}\boldsymbol{\beta}$ can be viewed as linearly combining the columns of $\mathbf{Q}$, which have the same span as the columns of $\mathbf{B}$, so that we can write any $\mathbf{z} \in \operatorname{span}\{\operatorname{col}(\mathbf{B})\}$ as a linear combination of the columns of $\mathbf{Q}$, i.e., $\mathbf{z} = \mathbf{Q}\boldsymbol{\beta}$. We also have $\mathbf{z} = \mathbf{B}\boldsymbol{\alpha}$, so it follows that $\mathbf{Q}\boldsymbol{\beta} = \mathbf{B}\boldsymbol{\alpha}$. If we fit the linear regression in the $\mathbf{Q}$-basis, we must then convert the coefficients $\boldsymbol{\beta}$ back to the $\mathbf{B}$-basis, which is accomplished as

$$\mathbf{B}\boldsymbol{\alpha} = \mathbf{Q}\boldsymbol{\beta} \implies \boldsymbol{\alpha} = \mathbf{B}^{-1}\mathbf{Q}\boldsymbol{\beta} = \left(\mathbf{QR}\right)^{-1}\mathbf{Q}\boldsymbol{\beta} = \mathbf{R}^{-1}\mathbf{Q}^{-1}\mathbf{Q}\boldsymbol{\beta} = \mathbf{R}^{-1}\boldsymbol{\beta},$$

where we have replaced $\mathbf{B}$ by its $\mathbf{QR}$ decomposition. Now, $\mathbf{Q}$ is orthonormal and $\mathbf{R}$ is upper triangular and invertible, so that the decomposition provides both computational stability (through a possibly lower condition number) and performance (due to the upper triangular shape of $\mathbf{R}$). We now consider the problem of finding the decomposition. Letting $\left\{\mathbf{v}^{(i)}\right\}_{i=1}^k \in \mathbb{R}^n$, we wish to find $\left\{\mathbf{q}^{(i)}\right\}_{i=1}^k \in \mathbb{R}^n$ such that the $\mathbf{q}^{(i)}$ are orthonormal and

$$\operatorname{span}\left\{\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, \mathbf{v}^{(\ell)}\right\} = \operatorname{span}\left\{\mathbf{q}^{(1)}, \mathbf{q}^{(2)}, \ldots, \mathbf{q}^{(\ell)}\right\}, \quad \ell \in \{1, 2, \ldots, k\}.$$

We can find the $\mathbf{q}^{(i)}$ through Gram-Schmidt orthogonalization.

(1) Set $\mathbf{q}^{(1)} = \mathbf{v}^{(1)}$.
(2) Normalize $\mathbf{q}^{(1)}$, i.e., set $\mathbf{q}^{(1)} = \mathbf{q}^{(1)} / \left\|\mathbf{q}^{(1)}\right\|_2$.
(3) To produce $\mathbf{q}^{(2)}$, we will subtract from $\mathbf{v}^{(2)}$ its projection onto $\mathbf{q}^{(1)}$, i.e., we will set $\mathbf{q}^{(2)} = \mathbf{v}^{(2)} - \left(\mathbf{v}^{(2)} \cdot \mathbf{q}^{(1)}\right)\mathbf{q}^{(1)}$.
(4) Normalize $\mathbf{q}^{(2)}$, i.e., set $\mathbf{q}^{(2)} = \mathbf{q}^{(2)} / \left\|\mathbf{q}^{(2)}\right\|_2$.
(5) To produce $\mathbf{q}^{(3)}$, we will subtract from $\mathbf{v}^{(3)}$ its projections onto $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$, i.e., we will set $\mathbf{q}^{(3)} = \mathbf{v}^{(3)} - \left(\mathbf{v}^{(3)} \cdot \mathbf{q}^{(1)}\right)\mathbf{q}^{(1)} - \left(\mathbf{v}^{(3)} \cdot \mathbf{q}^{(2)}\right)\mathbf{q}^{(2)}$.
(6) Normalize $\mathbf{q}^{(3)}$, i.e., set $\mathbf{q}^{(3)} = \mathbf{q}^{(3)} / \left\|\mathbf{q}^{(3)}\right\|_2$.
(7) Continue in this fashion until we construct $\mathbf{q}^{(k)}$.

Algorithm 9 provides a formal statement of the Gram-Schmidt iteration.

Algorithm 9 Gram-Schmidt Orthogonalization

Require: $\left\{\mathbf{v}^{(i)}\right\}_{i=1}^{k} \in \mathbb{R}^n$
1: **for** $i = 1, 2, \ldots, k$ **do**
2: $\quad \mathbf{w}^{(i)} \leftarrow \mathbf{v}^{(i)}$
3: $\quad$ **for** $j = 1, 2, \ldots, i-1$ **do**
4: $\quad\quad \mathbf{w}^{(i)} \leftarrow \mathbf{w}^{(i)} - \left(\mathbf{v}^{(i)} \cdot \mathbf{q}^{(j)}\right) \mathbf{q}^{(j)}$
5: $\quad$ **end for**
6: $\quad \mathbf{q}^{(i)} \leftarrow \mathbf{w}^{(i)} / \left\|\mathbf{w}^{(i)}\right\|_2$
7: **end for**
$\quad$ **return** $\left\{\mathbf{q}^{(i)}\right\}_{i=1}^{k}$

We now consider an intuitive argument for why Gram-Schmidt produces an orthonormal basis. Suppose that we construct $\mathbf{q}^{(1)} = \mathbf{v}^{(1)} / \left\|\mathbf{v}^{(1)}\right\|_2$. We will construct $\mathbf{q}^{(2)}$ by setting it equal to the difference of $\mathbf{v}^{(2)}$ and the projection of $\mathbf{v}^{(2)}$ onto $\mathbf{q}^{(1)}$. Set $\mathbf{u}$ to be this projection, i.e., $\mathbf{u} = \left(\mathbf{v}^{(2)} \cdot \mathbf{q}^{(1)}\right) \mathbf{q}^{(1)}$, subtract $\mathbf{u}$ from $\mathbf{v}^{(2)}$, and normalize the result to produce $\mathbf{q}^{(2)}$. Figure 17.4.1 depicts the initial steps of the iteration.

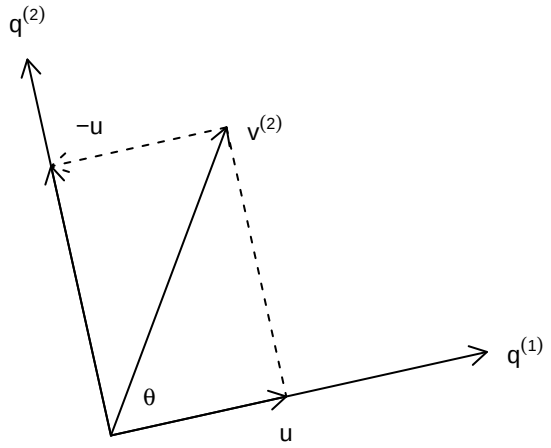


FIGURE 17.4.1. Example of Gram-Schmidt Orthogonalization

Clearly $\mathbf{u}$ lies along $\mathbf{q}^{(1)}$, so we will have $\mathbf{u} = c\mathbf{q}^{(1)}$ for some $c \in \mathbb{R}$. Observing that the projection of $\mathbf{v}^{(2)}$ onto $\mathbf{q}^{(1)}$ forms a right angle with $\mathbf{q}^{(1)}$, and letting θ be the angle between $\mathbf{u}$ and $\mathbf{v}^{(2)}$, we see that $c = \|\mathbf{u}\|_2$, and that $\cos\theta = \|\mathbf{u}\|_2 / \left\|\mathbf{v}^{(2)}\right\|_2 \implies c = \left\|\mathbf{v}^{(2)}\right\|_2 \cos\theta$. Thus, $\mathbf{u} = \left(\left\|\mathbf{v}^{(2)}\right\|_2 \cos\theta\right) \mathbf{q}^{(1)}$. Now, the dot product of $\mathbf{v}^{(2)}$ and $\mathbf{q}^{(1)}$ can be written as

$$\mathbf{v}^{(2)} \cdot \mathbf{q}^{(1)} = \left\|\mathbf{v}^{(2)}\right\|_2 \left\|\mathbf{q}^{(1)}\right\|_2 \cos\theta = \left\|\mathbf{v}^{(2)}\right\|_2 \cdot 1 \cdot \cos\theta = \left\|\mathbf{v}^{(2)}\right\|_2 \cos\theta,$$

where the penultimate equality follows because $\mathbf{q}^{(1)}$ has unit norm. Then, we see that $\left\|\mathbf{v}^{(2)}\right\|_2 \cos\theta = \mathbf{v}^{(2)} \cdot \mathbf{q}^{(1)}$, as given in algorithm 9.

Now, $\mathbf{q}^{(1)}$ is parallel to $\mathbf{v}^{(1)}$, hence has the same span. Noting that in our construction $\mathbf{q}^{(2)} = \mathbf{v}^{(2)} - \left(\mathbf{v}^{(2)} \cdot \mathbf{q}^{(1)}\right) \mathbf{q}^{(1)}$, the term $\mathbf{v}^{(2)} \cdot \mathbf{q}^{(1)}$ is a constant, it is clear that $\mathbf{q}^{(2)}$ is a linear combination of $\mathbf{v}^{(2)}$ and $\mathbf{q}^{(1)}$, $\mathbf{q}^{(2)} \in \operatorname{span}\left\{\mathbf{v}^{(2)}, \mathbf{q}^{(1)}\right\}$. We have shown that $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$ are orthogonal, hence linearly independent,

so it follows that the span of $\mathbf{q}^{(1)}$ and $\mathbf{q}^{(2)}$ must be the same as the span of $\mathbf{v}^{(2)}$ and $\mathbf{q}^{(1)}$. We can argue in a similar fashion for the remaining $\mathbf{q}^{(i)}$.

We can also consider reconstructing the $\mathbf{v}^{(i)}$ from the $\mathbf{q}^{(i)}$. We have

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & \cdots & \mathbf{v}^{(k)} \end{bmatrix} = \begin{bmatrix} \mathbf{q}^{(1)} & \mathbf{q}^{(2)} & \cdots & \mathbf{q}^{(k)} \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1k} \\ 0 & r_{22} & \cdots & r_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{kk} \end{bmatrix} = \mathbf{QR},$$

where we recall that $\mathbf{R} \in \mathbf{M}_{k,k}(\mathbb{R})$ is upper triangular. Now, the first column of $\mathbf{V}$, i.e., $\mathbf{v}^{(1)}$, is equal to the first column of $\mathbf{QR}$, which is simply $r_{11}\mathbf{q}^{(1)}$, and the second column is given by $r_{12}\mathbf{q}^{(1)} + r_{22}\mathbf{q}^{(2)}$. So, we see that to reconstruct the jth column of $\mathbf{V}$, we require the first j columns of $\mathbf{QR}$, which in turn requires the vectors $\left\{\mathbf{q}^{(i)}\right\}_{i=1}^{j}$. Finally, we observe that $\mathbf{Q}$ will be invertible so long as its diagonal entries are non-zero, which will be true provided the $\mathbf{v}^{(i)}$ are not linearly dependent.

17.5. Numerical differentiation

Let f be a real-valued function of $x \in \mathbb{R}$, and consider the problem of computing the first and second derivatives of f, $f'(x)$ and $f''(x)$. More generally, if f is a real-valued function of $\mathbf{x} \in \mathbb{R}^d$, we can compute the gradient and Hessian of f, $\nabla f(\mathbf{x})$ and $\mathbf{H}_f(\mathbf{x})$, by computing the derivatives in one dimension at a time. We will compute such a derivative by the method of *finite differences*. Observe that

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \approx \frac{f(x+h) - f(x)}{h},$$

where the final term is called the *forward finite difference* (forward because we are evaluating f and $x+h$ for some $h > 0$). The error associated with the forward finite difference is

$$E_{\text{fwd}}(h) = \left| f'(x) - \frac{f(x+h) - f(x)}{h} \right|.$$

A third-order Taylor series for $f(x+h)$ around the base point x is

$$f(x+h) \approx f(x) + f'(x)(x+h-x) + \frac{1}{2!} f''(x)(x+h-x)^2 + \mathcal{O}\left(h^3\right), \tag{17.5.1}$$

so that

$$E_{\text{fwd}}(h) = \left| f'(x) - \frac{f(x) + hf'(x) + h^2 f''(x)/2 + \mathcal{O}\left(h^3\right) - f(x)}{h} \right| = \left| -\frac{h}{2} f''(x) + \mathcal{O}\left(h^3\right) \right|.$$

When h is small, the first term will tend to dominate the $\mathcal{O}\left(h^3\right)$ term, so that the error associated with the forward finite difference is $\mathcal{O}(h)$. We can similarly consider the *central finite difference*, which is given by

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} \approx \frac{f(x+h) - f(x-h)}{2h}.$$

The error associated with the central finite difference is

$$E_{\text{ctr}}(h) = \left| f'(x) - \frac{f(x+h) - f(x-h)}{2h} \right|.$$

A third-order Taylor series for for $f(x-h)$ around the base point x is

$$f(x-h) \approx f(x) + f'(x)(x-h-x) + \frac{1}{2!} f''(x)(x-h-x)^2 + \frac{1}{3!} f'''(x)(x-h-x)^3. \tag{17.5.2}$$

Subtracting (17.5.2) from (17.5.1) gives

$$\begin{aligned} f(x+h) &\approx f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(x) \\ -f(x-h) &\approx f(x) - hf'(x) + \frac{h^2}{2} f'(x) - \frac{h^3}{6} f'''(x) \\ &\approx 0 + 2hf'(x) + 0 + \frac{2h^3}{6} f'''(x), \end{aligned}$$

so that

$$E_{\text{ctr}}(h) = \left| f'(x) - \frac{2hf'(x) + h^3 f'''(x)/3}{2h} \right| = \left| -\frac{h^2}{6} f'''(x) \right|,$$

i.e., the error associated with the central finite difference is $\mathcal{O}\left(h^2\right)$. While the central finite difference is more accurate than the forward finite difference, the forward finite difference requires only a single evaluation of f beyond $f(x)$ and is thus a little faster.

We now consider choosing h, which we might naïvely choose to be arbitrarily small. Recall that what we actually compute is not in general the exact representation of f, but its floating-point representation, so that the error associated with the forward finite difference becomes

$$\begin{aligned} E_{\text{fwd,ro}}(h) &= \left| f'(x) - \text{fl}\left(\frac{f(x+h) - f(x)}{h} \right) \right| \\ &= \left| f'(x) - \frac{\text{fl}\left(f(x+h)\right) - \text{fl}\left(f(x)\right)}{h} \right| \\ &= \left| f'(x) - \frac{f(x+h) \pm \epsilon_1 - \left[f(x) \pm \epsilon_2\right]}{h} \right|, \end{aligned}$$

where ϵ_1 and ϵ_2 denote the round-off errors (each on the order of machine epsilon ϵ_{mach}) associated with the floating-point representation of the respective function evaluations. The ϵ_i may not have the same sign, so without loss of generality, we have

$$E_{\text{fwd,ro}}(h) = \left| f'(x) - \frac{f(x+h) - f(x)}{h} - \frac{\epsilon_1 + \epsilon_2}{h} \right| = \left| -\frac{h}{2} f''(x) + \mathcal{O}\left(h^3\right) - \frac{\epsilon_1 + \epsilon_2}{h} \right|.$$

Noting that $h > 0$, we have

$$\left| \frac{\epsilon_1 + \epsilon_2}{h} \right| \leq \frac{2\epsilon_{\text{mach}}}{h} \implies E_{\text{fwd,ro}}(h) \leq \left| -\frac{h}{2} f''(x) + \mathcal{O}\left(h^3\right) + \frac{2\epsilon_{\text{mach}}}{h} \right|.$$

Now, as h goes to zero, the first two terms will go to zero, but the term $2\epsilon_{\text{mach}}/h$ will go to infinity. Conversely, as h goes to infinity, the term $2\epsilon_{\text{mach}}/h$ will go to zero, but the first two terms will go to infinity. I.e., we must choose h to simultaneously minimize these two sources of error. Ignoring the $\mathcal{O}\left(h^3\right)$ term and taking the derivative of $E_{\text{fwd,ro}}(h)$ gives

$$E'_{\text{ro,fwd}}(h) = \frac{1}{2} f''(x) - \frac{2\epsilon_{\text{mach}}}{h^2}.$$

Setting this equal to zero, we have

$$\frac{2\epsilon_{\text{mach}}}{h^2} = \frac{1}{2} f''(x) \implies h^2 f''(x) = 4\epsilon_{\text{mach}} \implies h^2 = \frac{4\epsilon_{\text{mach}}}{f''(x)} \implies h = \sqrt{\frac{4\epsilon_{\text{mach}}}{f''(x)}}.$$

We see that the h that minimizes the error of the approximation with round-off error is proportional to the square root of ϵ_{mach}. Taking $\epsilon_{\text{mach}} = 10^{-16}$, we have

$$h \approx \sqrt{\epsilon_{\text{mach}}} = \sqrt{10^{-16}} = 10^{-8},$$

i.e., for the forward finite difference, we should take h to be approximately 10^{-8}. Similarly, the error associated with the central finite difference becomes

$$\begin{aligned} E_{\text{ctr,ro}}(h) &= \left| f'(x) - \text{fl}\left(\frac{f(x+h) - f(x-h)}{2h} \right) \right| \\ &= \left| f'(x) - \frac{\text{fl}\left(f(x+h)\right) - \text{fl}\left(f(x-h)\right)}{2h} \right| \\ &= \left| f'(x) - \frac{f(x+h) \pm \epsilon_1 - \left[f(x-h) \pm \epsilon_2\right]}{2h} \right|, \end{aligned}$$

where ϵ_1 and ϵ_2 again denote the round-off errors associated with the floating-point representation of the respective function evaluations. Proceeding as above, we have

$$E_{\text{ctr,ro}}(h) = \left| f'(x) - \frac{f(x+h) - f(x-h)}{2h} - \frac{\epsilon_1 + \epsilon_2}{2h} \right|$$

$$
\begin{aligned}
&= \left| -\frac{h^2}{6} f'''(x) - \frac{\epsilon_1 + \epsilon_2}{2h} \right| \\
&\leq \left| -\frac{h^2}{6} f'''(x) - \frac{2\epsilon_{\text{mach}}}{2h} \right| \\
&= \left| -\frac{h^2}{6} f'''(x) - \frac{\epsilon_{\text{mach}}}{h} \right|.
\end{aligned}
$$

Taking the derivative gives

$$E'_{\text{ctr,ro}}(h) = -\frac{h}{3} f'''(x) + \frac{\epsilon_{\text{mach}}}{h^2}.$$

Setting this equal to zero, we have

$$\frac{\epsilon_{\text{mach}}}{h^2} = \frac{h}{3} f'''(x) \implies h^3 f'''(x) = 3\epsilon_{\text{mach}} \implies h^3 = \frac{3\epsilon_{\text{mach}}}{f'''(x)} \implies h = \left[\frac{3\epsilon_{\text{mach}}}{f'''(x)} \right]^{1/3}.$$

We see that the h that minimizes the error of the approximation with round-off error is proportional to the cube root of ϵ_{mach}. Taking $\epsilon_{\text{mach}} = 10^{-16}$, we have

$$h \approx (\epsilon_{\text{mach}})^{1/3} = \left(10^{-16}\right)^{1/3} = 10^{-16/3},$$

i.e., for the central finite difference, we should take h to be approximately 10^{-5}.

We can also use finite differences to calculate the second derivative of f as

$$f''(x) \approx \frac{f(x+h) - 2f(x) - f(x-h)}{h^2},$$

by applying Taylor series to the terms $f(x+h)$ and $f(x-h)$.

17.6. Numerical integration

Let f be a real-valued function of $x \in \mathbb{R}$, and consider the problem of computing the definite integral of f over some interval (a, b) on which f is continuous, i.e., $\int_a^b f(x)\,\mathrm{d}x$. We can we can approximate the integral of f over (a, b) by a Riemann sum. Let $x_0 = a$ and let $x_n = b$, and divide the interval into $n-1$ subintervals of equal width Δx, so that the ith subinterval is given by $(x_i, x_i + 1)$. For the ith subinterval, construct a rectangle whose width is $\Delta x = x_{i+1} - x_i$ and whose height is $f(x_i)$. Then, the Riemann approximation is

$$
\begin{aligned}
\int_a^b f(x)\,\mathrm{d}x &= \int_{x_0}^{x_1} f(x)\,\mathrm{d}x + \int_{x_1}^{x_2} f(x)\,\mathrm{d}x + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,\mathrm{d}x \\
&\approx f(x_0)\,\Delta x + f(x_1)\,\Delta x + \cdots + f(x_{n-1})\,\Delta x \\
&= \sum_{i=0}^{n-1} f(x_i)\,\Delta x,
\end{aligned}
$$

which we depict in figure 17.6.1.

We now consider the error associated with Riemann integration. Letting $x_0 = a$ and $x_n = b$, it follows from the linearity of integration that

$$\int_a^b f(x)\,\mathrm{d}x = \int_{x_0}^{x_1} f(x)\,\mathrm{d}x + \int_{x_1}^{x_2} f(x)\,\mathrm{d}x + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,\mathrm{d}x = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x.$$

We will approximate each term in this sum by $f(x_i)\,\Delta x$, so that the error is

$$E = \left| \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x - \sum_{i=0}^{n-1} f(x_i)\,\Delta x \right| = \left| \sum_{i=0}^{n-1} \left(\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}x - f(x_i)\,\Delta x \right) \right|.$$

Now, a first-order Taylor series for $f(x)$ around the base point x_i is $f(x) \approx f(x_i) + f'(x_i)(x - x_i)$, so that the error becomes

$$E \approx \left| \sum_{i=0}^{n-1} \left(\int_{x_i}^{x_{i+1}} f(x_i) + f'(x_i)(x - x_i)\,\mathrm{d}x - f(x_i)\,\Delta x \right) \right|$$

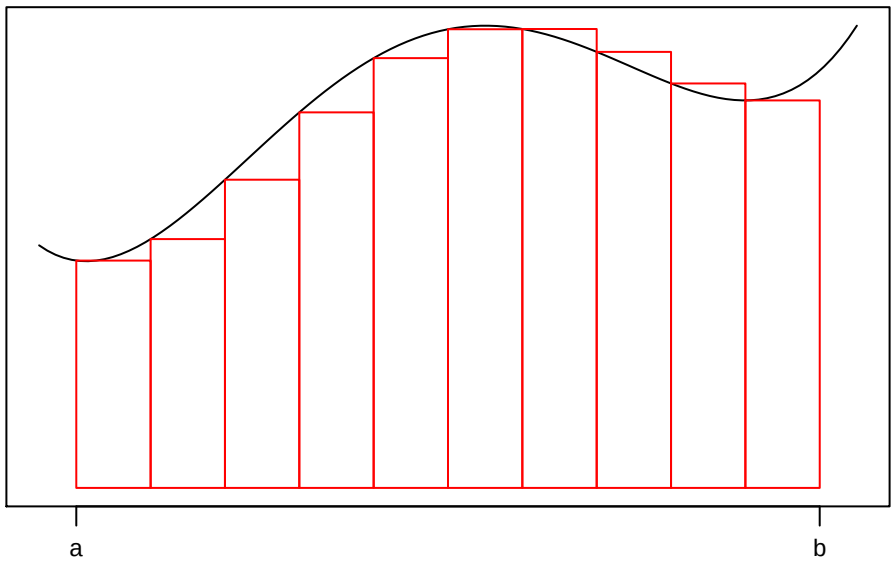


FIGURE 17.6.1. Riemann integration

$$
\begin{aligned}
&= \left| \sum_{i=0}^{n-1} \left(\left[f(x_i)\, x + \frac{f'(x_i)\,(x - x_i)^2}{2} \right|_{x_i}^{x_{i+1}} - f(x_i)\,\Delta x \right) \right| \\
&= \left| \sum_{i=0}^{n-1} \left(f(x_i)\, x_{i+1} + \frac{f'(x_i)\,(x_{i+1} - x_i)^2}{2} - f(x_i)\, x_i - \frac{f'(x_i)\,(x_i - x_i)^2}{2} - f(x_i)\,\Delta x \right) \right| \\
&= \left| \sum_{i=0}^{n-1} \left(f(x_i)\, x_{i+1} + \frac{f'(x_i)\,(\Delta x)^2}{2} - f(x_i)\, x_i - 0 - f(x_i)\,\Delta x \right) \right| \\
&= \left| \sum_{i=0}^{n-1} \left(f(x_i)\,(x_{i+1} - x_i) + \frac{f'(x_i)\,(\Delta x)^2}{2} - f(x_i)\,\Delta x \right) \right| \\
&= \left| \sum_{i=0}^{n-1} \left(f(x_i)\,\Delta x + \frac{f'(x_i)\,(\Delta x)^2}{2} - f(x_i)\,\Delta x \right) \right| \\
&= \left| \sum_{i=0}^{n-1} \frac{f'(x_i)\,(\Delta x)^2}{2} \right|.
\end{aligned}
$$

Now, $f'(x_i)$ is a constant, which we denote by c, so that

$$E \approx \left| \sum_{i=0}^{n-1} \frac{c\,(\Delta x)^2}{2} \right| = \left| \frac{c}{2} \sum_{i=0}^{n} (\Delta x)^2 \right| = \left| \frac{cn}{2} (\Delta x)^2 \right|.$$

The product of the number of intervals and the width of each interval is equal to the width of the interval (a, b), i.e., $n\Delta x = b - a$, so that

$$E \approx \left| \frac{c}{2} \left(\frac{b-a}{\Delta x} \right) (\Delta x)^2 \right| = \left| \frac{c}{2} (b-a)\,\Delta x \right|,$$

so that Riemann integration has error $\mathcal{O}(\Delta x)$.

We can also approximate the integral of f over the interval (a, b) by the trapezoid rule. For the ith subinterval, rather than construct a rectangle as in the Riemann sum approximation, we will construct a trapezoid whose vertices are given by $(x_i, x_{i+1}, f(x_{i+1}), f(x_i))$. The area of a trapezoid is equal to the product of the mean of the lengths of its parallel sides and its height (the perpendicular distance between the parallel sides). It

follows that the area of the ith trapezoid is

$$A_i = \frac{f(x_i) + f(x_{i+1})}{2} \Delta x,$$

so that the trapezoid rule approximation is

$$\begin{aligned}
\int_a^b f(x)\,\mathrm{d}\,x &= \int_{x_0}^{x_1} f(x)\,\mathrm{d}\,x + \int_{x_1}^{x_2} f(x)\,\mathrm{d}\,x + \cdots + \int_{x_{n-1}}^{x_n} f(x)\,\mathrm{d}\,x \\
&\approx \frac{f(x_0) + f(x_1)}{2} \Delta x + \frac{f(x_1) + f(x_2)}{2} \Delta x + \cdots + \frac{f(x_{n-1}) + f(x_n)}{2} \Delta x \\
&= \sum_{i=0}^{n-1} \frac{f(x_i) + f(x_{i+1})}{2} \Delta x,
\end{aligned}$$

which we depict in figure 17.6.2.

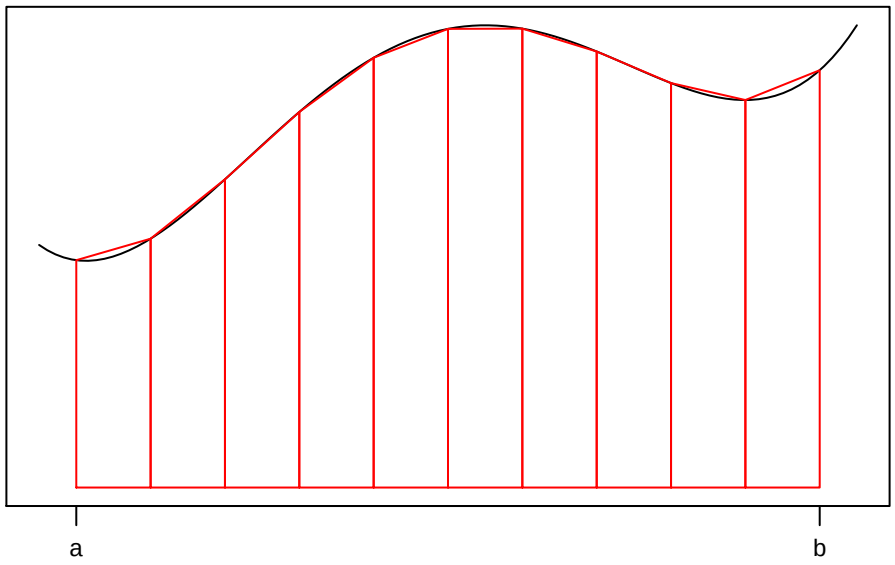


FIGURE 17.6.2. Trapezoid rule integration

We now consider the error associated with trapezoid rule integration. Again letting $x_0 = a$ and $x_n = b$, we have

$$\int_a^b f(x)\,\mathrm{d}\,x = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}\,x.$$

We will approximate each term in this sum by $(f(x_i) + f(x_{i+1}))\,\Delta x/2$, so that the error is

$$E = \left| \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}\,x - \sum_{i=0}^{n-1} \frac{f(x_i) + f(x_{i+1})}{2} \Delta x \right| = \left| \sum_{i=0}^{n-1} \left(\int_{x_i}^{x_{i+1}} f(x)\,\mathrm{d}\,x - \frac{f(x_i) + f(x_{i+1})}{2} \Delta x \right) \right|.$$

Applying a first-order Taylor series for $f(x)$ around the base point x_i and a first-order Taylor series for $f(x_{i+1})$ around x_i gives

$$\begin{aligned}
E &\approx \left| \sum_{i=0}^{n-1} \left(\int_{x_i}^{x_{i+1}} f(x_i) + f'(x_i)(x - x_i)\,\mathrm{d}\,x - \frac{f(x_i) + f(x_i) + f'(x_i)(x_{i+1} - x_i)}{2} \Delta x \right) \right| \\
&= \left| \sum_{i=0}^{n-1} \left(\left[f(x_i)\,x + \frac{f'(x_i)(x - x_i)^2}{2} \right|_{x_i}^{x_{i+1}} - \frac{2f(x_i) + f'(x_i)\,\Delta x}{2} \Delta x \right) \right|
\end{aligned}$$

$$= \left| \sum_{i=0}^{n-1} \left(f(x_i) x_{i+1} + \frac{f'(x_i)(x_{i+1} - x_i)^2}{2} - f(x_i) x_i - \frac{f'(x_i)(x_i - x_i)^2}{2} - f(x_i) \Delta x - \frac{f'(x_i)(\Delta x)^2}{2} \right) \right|$$

$$= \left| \sum_{i=0}^{n-1} \left(f(x_i)(x_{i+1} - x_i) + \frac{f'(x_i)(\Delta x)^2}{2} - 0 - f(x_i) \Delta x - \frac{f'(x_i)(\Delta x)^2}{2} \right) \right|$$

Part 5

Deterministic mathematical models

CHAPTER 18

Optimization

18.1. Classical multivariable optima

A real-valued, differentiable function f of a real variable x on a closed interval $[a, b]$ achieves both a maximum and a minimum on $[a, b]$, and any extreme must be at an endpoint or an interior *critical point* x_0 where $f'(x_0) = 0$. To understand the behavior of f around x_0, we can apply a second-order Taylor series expansion, which gives

$$f(x) \approx f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2.$$

Now, by assumption, x_0 is a critical point, hence $f'(x_0) = 0$, so that

$$f(x) \approx f(x_0) + \frac{1}{2} f''(x_0)(x - x_0)^2.$$

If the f'' term is positive, then $f(x)$ will be greater than $f(x_0)$, so that x_0 will be a local minimum. If the f'' term is negative, then $f(x)$ will be less than $f(x_0)$, so that x_0 will be a local maximum. The quantity $(x - x_0)^2$ will be nonnegative, hence we can characterize x_0 by inspecting $f''(x_0)$, which leads to the single variable *second derivative test*: if $f''(x_0) > 0$, then x_0 is a local minimum, and if $f''(x_0) < 0$, then x_0 is a local maximum.

If f is a function of two variables, e.g., x and y, then a stationary or critical point occurs at the point (x_0, y_0) if both partial derivatives are zero, i.e., if $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$. To classify the critical point, we first calculate determinant of the Hessian matrix of f i.e.,

$$D(x, y) = \det\left(\begin{bmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{bmatrix}\right) = f_{xx}(x, y) f_{yy}(x, y) - f_{xy}(x, y) f_{yx}(x, y) = f_{xx}(x, y) f_{yy}(x, y) - (f_{xy}(x, y))^2,$$

where the final equality follows from our assumption that f has continuous second partial derivatives. This leads the two-variable second derivative test: if (x_0, y_0) is a critical point of f, i.e., $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$, and

(1) if $D(x_0, y_0) > 0$ and $f_{xx}(x_0, y_0) < 0$, then (x_0, y_0) is a local maximum.
(2) if $D(x_0, y_0) > 0$ and $f_{xx}(x_0, y_0) > 0$, then (x_0, y_0) is a local minimum.
(3) if $D(x_0, y_0) < 0$, then (x_0, y_0) is a saddle point.
(4) if $D(x_0, y_0) = 0$, then the test is inconclusive.

EXAMPLE 18.1. Find and characterize the critical points of the following functions.

(1) Let $f(x, y) = x^2 - y^2$. We have $f_x = 2x$, which will be equal to zero if and only if $x = 0$, and $f_y = -2y$, which will similarly be zero only in the case that $y = 0$. Thus, $(0, 0)$ is the single critical point of f. We have

$$f_{xx} = 2, \quad f_{yy} = -2, \quad \text{and} \quad f_{xy} = f_{yx} = 0,$$

so that $D(0, 0) = -4 - 0 = -4$. Applying the second derivative test, we see that $(0, 0)$ is a saddle point.

(2) Let $f(x, y) = x^2 y^2 - y^2 - x^2 + 1$. We have $f_x = 2xy^2 - 2x$ and $f_y = 2yx^2 - 2y$. Setting f_x equal to zero, we have

$$0 = 2xy^2 - 2x = 2x\left(y^2 - 1\right) = 2x(y - 1)(y + 1),$$

which will be zero if either $x = 0$ or $y = \pm 1$. If $x = 0$, then setting $f_y = 0$ and substituting yields

$$f_y(0, y) = 2y \cdot 0 - 2y = -2y = 0 \implies y = 0.$$

If $y = 1$, then

$$f_y(x,1) = 2x^2 - 2 = 0 \implies 2x^2 = 2 \implies x = \pm 1,$$

and if $y = -1$, then

$$f_y(x,-1) = -2x^2 + 2 = 0 \implies 2x^2 = 2 \implies x = \pm 1.$$

Thus, we obtain the critical points $\{(0,0),(1,1),(-1,1),(1,-1),(-1,-1)\}$. We have

$$f_{xx} = 2y^2 - 2 = 2\left(y^2-1\right), \quad f_{yy} = 2x^2 - 2 = 2\left(x^2-1\right), \quad \text{and} \quad f_{xy} = 4xy,$$

so that

$$D(x,y) = f_{xx}f_{yy} - f_{xy}^2 = 4\left(y^2-1\right)\left(x^2-1\right) - 16x^2y^2 = 4\left(\left(y^2-1\right)\left(x^2-1\right) - 4x^2y^2\right).$$

Then, $D(0,0) = 4$ and $f_{xx}(0,0) = -2$, so that $(0,0)$ is a local maximum. Observe that x and y appear in $D(x,y)$ only in squared form, hence the result obtained for $(1,1)$ will be the same as the result for all $(\pm 1, \pm 1)$. Then, $D(1,1) = -16$, so that each of $(\pm 1, \pm 1)$ is a saddle point.

18.2. Calculus of variations

The calculus of variations, or variational method, lays out a pathway to finding extremals for functionals, which are real-valued mappings defined on sets of functions. The simplest example is perhaps the arc length of a function $y = f(x)$ from $(a, f(a))$ to $(b, f(b))$ on an interval $[a,b]$, which is given by

$$J[f] = \int_a^b \sqrt{1 + (f'x)^2}\, \mathrm{d}\, x,$$

where $J[f]$ is the corresponding arc length functional. In one dimension, the function that minimizes the arc length is a straight line, though in higher dimensions, the picture becomes more complicated.

18.2.1. The Lagrangian. We will be interested in functionals of the form

$$J[f] = \int_a^b L(x, f(x), f'(x))\, \mathrm{d}\, x$$

for functions in some admissible set $\mathcal{A}$ and where $L = L(x_1, x_2, x_3)$ is known as the *Lagrangian* associated with J. L is generally assumed to be twice differentiable in all variables. We also assume that $\mathcal{A}$ is a normed linear space, e.g., $\mathcal{C}^1[a,b]$. Then, to say that J has a local minimum at $f_0 \in \mathcal{A}$ will mean that $J[f_0] \leq J[f]$ for all f in a norm-based neighborhood of f_0.

For a given functional J and a point $f_0 \in \mathcal{A}$, let h be a function such that $f_0 + \epsilon h \in \mathcal{A}$ for all sufficiently small and positive ϵ. Then, the total change in J at f_0 due to ϵh, denoted ΔJ, is defined as

$$\Delta J[f_0, h] \coloneqq J[f_0 + h] - J[f_0].$$

Definition 18.2. For some functional J, a point in the admissible set $f_0 \in \mathcal{A}$, and a function h that satisfies $f_0 + \epsilon h \in \mathcal{A}$, the *first variation* or *Gâteaux derivative* of J at f_0 in the direction h is defined as

$$\delta J[f_0, h] \coloneqq \left[\frac{\mathrm{d}}{\mathrm{d}\,\epsilon} J[f_0 + \epsilon h]\right]\Bigg|_{\epsilon=0} = \lim_{\epsilon \to 0} \frac{J[f_0 + \epsilon h] - J[f_0]}{\epsilon} = \lim_{\epsilon \to 0} \frac{\Delta J[f_0, \epsilon h]}{\epsilon}.$$

Example 18.3. Find the Gâteaux derivative for the functional

$$J[f] = \int_0^1 1 + (f'(x))^2\, \mathrm{d}\, x,$$

and find the extremal for $h(x) = x(1-x)$.

Noting that the derivative of $f(x) + \epsilon h(x)$ is $f'(x) + \epsilon h'(x)$, we have

$$\begin{aligned} J[f + \epsilon h] &= \int_0^1 1 + \left(\left(f(x) + \epsilon h(x)\right)'\right)^2 \mathrm{d}\, x \\ &= \int_0^1 1 + \left(f'(x) + \epsilon h'(x)\right)^2 \mathrm{d}\, x \\ &= \int_0^1 1 + \left(f'(x)\right)^2 + 2\epsilon f'(x) h'(x) + \epsilon^2 \left(h'(x)\right)^2 \mathrm{d}\, x \end{aligned}$$

$$= \int_0^1 \left(1 + (f'(x))^2\right) \mathrm{d}\, x + 2\epsilon \int_0^1 f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon^2 \int_0^1 (h'(x))^2\, \mathrm{d}\, x,$$

so that

$$\Delta J[f, \epsilon h] = J[f + \epsilon h] - J[f] = 2\epsilon \int_0^1 f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon^2 \int_0^1 (h'(x))^2\, \mathrm{d}\, x.$$

Then, the Gâteaux derivative is

$$\begin{aligned}
\delta J[f, h] &= \lim_{\epsilon \to 0} \frac{\Delta J[f, \epsilon h]}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(2\epsilon \int_0^1 f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon^2 \int_0^1 (h'(x))^2\, \mathrm{d}\, x\right) \\
&= \lim_{\epsilon \to 0} \left(2 \int_0^1 f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon \int_0^1 (h'(x))^2\, \mathrm{d}\, x\right) \\
&= 2 \int_0^1 f'(x)\, h'(x)\, \mathrm{d}\, x + 0 \\
&= 2 \int_0^1 f'(x)\, h'(x)\, \mathrm{d}\, x.
\end{aligned}$$

For $h(x) = x(1 - x)$, we have $h'(x) = 1 - 2x$. Setting $\delta J[f, h]$ equal to zero gives

$$0 = 2 \int_0^1 f'(x)\, (1 - 2x)\, \mathrm{d}\, x.$$

Suppose that $f(x) = c_1 x + c_2$ for some $c_1, c_2 \in \mathbb{R}$, i.e., f is linear. Then, $f'(x) = c_1$, so that

$$0 = 2 \int_0^1 c_1 (1 - 2x)\, \mathrm{d}\, x = 2c_1 \int_0^1 1 - 2x\, \mathrm{d}\, x = 2c_1 \left[x - x^2\right|_0^1 = 2c_1 (0 - 0) = 0,$$

i.e., the equality holds. It follows that $f(x) = c_1 x + c_2$ is the minimal function.

Example 18.4. Find the Gâteaux derivative for the functional

$$J[f] = \int_0^1 (1 + x)\, (f'(x))^2\, \mathrm{d}\, x,$$

and find the extremal for $h(x) = x(1 - x)$.

We have

$$\begin{aligned}
J[f + \epsilon h] &= \int_0^1 (1 + x)\, (f'(x) + \epsilon h'(x))^2\, \mathrm{d}\, x \\
&= \int_0^1 (1 + x) \left[(f'(x))^2 + 2\epsilon f'(x)\, h'(x) + \epsilon^2 (h'(x))^2\right] \mathrm{d}\, x \\
&= \int_0^1 (1 + x)\, (f'(x))^2\, \mathrm{d}\, x + 2\epsilon \int_0^1 (1 + x)\, f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon^2 \int_0^1 (1 + x)\, (h'(x))^2\, \mathrm{d}\, x,
\end{aligned}$$

so that

$$\Delta J[f, \epsilon h] = J[f + \epsilon h] - J[f] = 2\epsilon \int_0^1 (1 + x)\, f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon^2 \int_0^1 (1 + x)\, (h'(x))^2\, \mathrm{d}\, x.$$

Then,

$$\begin{aligned}
\delta J[f, h] &= \lim_{\epsilon \to 0} \frac{\Delta J[f, \epsilon h]}{\epsilon} \\
&= \lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(2\epsilon \int_0^1 (1 + x)\, f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon^2 \int_0^1 (1 + x)\, (h'(x))^2\, \mathrm{d}\, x\right) \\
&= \lim_{\epsilon \to 0} \left(2 \int_0^1 (1 + x)\, f'(x)\, h'(x)\, \mathrm{d}\, x + \epsilon \int_0^1 (1 + x)\, (h'(x))^2\, \mathrm{d}\, x\right) \\
&= 2 \int_0^1 (1 + x)\, f'(x)\, h'(x)\, \mathrm{d}\, x.
\end{aligned}$$

For $h(x) = x(1-x)$, we have $h'(x) = 1-2x$. Setting $\delta J[f,h]$ equal to zero gives

$$0 = 2\int_0^1 (1+x) f'(x)(1-2x)\,\mathrm{d}x.$$

Suppose $f(x) = c\log(1+x)$ for some $c \in \mathbb{R}$, so that $f'(x) = c/(1+x)$. Then,

$$0 = 2\int_0^1 (1+x)\left(\frac{c}{1+x}\right)(1-2x)\,\mathrm{d}x = 2c\int_0^1 (1-2x)\,\mathrm{d}x = 2c\cdot 0 = 0,$$

i.e., the equality holds. It follows that $f(x) = c\log(1+x)$ is the minimal function.

18.2.2. Euler's equation. Suppose that a functional has a Lagrangian of the form $L(x, f(x), f'(x))$. Euler's method can be used to find the extremals of the functional. We begin by stating a lemma necessary for the proof of the method.

LEMMA 18.5. *The Gâteaux derivative of a functional J with Lagrangian L can be approximated by the equation*

$$\delta J[f,h] = \lim_{\epsilon\to 0}\frac{\Delta J[f,\epsilon h]}{\epsilon} \approx \int_a^b h(x)\left[L_f(x,f(x),f'(x)) + h'(x)L_{f'}(x,f(x),f'(x))\right]\mathrm{d}x,$$

where L_f and $L_{f'}$ denote the derivatives of L with respect to f and f', respectively.

PROOF. [proof goes here]Letting $y = f(x)$ and $z = f'(x)$, we will apply a second-order Taylor series expansion to the Lagrangian around the base point $(x, y+\Delta y, z+\Delta z)$, i.e.,

$$L(x, y+\Delta y, z+\Delta z) \approx L()$$

□

THEOREM 18.6 (Euler's method). *The extremals of the functional*

$$J[f] = \int_a^b L(x, f(x), f'(x))\,\mathrm{d}x$$

are obtained by solving

$$L_f - \frac{\mathrm{d}}{\mathrm{d}x}L_{f'} = 0.$$

PROOF. [proof goes here] □

EXAMPLE 18.7. Use Euler's method to find the minimal for the functional

$$J[f] = \int_0^1 x f(x) f'(x)\,\mathrm{d}x.$$

The Lagrangian is $L = xff'$, so that

$$L_f = xf', \quad L_{f'} = xf, \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}x}L_{f'} = f + xf'.$$

Then, Euler's method implies that the extremals of $J[f]$ are obtained by solving

$$0 = L_f - \frac{\mathrm{d}}{\mathrm{d}x}L_{f'} = xf' - f - xf' = -f \implies f(x) = 0.$$

Thus, the only possibility for the extremal of $J[f]$ is $f(x) = 0$.

EXAMPLE 18.8. Use Euler's method to find the minimal for the functional

$$J[f] = \int_0^1 (f'(x))^2 + 3f(x) + 2x\,\mathrm{d}x.$$

The Lagrangian is $L = (f')^2 + 3f + 2x$, so that

$$L_f = 3, \quad L_{f'} = 2f', \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}x}L_{f'} = 2f''.$$

Then, Euler's method implies that the extremals of $J[f]$ are obtained by solving

$$0 = L_f - \frac{\mathrm{d}}{\mathrm{d}\,x} L_{f'} = 3 - 2f'' \implies f''(x) = \frac{3}{2}.$$

We have $b = c = 0$, so that the roots of the characteristic equation are $r_1 = r_2 = 0$. Then, the homogeneous solution is $f_h(x) = c_1 \mathrm{e}^0 + x c_2 \mathrm{e}^0 = c_1 + x c_2$ for some $c_1, c_2 \in \mathbb{R}$, where we have introduced a factor of x into the second term to create linear independence between the solutions. We will apply the method of undetermined coefficients to find the particular solution. The forcing term is $3/2$, so we will seek $f_p(x)$ as a multiple of the lowest integral power of x such that 1 (ignoring the multiplicative constant $3/2$) does not duplicate any of the terms in f_h. We see that f_h includes both a constant and first-order term, so we must multiply 1 by x^2 to avoid duplication. Then, $f_p(x) = Ax^2$, so that $f_p' = 2Ax$ and $f_p'' = 2A$. Then, the nonhomogeneous system becomes

$$2A = \frac{3}{2} \implies A = \frac{3}{4},$$

so that the general solution is

$$f(x) = f_h(x) + f_p(x) = \frac{3}{4}x^2 + c_2 x + c_1.$$

When the Lagrangian takes certain forms, the Euler's method solution may be simplified.

18.2.2.1. *Case 1: Lagrangian does not depend on $f'(x)$.* In this case, $L_{f'} = 0$, so that Euler's equation becomes

$$0 = L_f - \frac{\mathrm{d}}{\mathrm{d}\,x} L_{f'} = L_f.$$

Example 18.9. Use Euler's method to find the minimal for the functional

$$J[f] = \int_0^1 x^2 + (f(x))^2 \,\mathrm{d}\,x.$$

The Lagrangian $L = x^2 + f^2$ does not depend on f', hence the minimal of $J[f]$ is obtained by solving $0 = L_f = 2f$, so that $f(x) = 0$ is the minimal.

Example 18.10. Use Euler's method to find the minimal for the functional

$$J[f] = \int_0^1 x^2 (f(x))^2 + 4f(x) \,\mathrm{d}\,x.$$

The Lagrangian $L = x^2 f^2 + 4f$ does not depend on f', hence the minimal of $J[f]$ is obtained by solving

$$0 = L_f = 2x^2 f + 4 \implies x^2 f(x) = -2,$$

so that the minimal is $f(x) = -2/x^2$.

18.2.2.2. *Case 2: Lagrangian does not depend on $f(x)$.* In this case, $L_f = 0$, so that Euler's equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}\,x} L_{f'} = 0 \implies \int \left(\frac{\mathrm{d}}{\mathrm{d}\,x} L_{f'} \right) \mathrm{d}\,x = \int 0 \,\mathrm{d}\,x \implies L_{f'} = C$$

for some $C \in \mathbb{R}$. This case also applies when the Lagrangian depends only on $L_{f'}$.

Example 18.11. Use Euler's method to find the minimal for the functional

$$J[f] = \int_0^1 \frac{(f'(x))^2}{x^3} \,\mathrm{d}\,x.$$

The Lagrangian $L = (f')^2 / x^3$ does not depend on f, hence the minimal of $J[f]$ is obtained by solving

$$C = L_{f'} = \frac{2f'}{x^3} \implies 2f' = Cx^3 \implies \int f'(x) \,\mathrm{d}\,x = \frac{C}{2} \int x^3 \,\mathrm{d}\,x,$$

which leads to the solution

$$f(x) = \frac{C}{2} \left(\frac{1}{4} x^4 + C_2 \right) = \frac{C}{8} x^4 + C_2 = C_1 x^4 + C_2,$$

where $C_1 = C/8$.

EXAMPLE 18.12 (Arc length). Use Euler's method to find the minimal for the functional

$$J[f] = \int_a^b \sqrt{1 + (f'(x))^2}\,\mathrm{d}\,x.$$

The Lagrangian $L = \sqrt{1 + (f')^2}$ does not depend on f, hence the minimal of $J[f]$ is obtained by solving

$$C = L_{f'} = \frac{2f'}{2\sqrt{1 + (f')^2}} = \frac{f'}{\sqrt{1 + (f')^2}},$$

which implies that

$$\begin{aligned} f' &= C\sqrt{1 + (f')^2} \\ \implies (f')^2 &= C^2\left(1 + (f')^2\right) \\ \implies C^2 &= (f')^2 - C^2 (f')^2 \\ &= (f')^2 \left(1 - C^2\right) \\ \implies f' &= \sqrt{\frac{C^2}{1 - C^2}}. \end{aligned}$$

Setting $C_1 = \sqrt{C^2/(1 - C^2)}$, we have the ODE

$$f'(x) = C_1 \implies \int f'(x)\,\mathrm{d}\,x = \int C_1\,\mathrm{d}\,x,$$

which leads to the solution $f(x) = C_1 x + C_2$, i.e., the shortest curve between two points (in the Cartesian plane) is a line.

18.2.2.3. *Case 3: Lagrangian does not depend on x.* In this case, we can apply the chain rule when differentiating $L_{f'}$ with respect to x, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}\,x} L_{f'} = \frac{\partial L_{f'}}{\partial f}\frac{\partial f}{\partial x} + \frac{\partial L_{f'}}{\partial f'}\frac{\partial f'}{\partial x} = L_{f'f} \cdot f' + L_{f'f'} \cdot f''.$$

Then, the extremals may be obtained by using Euler's method, i.e.,

$$0 = L_f - \frac{\mathrm{d}}{\mathrm{d}\,x} L_{f'} = L_f - L_{f'f} \cdot f' - L_{f'f'} \cdot f''.$$

Alternatively, we can multiply this equation by f' to obtain
FIX THIS DERIVATION

$$0 \cdot f' = f' \cdot L_f - L_{f'f} \cdot f' - L_{f'f'} \cdot f'' \implies 0 = L_f \cdot f' - L_{f'f} \cdot (f')^2 - L_{f'f'} \cdot f''f' = \frac{\mathrm{d}}{\mathrm{d}\,x}\left(L - L_{f'} f'\right),$$

which implies that the extremals may equivalently be obtained by solving

$$\int \frac{\mathrm{d}}{\mathrm{d}\,x}\left(L - L_{f'} f'\right)\mathrm{d}\,x = \int 0\,\mathrm{d}\,x \implies L - L_{f'} f' = C.$$

EXAMPLE 18.13. Use Euler's method to find the minimal for the functional

$$J[f] = \int_a^b (f(x))^2 + (f'(x))^2\,\mathrm{d}\,x.$$

The Lagrangian $L = f^2 + (f')^2$ does not depend on x, hence the minimal of $J[f]$ is obtained by solving

$$0 = L_f - L_{f'f} \cdot f' - L_{f'f'} \cdot f'' = 2f - 0 \cdot f' - 2f'' = -2f'' + 2f \implies f'' - f = 0.$$

This is a homogeneous second-order ODE, which can be solved by finding the roots of the characteristic equation. We have $b = 0$ and $c = -1$, so that

$$r^2 + br + c = r^2 - 1 = 0 \implies r^2 = 1 \implies r = \pm 1,$$

i.e., the roots are $r_1 = 1$ and $r_2 = -1$. Then, the general solution is $f(x) = c_1 \mathrm{e}^x + c_2 \mathrm{e}^{-x}$.

EXAMPLE 18.14. Use Euler's method to find the minimal for the functional

$$J[f] = \int_a^b 5f(x) f'(x) + (f'(x))^2 - 4(f(x))^2 \,\mathrm{d}\,x.$$

The Lagrangian $L = 5ff' + (f')^2 - 4f^2$ does not depend on x, hence the minimal of $J[f]$ is obtained by solving

$$0 = L_f - L_{f'f} \cdot f' - L_{f'f'} \cdot f'' = 5f' - 8f - 5f' - 2f'' = -2f'' - 8f \implies f'' + 4f = 0.$$

The roots of the characteristic equation are

$$r^2 + br + c = r^2 + 4 = 0 \implies r^2 = -4 \implies r = \pm\sqrt{-4} = \pm 2\iota,$$

i.e., the roots are $r_1 = 2\iota$ and $r_2 = -2\iota$. Then, the general solution is

$$\begin{aligned} f(x) &= c_1 \mathrm{e}^{2\iota x} + c_2 \mathrm{e}^{-2\iota x} \\ \text{(Euler's Formula)} \qquad &= c_1(\cos(2x) + \iota \sin(2x)) + c_2(\cos(-2x) + \iota \sin(-2x)) \\ &= c_1 \cos(2x) + \iota c_1 \sin(2x) + c_2 \cos(2x) - \iota c_2 \sin(2x) \\ &= (c_1 + c_2)\cos(2x) + \iota(c_1 - c_2)\sin(2x) \\ &= c_1 \cos(2x) + c_2 \sin(2x), \end{aligned}$$

where we have redefined the arbitrary constants c_1 and c_2.

18.2.3. Boundary conditions. The resultant solutions to the functionals can be written in general form, e.g., $f(x) = c_1 x + c_2$. To find an explicit solution, we will require additional information in the form of boundary conditions.

DEFINITION 18.15. For a functional of the form

$$J[f] = \int_a^b L(x, f(x), f'(x)) \,\mathrm{d}\,x,$$

the *boundary conditions* are defined as $f(a) = y_0$ and $f(b) = y_1$.

EXAMPLE 18.16. Use Euler's method to find the minimal for the functional

$$J[f] = \int_a^b \sqrt{1 + (f'(x))^2} \,\mathrm{d}\,x$$

that satisfies the boundary conditions $f(0) = 0$ and $f(1) = 1$.
The minimal is $f(x) = c_1 x + c_2$. We have

$$0 = f(0) = c_2 \implies c_2 = 0 \quad \text{and} \quad 1 = f(1) = c_1 + c_2 = c_1 \implies c_1 = 1,$$

so that the particular solution is $f(x) = x$. This accords with our geometric intuition: the line that passes through $(0, 0)$ and $(1, 1)$ must have slope 1 and intercept 0.

The concept of a boundary condition will differ somewhat from the initial conditions used to find particular solutions to ODEs. In particular, it is possible that the function that minimizes the functional fails to satisfy the boundary conditions.

EXAMPLE 18.17. Use Euler's method to find the minimal for the functional

$$J[f] = \int_0^1 x f(x) f'(x) \,\mathrm{d}\,x$$

that satisfies the boundary conditions $f(0) = 0$ and $f(1) = 1$.
The minimal is $f(x) = 0$, which we see satisfies the first boundary condition, but not the second.

18.2.4. Applied examples.

EXAMPLE 18.18 (Cost functional).

EXAMPLE 18.19 (Brachistochrone problem).

18.3. Lagrange multipliers

The objective of the method of Lagrange multipliers is to optimize a function f subject to a constraint of the form $\varphi = 0$, where φ is a function of the same variables as f. We will assume that, in the two-dimensional case, the partial derivative with respect to y evaluated at an extremum is non-zero, and in the three-dimensional case, that the partial derivative with respect to z evaluated at an extremum is non-zero.

18.3.1. Two-dimensional case. Suppose that we would like to find the extrema of $f(x,y)$ subject to the constraint $\varphi(x,y) = 0$. Let $\lambda \in \mathbb{R}$ be the Lagrange multiplier, and form the auxiliary function $F(x,y,\lambda) = f(x,y) + \lambda\varphi(x,y)$. The critical points of F are solutions of the simultaneous equations $F_x = 0$, $F_y = 0$, and $F_\lambda = 0$. Observe that the condition $F_\lambda = 0$ is equivalent to the constraint $\varphi(x,y) = 0$.

EXAMPLE 18.20. Find the extremes of $f(x,y) = x^2 + xy + y^2$ subject to $y = x - 1$.

We have $\varphi(x,y) = y - x + 1$, so that the auxiliary function is

$$F(x,y,\lambda) = f(x,y) + \lambda\varphi(x,y) = x^2 + xy + y^2 + \lambda(y - x + 1).$$

The optima $(x^\star, y^\star)$ will occur at points that satisfy

$$\{(x,y) : F_x(x,y) = 0, F_y(x,y) = 0, \text{ and } F_\lambda(x,y) = 0\},$$

where F_z denotes the partial derivative of F with respect to z. We have

$$F_x = 2x + y - \lambda, \quad F_y = x + 2y + \lambda, \quad \text{and} \quad F_\lambda = y - x + 1.$$

Setting these equal to zero and adding the first equation to the second yields

$$0 + 0 = F_x + F_y = 2x + y - \lambda + x + 2y + \lambda = 3x + 3y \implies 3x = -3y \implies x = -y.$$

Substituting into the third equation gives

$$0 = y + y + 1 = 2y + 1 \implies y = -\frac{1}{2} \implies x = \frac{1}{2}.$$

Observe that the point $(1,0)$ satisfies $y = x - 1$, and observe also that $f(1,0) = 1$. The solution $(x^\star, y^\star) = (1/2, -1/2)$ obtained by the method of Lagrange multipliers is unique, and we have

$$f(x^\star, y^\star) = \frac{1}{4} - \frac{1}{4} + \frac{1}{4} = \frac{1}{4} < 1,$$

so it follows that $(1/2, -1/2)$ is a minimum.

EXAMPLE 18.21. Find and characterize the extremes of $f(x,y) = (x-1)^2 y^2 - y^2 - (x-1)^2 + 1$ subject to $x + y = 1$.

We have $\varphi(x,y) = x + y - 1$, so that the auxiliary function is

$$F(x,y,\lambda) = f(x,y) + \lambda\varphi(x,y) = (x-1)^2 y^2 - y^2 - (x-1)^2 + \lambda(x + y - 1).$$

The partial derivatives of F are

$$F_x = 2y^2(x-1) - 2(x-1) + \lambda = 2(x-1)\left(y^2 - 1\right) + \lambda,$$

$$F_y = 2y(x-1)^2 - 2y + \lambda = 2y\left((x-1)^2 - 1\right) + \lambda, \quad \text{and} \quad F_\lambda = x + y - 1.$$

Setting $F_\lambda = 0$ yields $-y = x - 1$. Setting $F_x = 0$ and substituting, we have

$$0 = -2y\left(y^2 - 1\right) + \lambda \implies \lambda = 2y\left(y^2 - 1\right).$$

Noting that $(1-x)^2 = (x-1)^2 = (-y)^2 = y^2$, setting $F_y = 0$ and substituting gives

$$0 = 2y\left(y^2 - 1\right) + 2y\left(y^2 - 1\right) = 4y\left(y^2 - 1\right) \implies y\left(y^2 - 1\right) = y(y-1)(y+1) = 0,$$

so that $y^\star \in \{1, 0, -1\}$, hence the extrema of f are $\{(0,1), (1,0), (2,-1)\}$. Define the *bordered Hessian matrix* by

$$\mathbf{H}_F(x,y,\lambda) = \begin{bmatrix} 0 & -g_x & -g_y \\ -g_x & F_{xx} & F_{xy} \\ -g_y & F_{yx} & F_{yy} \end{bmatrix},$$

where $g(x,y) = x+y$, i.e., $g(x,y)-1 = \varphi(x,y)$. If $(x^\star, y^\star, \lambda^\star)$ is a critical point of F obtained by the method of Lagrange multipliers, it can be shown that $(x^\star, y^\star, \lambda^\star)$ is a local maximum if $\det(\mathbf{H}_F(x^\star, y^\star, \lambda^\star)) > 0$, and that $(x^\star, y^\star, \lambda^\star)$ is a local minimum if $\det(\mathbf{H}_F(x^\star, y^\star, \lambda^\star)) < 0$. The determinant of $\mathbf{H}_F$ is

$$\begin{aligned}
\det(\mathbf{H}_F) &= \det\left(\begin{bmatrix} 0 & -g_x & -g_y \\ -g_x & F_{xx} & F_{xy} \\ -g_y & F_{yx} & F_{yy} \end{bmatrix}\right) \\
&= 0\begin{vmatrix} F_{xx} & F_{xy} \\ F_{yx} & F_{yy} \end{vmatrix} - (-g_x)\begin{vmatrix} -g_x & -g_y \\ F_{yx} & F_{yy} \end{vmatrix} + (-g_y)\begin{vmatrix} -g_x & -g_y \\ F_{xx} & F_{xy} \end{vmatrix} \\
&= g_x\begin{vmatrix} -g_x & -g_y \\ F_{yx} & F_{yy} \end{vmatrix} - g_y\begin{vmatrix} -g_x & -g_y \\ F_{xx} & F_{xy} \end{vmatrix} \\
&= g_x(-g_xF_{yy} + g_yF_{yx}) - g_y(-g_xF_{xy} + g_yF_{xx}) \\
&= -g_x^2F_{yy} + g_xg_yF_{yx} + g_yg_xF_{xy} - g_y^2F_{xx} \\
(F_{xy} = F_{yx}) \qquad &= 2g_xg_yF_{xy} - g_x^2F_{yy} - g_y^2F_{xx}.
\end{aligned}$$

We have $g_x = 1$, $g_y = 1$,

$$F_{xx} = 2\left(y^2-1\right), \quad F_{yy} = 2\left((x-1)^2-1\right), \quad \text{and} \quad F_{xy} = 4y(x-1) = F_{yx},$$

where the final equality follows from the fact that F is a polynomial, hence has continuous second partial derivatives. Thus,

$$\det(\mathbf{H}_F) = 8y(x-1) - 2\left((x-1)^2-1\right) - 2\left(y^2-1\right).$$

For $(0,1)$, we have

$$\det(\mathbf{H}_F(0,1)) = 8(-1) - 2\left((-1)^2-1\right) - 2(1-1) = -8-0-0 = -8,$$

which implies that $(0,1)$ is a local minimum. For $(1,0)$, we have

$$\det(\mathbf{H}_F(1,0)) = 0 - 2(0-1) - 2(0-1) = 2+2 = 4,$$

which implies that $(1,0)$ is a local maximum. For $(2,-1)$, we have

$$\det(\mathbf{H}_F(2,-1)) = -8(1) - 2(1-1) - 2(0) = -8,$$

which implies that $(2,-1)$ is a local minimum.

18.3.2. Three-dimensional case. Suppose that we would like to find the extrema of $f(x,y,z)$ subject to a constraint of the form $\varphi(x,y,z) = 0$. Similar to the two-dimensional case, we will form the auxiliary function $F(x,y,z,\lambda) = f(x,y,z) + \lambda\varphi(x,y,z)$. The critical points of F are solutions of the simultaneous equations $F_x = 0$, $F_y = 0$, $F_z = 0$, and $F_\lambda = 0$.

EXAMPLE 18.22. Minimize $x^2+y^2+z^2$ subject to $2x+3y-z = 1$.

We have $\varphi(x,y,z) = 2x+3y-z-1$, so that the auxiliary function is

$$F(x,y,z,\lambda) = f(x,y,z) + \lambda\varphi(x,y,z) = x^2+y^2+z^2+\lambda(2x+3y-z-1).$$

We have

$$F_x = 2x+2\lambda, \quad F_y = 2y+3\lambda, \quad F_z = 2z-\lambda, \quad \text{and} \quad F_\lambda = 2x+3y-z-1.$$

Setting these equal to zero, the first three equations yield

$$x = -\lambda, \quad y = -\frac{3}{2}\lambda, \quad \text{and} \quad z = \frac{1}{2}\lambda.$$

Setting $F_\lambda = 0$ and substituting, we have

$$0 = -2\lambda - \frac{9}{2}\lambda - \frac{1}{2}\lambda - 1 \implies 7\lambda = -1 \implies \lambda = -\frac{1}{7} \implies (x,y,z) = \left(\frac{1}{7}, -\frac{3}{14}, -\frac{1}{14}\right).$$

Observe that the point $(0,0,-1)$ satisfies $2x+3y-z = 1$, and observe also that $f(0,0,-1) = 1$. The solution $(x^\star, y^\star, z^\star) = (1/7, -3/14, -1/14)$ obtained by the method of Lagrange multipliers is unique, and we have

$$f(x^\star, y^\star, z^\star) = \frac{1}{49} + \frac{9}{196} + \frac{1}{196} = \frac{14}{196} = \frac{1}{14} < 1,$$

so it follows that $(x^\star, y^\star, z^\star)$ is a minimum.

18.3.3. Economic motivation. We now present a classical consumer choice problem. FINISH PROBLEM

18.4. Variation subject to constraints

We can also apply the method of Lagrange multipliers to functionals of the form

$$J[f] = \int_a^b L(x, f(x), f'(x)) \, \mathrm{d}x$$

for functions f in the admissible set $\mathcal{A}$. We wish to find the extremals of the functional subject to a constraint of the form $\varphi(x, f) = 0$. Let

$$\tilde{L}(x, f, f') = L(x, f, f') + \lambda \varphi(\lambda, f),$$

and define the functional

$$F[f, \lambda] \coloneqq \int_a^b L(x, f(x), f'(x)) \, \mathrm{d}x + \lambda \int_a^b \varphi(x, f(x)) \, \mathrm{d}x = \int_a^b \tilde{L}(x, f(x), f'(x)) \, \mathrm{d}x.$$

THEOREM 18.23. *The extremals of the functional*

$$J[f] = \int_a^b L(x, f(x), f'(x)) \, \mathrm{d}x$$

with constraint $\varphi(x, f) = 0$ such that $\varphi_x \neq 0$ and $\varphi_f \neq 0$ are obtained by solving

$$\frac{\mathrm{d}}{\mathrm{d}x} L_{f'} - L_f = \lambda \varphi_f.$$

PROOF. [proof goes here] □

EXAMPLE 18.24. Find the minimal for the functional

$$J[f] = \int_a^b (f(x))^2 + (f'(x))^2 \, \mathrm{d}x$$

with constraint $\varphi(x, f) = x^2 f - x$.
The Lagrangian is $L = f^2 + (f')^2$, so that

$$L_f = 2f, \quad L_{f'} = 2f', \quad \text{and} \quad \frac{\mathrm{d}}{\mathrm{d}x} L_{f'} = 2f''.$$

Noting that $\varphi_f = x^2$, it follows from theorem 18.23 that the minimal of $J[f]$ subject to $\varphi(x, f) = 0$ is obtained by solving

$$\frac{\mathrm{d}}{\mathrm{d}x} L_{f'} - L_f = \lambda \varphi_f \implies 2f'' - 2f = \lambda x^2 \implies f'' - f = \frac{\lambda}{2} x^2.$$

This is a second-order nonhomogeneous ODE, so we will apply the method of undetermined coefficients. We have $b = 0$ and $c = -1$, so that the characteristic equation becomes

$$r^2 + br + c = r^2 - 1 = 0 \implies r^2 = 1 \implies r = \pm 1,$$

i.e., the roots are $r_1 = 1$ and $r_2 = -1$. Then, the homogeneous solution is

$$f_h(x) = c_1 \mathrm{e}^x + c_2 \mathrm{e}^{-x}.$$

The forcing term x^2 has derivatives (ignoring multiplicative constants) of x and 1, so we will seek the particular solution as a linear combination of these terms, i.e., $f_p = Ax^2 + Bx + C$, so that $f_p' = 2Ax + B$ and $f_p'' = 2A$. We observe that none of the terms in f_p is duplicated in f_h, so that the nonhomogeneous system becomes

$$\frac{\lambda}{2} x^2 = 2A - Ax^2 - Bx - C = -Ax^2 - Bx + (2A - C) \implies A = -\frac{\lambda}{2}, B = 0, C = 2A = -\lambda.$$

Then, the general solution is given by

$$f(x) = f_h(x) + f_p(x) = c_1 \mathrm{e}^x + c_2 \mathrm{e}^{-x} - \frac{\lambda}{2} x^2 - \lambda.$$

CHAPTER 19

Ordinary differential equations

19.1. First order equations

A differential equation of the form $x'(t) = f(x(t), t)$, where $x'(t) = \mathrm{d}x/\mathrm{d}t$, is called a *first order equation* because the first derivative is the highest order derivative that appears in the equation. We may also have the *initial condition* $x(t_0) = x_0$.

19.1.1. Separable equations. If we can collect all the terms involving x on one side of the equation and all the terms involving t on the other, then the equation is said to be *separable*. Equivalently, a separable equation is one that may be written as $f(x,t) = M(x)N(t)$, where M and N are functions of x only and t only, respectively. Separable equations can be solved by collecting terms in this manner, then integrating both sides to obtain the *general solution*. If in addition we have an initial condition, we may determine the constant of integration to obtain the *particular solution*.

EXAMPLE 19.1. Solve $x' = t$, with $x(0) = 1$.

We obtain the general solution

$$\frac{\mathrm{d}x}{\mathrm{d}t} = t \implies 1\,\mathrm{d}x = t\,\mathrm{d}t \implies \int 1\,\mathrm{d}x = \int t\,\mathrm{d}t \implies x(t) = \frac{t^2}{2} + C,$$

where $C \in \mathbb{R}$ has "absorbed" the constants of integration on both sides of the equality. Applying the initial value condition, we have

$$x(0) = \frac{0^2}{2} + C = 1 \implies C = 1,$$

so that the particular solution is $x(t) = t^2/2 + 1$.

19.1.2. First order linear equations. A first order differential equation of the form $x' + h(t)x = g(t)$ is called a *first order linear equation* because the equation is linear in x. Such equations can be solving by the *method of integrating factor*. Let $H(t)$ be an antiderivative of $h(t)$, and observe that

$$\begin{aligned}
\mathrm{e}^{H(t)}(x' + h(t)x) &= \mathrm{e}^{H(t)}x' + \mathrm{e}^{H(t)}h(t)x \\
(h(t) = H'(t)) \qquad &= \mathrm{e}^{H(t)}\frac{\mathrm{d}}{\mathrm{d}t}x + x\mathrm{e}^{H(t)}H'(t) \\
(\text{Chain Rule}) \qquad &= \mathrm{e}^{H(t)}\frac{\mathrm{d}}{\mathrm{d}t}x + x\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{e}^{H(t)} \\
(\text{Product Rule}) \qquad &= \left(x\mathrm{e}^{H(t)}\right)'.
\end{aligned}$$

Thus, the general solution is obtained by solving

$$\begin{aligned}
\mathrm{e}^{H(t)}(x' + h(t)x) &= \left(x\mathrm{e}^{H(t)}\right)' \\
&= \mathrm{e}^{H(t)}g(t) \\
\implies \int \left(x\mathrm{e}^{H(t)}\right)'\mathrm{d}t &= \int \mathrm{e}^{H(t)}g(t)\,\mathrm{d}t \\
(\text{Fundamental Theorem of Calculus}) \qquad \implies x\mathrm{e}^{H(t)} &= \int \mathrm{e}^{H(t)}g(t)\,\mathrm{d}t \\
\implies x(t) &= \mathrm{e}^{-H(t)}\int \mathrm{e}^{H(t)}g(t)\,\mathrm{d}t.
\end{aligned}$$

EXAMPLE 19.2. Solve $x' - 7x = 1$, with $x(0) = 0$.

Let $h(t) = -7$, and observe that $H(t) = -7t$ is an antiderivative of $h(t)$. Multiplying both sides by $\mathrm{e}^{H(t)}$ and applying the above result gives

$$\mathrm{e}^{H(t)}(x' - 7x) = \mathrm{e}^{H(t)} \implies \left(x\mathrm{e}^{-7t}\right)' = \mathrm{e}^{-7t}.$$

Integrating both sides with respect to t gives the general solution

$$\int \left(x\mathrm{e}^{-7t}\right)' \mathrm{d}\,t = \int \mathrm{e}^{-7t}\,\mathrm{d}\,t \implies x(t) = \mathrm{e}^{7t}\left(-\frac{1}{7}\mathrm{e}^{-7t} + C\right) = -\frac{1}{7} + C\mathrm{e}^{7t}.$$

Applying the initial value condition, we have

$$x(0) = -\frac{1}{7} + C\mathrm{e}^{0} = 0 \implies C = \frac{1}{7},$$

so that the particular solution is

$$x(t) = \frac{1}{7}\left(\mathrm{e}^{7t} - 1\right).$$

EXAMPLE 19.3 (Terminal velocity). Suppose a sky diver with terminal velocity $v_T = 200$ ft/sec steps out of a hovering helicopter and drops straight down subject to $v'(t) = \alpha(v_T - v(t))$. Find the velocity $v(t)$ for all t.

The sky diver is subject to gravitational acceleration and air resistance, with the resulting acceleration governed by $v'(t) = \alpha(v_T - v(t))$, and where we will take the constant of gravitational acceleration to be $G = 32$ ft/sec^2. Taking downward motion to be positive, we see that

$$\alpha v_T = G \implies \alpha = \frac{G}{v_T} = \frac{32 \text{ ft/sec}^2}{200 \text{ ft/sec}} = \frac{32}{200}\text{sec}^{-1}.$$

Then,

$$v'(t) = \frac{G}{v_T}(v_T - v(t)) = G - \frac{G}{v_T}v(t) \implies v'(t) + \frac{G}{v_T}v(t) = G.$$

This is a first order linear equation with $h(t) = G/v_T$ and $g(t) = G$. Observe that $H(t) = Gt/v_T$ is an antiderivative of $h(t)$. Multiplying both sides by $\mathrm{e}^{H(t)}$ and applying the result above gives

$$\mathrm{e}^{H(t)}\left(v' + \frac{G}{v_T}v(t)\right) = \mathrm{e}^{H(t)}G \implies \left(v(t)\,\mathrm{e}^{Gt/v_T}\right)' = G\mathrm{e}^{Gt/v_T}.$$

Integrating both sides with respect to t gives the general solution

$$\int \left(v(t)\,\mathrm{e}^{Gt/v_T}\right)' \mathrm{d}\,t = \int G\mathrm{e}^{Gt/v_T}\,\mathrm{d}\,t \implies v(t) = \mathrm{e}^{-Gt/v_T}\left(v_T\mathrm{e}^{Gt/v_T} + C\right) = v_T + C\mathrm{e}^{-Gt/v_T}.$$

At time $t = 0$, the sky diver is just stepping out of the helicopter, hence has velocity $v_0 = 0$, so that

$$v(0) = v_T + C\mathrm{e}^{-G\cdot 0/v_T} = 0 \implies C = -v_T.$$

Then, the particular solution is

$$\begin{aligned} v(t) &= v_T - v_T\mathrm{e}^{-Gt/v_T} \\ &= v_T\left(1 - \mathrm{e}^{-Gt/v_T}\right) \\ &= 200\frac{\text{ft}}{\text{sec}}\left(1 - \exp\left\{-32\frac{\text{ft}}{\text{sec}^2}\cdot t \text{ sec}\cdot\frac{1}{200}\frac{\text{sec}}{\text{ft}}\right\}\right) \\ &= 200\frac{\text{ft}}{\text{sec}}\left(1 - \mathrm{e}^{-32t/200}\right) \\ &= 200\left(1 - \mathrm{e}^{-0.16t}\right)\frac{\text{ft}}{\text{sec}}. \end{aligned}$$

19.2. Second order equations

The general second order linear equation is of the form

$$y'' + p(x)\, y' + q(x)\, y = g(x).$$

The equation is called *homogeneous* if $g(x) = 0$ and nonhomogeneous otherwise.

19.2.1. Homogeneous, constant coefficients. Suppose that $p(x) = b$ and $q(x) = c$, where $b, c \in \mathbb{R}$, and suppose that $g(x) = 0$. Then, solutions are of the form e^{rx}, where r is a root of the characteristic equation $r^2 + br + c = 0$. There will generally be two roots, denoted r_1 and r_2, so that the general solution has the form $y(x) = c_1\mathrm{e}^{r_1 x} + c_2\mathrm{e}^{r_2 x}$. In the case that $r_1 = r_2$, then the second solution defaults to $x\mathrm{e}^{rx}$, where r is the single root.

Example 19.4. Solve $y'' - y' - 2y = 0$, with $y(0) = y'(0) = 1$.

This is a homogenous equation with constant coefficients $b = -1$ and $c = -2$, so that solutions are of the form e^{rx}, where r is a root of the characteristic equation

$$r^2 + br + c = r^2 - r - 2 = (r-2)(r+1) = 0,$$

so that the roots are $r_1 = 2$ and $r_2 = -1$. Then, the general solution is $y(x) = c_1\mathrm{e}^{2x} + c_2\mathrm{e}^{-x}$, so that $y' = 2c_1\mathrm{e}^{2x} - c_2\mathrm{e}^{-x}$. Applying the initial value conditions, we have

$$y(0) = c_1\mathrm{e}^0 + c_2\mathrm{e}^0 = 1 \implies c_1 = 1 - c_2 \quad \text{and} \quad y'(0) = 2c_1\mathrm{e}^0 - c_2\mathrm{e}^0 = 1 \implies c_1 = \frac{1+c_2}{2},$$

so that

$$1 - c_2 = \frac{1+c_2}{2} \implies 2 - 2c_2 = 1 + c_2 \implies 3c_2 = 1 \implies c_2 = \frac{1}{3} \text{ and } c_1 = \frac{1}{2} + \frac{1/3}{2} = \frac{2}{3}.$$

Then, the particular solution is

$$y(x) = \frac{2}{3}\mathrm{e}^{2x} + \frac{1}{3}\mathrm{e}^{-x}.$$

Example 19.5. Solve $y'' - 2y' + y = 0$.

This is a homogeneous equation with constant coefficients $b = -2$ and $c = 1$, so that solutions are of the form e^{rx}, where r is a root of the characteristic equation

$$r^2 + br + c = r^2 - 2r + 1 = (r-1)(r-1) = 0,$$

so that the single root (with multiplicity 2) is $r = 1$. Then, the general solution is $y(x) = c_1\mathrm{e}^x + c_2 x\mathrm{e}^x$. The constants c_1, c_2 can be found using initial conditions.

The roots of $r^2 + br + c = 0$ may be complex, e.g., $r_1 = \alpha + \iota\beta$ and $r_2 = \alpha - \iota\beta$, where $\iota = \sqrt{-1}$. In this case, the general solution has the form

$$y(x) = c_1\mathrm{e}^{\alpha x}\cos(\beta x) + c_2\mathrm{e}^{\alpha x}\sin(\beta x).$$

Example 19.6. Solve $y'' + 2y' + 5y = 0$, with $y(0) = 1$ and $y'(0) = 0$.

We have $b = 2$ and $c = 5$, so that the characteristic equation is

$$r^2 + br + c = r^2 + 2r + 5 = 0 \implies r = \frac{-2 \pm \sqrt{4-20}}{2} = \frac{-2 \pm 4\iota}{2} = -1 \pm 2\iota,$$

so that the roots are $r_1 = -1 + 2\iota$ and $r_2 = -1 - 2\iota$. Then, the general solution is

$$y(x) = c_1\mathrm{e}^{(-1+2\iota)x} + c_2\mathrm{e}^{(-1-2\iota)x} = c_1\mathrm{e}^{-x}\mathrm{e}^{2\iota x} + c_2\mathrm{e}^{-x}\mathrm{e}^{-2\iota x}.$$

Applying Euler's Formula gives

$$\begin{aligned} y(x) &= c_1\mathrm{e}^{-x}(\cos(2x) + \iota\sin(2x)) + c_2\mathrm{e}^{-x}(\cos(-2x) + \iota\sin(-2x)) \\ &= c_1\mathrm{e}^{-x}\cos(2x) + c_1\mathrm{e}^{-x}\iota\sin(2x) + c_2\mathrm{e}^{-x}\cos(2x) - c_2\mathrm{e}^{-x}\iota\sin(2x) \\ &= (c_1 + c_2)\,\mathrm{e}^{-x}\cos(2x) + \iota\,(c_1 - c_2)\,\mathrm{e}^{-x}\sin(2x). \end{aligned}$$

Noting that c_1 and c_2 are arbitrary constants, we may redefine them to yield $y(x) = c_1 \mathrm{e}^{-x} \cos(2x) + c_2 \mathrm{e}^{-x} \sin(2x)$. Then,

$$\begin{aligned} y' &= \frac{\mathrm{d}}{\mathrm{d}\,x} \mathrm{e}^{-x} \left(c_1 \cos(2x) + c_2 \sin(2x)\right) \\ &= -\mathrm{e}^{-x} \left(c_1 \cos(2x) + c_2 \sin(2x)\right) + \mathrm{e}^{-x} \left(-2c_1 \sin(2x) + 2c_2 \cos(2x)\right) \\ &= \mathrm{e}^{-x} \left(-c_1 \cos(2x) - c_2 \sin(2x) - 2c_1 \sin(2x) + 2c_2 \cos(2x)\right) \\ &= \mathrm{e}^{-x} \left((2c_2 - c_1) \cos(2x) - (2c_1 + c_2) \sin(2x)\right), \end{aligned}$$

so that applying the initial value conditions gives

$$y(0) = c_1 \mathrm{e}^0 \cos(0) + c_2 \mathrm{e}^0 \sin(0) = c_1 = 1$$

and

$$y'(0) = -\mathrm{e}^0 \left((2c_2 - c_1) \cos(0) - (2c_1 + c_2) \sin(0)\right) = -2c_2 + c_1 = -2c_2 + 1 = 0 \implies c_2 = \frac{1}{2}.$$

Then, the particular solution is

$$y(x) = \mathrm{e}^{-x} \cos(2x) + \frac{1}{2} \mathrm{e}^{-x} \sin(2x).$$

19.2.2. Nonhomogeneous equations. Suppose that $g(x) \neq 0$. The general solution will be of the form $y = y_h + y_p$, where y_p is a particular solution to the nonhomogeneous equation and y_h is the general solution of the homogeneous equation. Suppose that the forcing term $g(x)$ has the form $g(x) = K\mathrm{e}^{mx}$ for some $K, m \in \mathbb{R}$. Then, the particular solution will have the form $y_p = Ax^s \mathrm{e}^{mx}$, where $A \in \mathbb{R}$ and s is the smallest of $\{0, 1, 2\}$ such that $x^s \mathrm{e}^{mx}$ is not a solution of the homogeneous equation.

EXAMPLE 19.7. Find the solution to $y'' - 3y' + 2y = 10\mathrm{e}^{3x}$, with initial conditions $y(0) = y'(0) = 0$.

We have $b = -3$ and $c = 2$, so that the characteristic equation is

$$r^2 + br + c = r^2 - 3r + 2 = (r - 1)(r - 2) = 0,$$

so that the roots are $r_1 = 1$ and $r_2 = 2$. Then, the homogeneous solution is $y(x) = c_1 \mathrm{e}^x + c_2 \mathrm{e}^{2x}$. We have $m = 3$, so that $y_p = Ax^s \mathrm{e}^{3x}$. If $s = 0$, then $y_p = A\mathrm{e}^{3x}$, so that $y_p' = 3A\mathrm{e}^{3x}$ and $y_p'' = 9A\mathrm{e}^{3x}$. Then, the homogeneous equation becomes

$$y'' - 3y' + 2y = 9A\mathrm{e}^{3x} - 3\left(3A\mathrm{e}^{3x}\right) + 2\left(A\mathrm{e}^{3x}\right) = 2A\mathrm{e}^{3x} \neq 0,$$

i.e., $y_p = A\mathrm{e}^{3x}$ is not a solution of the homogeneous equation. Hence, we will take $s = 0$, so that the nonhomogeneous equation becomes

$$y'' - 3y' + 2y = 2A\mathrm{e}^{3x} = 10\mathrm{e}^{3x} \implies A = 5,$$

so that the general solution is $y(x) = c_1 \mathrm{e}^x + c_2 \mathrm{e}^{2x} + 5\mathrm{e}^{3x}$. Noting that

$$y'(x) = c_1 \mathrm{e}^x + 2c_2 \mathrm{e}^{2x} + 15\mathrm{e}^{3x},$$

applying the initial value conditions gives

$$y(0) = c_1 + c_2 + 5 = 0 \implies c_1 = -c_2 - 5$$

and

$$y'(0) = c_1 + 2c_2 + 15 = 0 \implies c_1 = -2c_2 - 15,$$

so that

$$-c_2 - 5 = -2c_2 - 15 \implies c_2 = -10 \implies c_1 = 5.$$

Then, the particular solution of the nonhomogeneous equation is

$$y(x) = 5\mathrm{e}^x - 10\mathrm{e}^{2x} + 5\mathrm{e}^{3x} = 5\mathrm{e}^x \left(1 - 2\mathrm{e}^x + \mathrm{e}^{2x}\right) = 5\mathrm{e}^x \left(\mathrm{e}^x - 1\right)^2.$$

EXAMPLE 19.8. Find the solution to $y'' - 3y' + 2y = 10\mathrm{e}^{2x}$, with initial conditions $y(0) = y'(0) = 0$.

The homogeneous solution is again $y(x) = c_1\mathrm{e}^{x} + c_2\mathrm{e}^{2x}$, and we also have $y_p = Ax^s\mathrm{e}^{3x}$. We have $m = 2$, so that $y_p = Ax^s\mathrm{e}^{2x}$. If $s = 0$, then $y_p = A\mathrm{e}^{2x}$, so that $y_p' = 2A\mathrm{e}^{2x}$ and $y_p'' = 4A\mathrm{e}^{2x}$. Then, the homogeneous equation becomes

$$y'' - 3y' + 2y = 4A\mathrm{e}^{2x} - 3\left(2A\mathrm{e}^{2x}\right) + 2\left(A\mathrm{e}^{2x}\right) = 0,$$

i.e., $y_p = A\mathrm{e}^{2x}$ is a solution of the homogeneous equation. If $s = 1$, then $y_p = Ax\mathrm{e}^{2x}$, so that

$$y_p' = A\left(\mathrm{e}^{2x} + 2x\mathrm{e}^{2x}\right) = A\mathrm{e}^{2x} + 2Ax\mathrm{e}^{2x} \text{ and } y_p'' = 2A\mathrm{e}^{2x} + 2A\left(\mathrm{e}^{2x} + 2x\mathrm{e}^{2x}\right) = 4A\mathrm{e}^{2x} + 4Ax\mathrm{e}^{2x}.$$

Then, the homogeneous equation becomes

$$y'' - 3y' + 2y = 4A\mathrm{e}^{2x} + 4Ax\mathrm{e}^{2x} - 3\left(A\mathrm{e}^{2x} + 2Ax\mathrm{e}^{2x}\right) + 2Ax\mathrm{e}^{2x} = A\mathrm{e}^{2x} \neq 0,$$

i.e., $y_p = Ax\mathrm{e}^{2x}$ is not a solution of the homogeneous equation. Hence, we will take $s = 1$, so that the nonhomogeneous equation becomes

$$y'' - 3y' + 2y = A\mathrm{e}^{2x} = 10\mathrm{e}^{2x} \implies A = 10,$$

so that the general solution is $y(x) = c_1\mathrm{e}^{x} + c_2\mathrm{e}^{2x} + 10x\mathrm{e}^{2x}$. Noting that

$$y'(x) = c_1\mathrm{e}^{x} + 2c_2\mathrm{e}^{x} + 10\left(\mathrm{e}^{2x} + 2x\mathrm{e}^{2x}\right),$$

applying the initial value conditions gives

$$y(0) = c_1 + c_2 = 0 \implies c_1 = -c_2$$

and

$$y'(0) = c_1 + 2c_2 + 10 = 0 \implies c_1 = -2c_2 - 10,$$

so that

$$-c_2 = -2c_2 - 10 \implies c_2 - 10 \implies c_1 = 10.$$

Then, the particular solution of the nonhomogeneous equation is

$$y(x) = 10\mathrm{e}^{x} - 10\mathrm{e}^{2x} + 10x\mathrm{e}^{2x}.$$

Now suppose that the forcing term $g(x)$ has the form $g(x) = \mathrm{e}^{mx}\left(K\cos(\gamma x) + M\sin(\gamma x)\right)$ for some $K, M, m, \gamma \in \mathbb{R}$. Then, the particular solution will have the form $y_p = x^s\mathrm{e}^{mx}\left(A\cos(\gamma x) + B\sin(\gamma x)\right)$, where $A, B \in \mathbb{R}$ and s is the smallest of $\{0, 1, 2\}$ such that no term in y_p is a solution of the homogeneous equation.

EXAMPLE 19.9. Solve $y'' - 2y' + y = 5\cos(x)$, with initial conditions $y(0) = y'(0) = 0$.

We have $b = -2$ and $c = 1$, so that the characteristic equation is

$$r^2 + br + c = r^2 - 2r + 1 = (r-1)(r-1) = 0,$$

so that the single root (with multiplicity 2) is $r = 1$. Then, the homogeneous solution is $y(x) = c_1\mathrm{e}^{x} + c_2x\mathrm{e}^{x}$. We have $m = 0$, $K = 5$, $M = 0$, and $\gamma = 1$, so that the particular solution will be of the form

$$y_p = x^s\left(A\cos(x) + B\sin(x)\right).$$

If $s = 0$, then $y_p = A\cos x + B\sin(x)$, so that $y_p' = -A\sin x + B\cos(x)$ and $y_p'' = -A\cos(x) - B\sin(x)$. We consider each term of the particular solution separately. For $A\cos(x)$, the homogeneous equation is

$$y'' - 2y' + y = -A\cos(x) + 2A\sin(x) + A\cos(x) = 2A\sin(x),$$

which will not in general be equal to zero. For $B\sin(x)$, the homogeneous equation is

$$y'' - 2y' + y = -B\sin(x) - 2B\cos(x) + B\sin(x) = -2B\cos(x),$$

which will also not in general be equal to zero. Hence, we will take $s = 0$, so that the nonhomogeneous equation becomes

$$\begin{aligned} y'' - 2y' + y &= -A\cos(x) - B\sin(x) - 2\left(-A\sin(x) + B\cos(x)\right) + A\cos(x) + B\sin(x) \\ &= 2A\sin(x) - 2B\cos(x) \\ &= 5\cos(x), \end{aligned}$$

which implies that $A = 0$ and $B = -5/2$. Then, the general solution is

$$y(x) = c_1 e^x + c_2 x e^x - \frac{5}{2}\sin(x).$$

Noting that

$$y'(x) = c_1 e^x + c_2\left(e^x + x e^x\right) - \frac{5}{2}\cos(x),$$

applying the initial value conditions gives

$$y(0) = c_1 = 0 \quad \text{and} \quad y'(0) = c_1 + c_2 - \frac{5}{2} = 0 \implies c_2 = \frac{5}{2}.$$

Then, the particular solution of the nonhomogeneous equation is

$$y(x) = \frac{5}{2} x e^x - \frac{5}{2}\sin(x).$$

19.3. Systems of ordinary differential equations

A system of ODEs is a vector equation of the form $\mathbf{x}'(t) = \mathbf{f}(t, \mathbf{x}(t))$, where both $\mathbf{x}(t)$ and $\mathbf{f}(t, \mathbf{x}(t))$ may have complex-valued components. To make the problem well posed, we typically include an initial condition of the form $\mathbf{x}(t_0) = \mathbf{x}_0$, where t_0 and $\mathbf{x}_0$ are constant.

19.3.1. Linear autonomous systems with constant coefficients. A *linear autonomous system* is one that is linear in the unknowns and in which the time variable t enters only implicitly through the unknowns. Such systems can be expressed in the form

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t),$$

where $\mathbf{A}$ is a matrix with constant coefficients. Suppose that $\mathbf{q}$ is an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ, and consider the function $\mathbf{x}(t) = e^{\lambda t}\mathbf{q}$. We have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \lambda e^{\lambda t}\mathbf{q} \quad \text{and} \quad \mathbf{A}\mathbf{x}(t) = \mathbf{A}\left(e^{\lambda t}\right)\mathbf{q} = e^{\lambda t}\mathbf{A}\mathbf{q} = \lambda e^{\lambda t}\mathbf{q} \implies \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \mathbf{A}\mathbf{x}(t),$$

i.e., $\mathbf{x}(t) = e^{\lambda t}\mathbf{q}$ is a solution to the system. The general solution of such a system is given by

$$\mathbf{x}(t) = c_1 e^{\lambda_1 t}\mathbf{q}_1 + \cdots + c_n e^{\lambda_n t}\mathbf{q}_n = \sum_{i=1}^{n} c_i e^{\lambda_i t}\mathbf{q}_i,$$

where c_i is an arbitrary coefficient (which may be determined given initial values) and λ_i is the ith eigenvalue of $\mathbf{A} \in \mathbf{M}_{n,n}(\mathbb{R})$ with corresponding eigenvector $\mathbf{q}_i$.

Example 19.10. Find the unique solution to the system

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{x}(t) = \begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix}\mathbf{x}(t), \quad \text{with } \mathbf{x}(0) = \begin{bmatrix} 3 \\ 5 \end{bmatrix}.$$

We begin by finding the eigenvalues of $\mathbf{A}$, which are obtained by finding the roots of the characteristic equation

$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{I}_2) &= \det\left(\begin{bmatrix} 1 & 1 \\ 4 & 1 \end{bmatrix} - \lambda\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) \\ &= \det\left(\begin{bmatrix} 1-\lambda & 1 \\ 4 & 1-\lambda \end{bmatrix}\right) \\ &= (1-\lambda)(1-\lambda) - 4 \\ &= \lambda^2 - 2\lambda - 3 \\ &= (\lambda - 3)(\lambda + 1). \end{aligned}$$

We see that the roots of the characteristic equation, i.e., the eigenvalues of the matrix, are $\lambda_1 = 3$ and $\lambda_2 = -1$. We now find the eigenvector corresponding to λ_1 by solving

$$\mathbf{A}\mathbf{x} - \lambda_1\mathbf{x} = \mathbf{A}\mathbf{x} - \lambda_1\mathbf{I}_2\mathbf{x} = (\mathbf{A} - \lambda_1\mathbf{I}_2)\mathbf{x} = \mathbf{0},$$

i.e.,

$$\left[\begin{array}{cc|c} 1-\lambda_1 & 1 & 0 \\ 4 & 1-\lambda_1 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} -2 & 1 & 0 \\ 4 & -2 & 0 \end{array}\right] \begin{array}{c} \div -2 \\ +2R_1 \end{array} \sim \left[\begin{array}{cc|c} 1 & -1/2 & 0 \\ 0 & 0 & 0 \end{array}\right] \implies x_1 = \frac{1}{2}x_2.$$

Choose $x_2 = 2$, so that the eigenvector corresponding to λ_1 is $\mathbf{x}_1 = \begin{bmatrix} 1 & 2 \end{bmatrix}^\mathsf{T}$. Similarly, we have

$$\left[\begin{array}{cc|c} 1-\lambda_2 & 1 & 0 \\ 4 & 1-\lambda_2 & 0 \end{array}\right] \sim \left[\begin{array}{cc|c} 2 & 1 & 0 \\ 4 & 2 & 0 \end{array}\right] \begin{array}{c} \div 2 \\ -2R_1 \end{array} \sim \left[\begin{array}{cc|c} 1 & 1/2 & 0 \\ 0 & 0 & 0 \end{array}\right] \implies x_1 = -\frac{1}{2}x_2.$$

Choose $x_2 = -2$, so that the eigenvector corresponding to λ_2 is $\mathbf{x}_2 = \begin{bmatrix} 1 & -2 \end{bmatrix}^\mathsf{T}$. Thus, the general solution to the autonomous system is

$$\mathbf{x}(t) = c_1 \mathrm{e}^{\lambda_1 t}\mathbf{x}_1 + c_2 \mathrm{e}^{\lambda_2 t}\mathbf{x}_2 = c_1 \mathrm{e}^{3t}\begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2 \mathrm{e}^{-t}\begin{bmatrix} 1 \\ -2 \end{bmatrix}.$$

Applying the initial value condition gives

$$\mathbf{x}(0) = c_1\begin{bmatrix} 1 \\ 2 \end{bmatrix} + c_2\begin{bmatrix} 1 \\ -2 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix} \implies \begin{bmatrix} 1 & 1 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} 3 \\ 5 \end{bmatrix},$$

so that we must solve the linear system

$$\left[\begin{array}{cc|c} 1 & 1 & 3 \\ 2 & -2 & 5 \end{array}\right] \begin{array}{c} \\ -2R_1 \end{array} \sim \left[\begin{array}{cc|c} 1 & 1 & 3 \\ 0 & -4 & -1 \end{array}\right] \begin{array}{c} +R_2/4 \\ \div -4 \end{array} \sim \left[\begin{array}{cc|c} 1 & 0 & 11/4 \\ 0 & 1 & 1/4 \end{array}\right],$$

i.e., $c_1 = 11/4$ and $c_2 = 1/4$. Then, the unique solution to the autonomous system is

$$\mathbf{x}(t) = \frac{11}{4}\mathrm{e}^{3t}\begin{bmatrix} 1 \\ 2 \end{bmatrix} + \frac{1}{4}\mathrm{e}^{-t}\begin{bmatrix} 1 \\ -2 \end{bmatrix} = \mathrm{e}^{3t}\begin{bmatrix} 11/4 \\ 11/2 \end{bmatrix} + \mathrm{e}^{-t}\begin{bmatrix} 1/4 \\ -1/2 \end{bmatrix}.$$

19.3.2. Higher order ODEs. A single differential equation of higher order can be converted to a system. We can express a higher order equation as

$$\frac{\mathrm{d}^n y}{\mathrm{d}\, t^n} = f\left(t, y, y', \ldots, y^{(n-1)}\right).$$

Define

$$\begin{aligned} x_1 &= y \\ x_2 &= y' \\ &\vdots \quad \vdots \\ x_n &= y^{(n-1)}, \end{aligned}$$

so that

$$x_n' = f(t, x_1, x_2, \ldots, x_n),$$

which yields the system of ODEs

$$\begin{cases} x_1' &= x_2 \\ x_2' &= x_3 \\ x_{n-1}' &= x_n \\ x_n' &= f(t, x_1, x_2, \ldots, x_n) \end{cases}.$$

Example 19.11. Transform the third order nonhomogeneous ODE $y''' = y'' + 2y' - \sin(t)$ into a first order linear system.

Set $x_1 = y$, set $x_2 = y'$, and set $x_3 = y''$, so that $x_3' = y''' = y'' + 2y' - \sin(t)$, to obtain the system

$$x_1' = x_2, \quad x_2' = y'' = x_3, \quad \text{and} \quad x_3' = x_3 + 2x_2 - \sin(t),$$

which can be written as the first order linear system

$$\begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} x_2 \\ x_3 \\ 2x_2 + x_3 - \sin(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 2 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -\sin(t) \end{bmatrix}.$$

19.4. Numerical solutions to differential equations

Not all differential equations can be solved in closed form. Consider $x' = \exp\left\{-x^2\right\}$, and observe that $\exp\left\{-x^2\right\}$ has no closed form antiderivative. Differential equations such as these are typically solved numerically.

19.4.1. Euler's method. Consider the slope field associated with the differential equation $y' = f(t, y)$. The most natural computational procedure is to choose an initial point (t_0, y_0) and then move from that point in the direction indicated by the slope field. After moving a short distance, re-evaluate the direction at the point (t_1, y_1), and move a short distance in that direction. Repeat this process for the desired solution interval. More formally, given the initial value problem for the first order equation $y' = f(t, y)$ with $y(t_0) = y_0$ and $t \in [a, b]$, define a grid of $n + 1$ points

$$a = t_0 < t_1 < \cdots < t_n = b$$

for the independent variable t with equal step size h. From the formal definition of the first derivative

$$\frac{\mathrm{d}}{\mathrm{d}\,t} y(t) = \lim_{h \to 0} \frac{y(t+h) - y(t)}{h},$$

we have the first-order approximation

$$\frac{\mathrm{d}}{\mathrm{d}\,t} y(t) \approx \frac{y(t+h) - y(t)}{h} - \frac{h}{2} y''(c)$$

for $c \in [t, t+h]$; this approximation is called the *two-point forward difference*. Disregarding the second term on the right-hand side leads to the further approximation

$$\frac{\mathrm{d}}{\mathrm{d}\,t} y(t) = y'(t) \approx \frac{y(t+h) - y(t)}{h} \implies y(t+h) - y(t) \approx h y'(t) = h f(t, y(t)).$$

Definition 19.12 (Euler's method). The solution y to the initial value problem $y' = f(t, y(t))$ with $y(t_0) = y_0$ and $t \in [a, b]$ can be found by the sequence

$$y_{j+1} = y_j + h f(t_j, y_j(t))$$

for $j \in \{0, 1, \ldots, n-1\}$, where $y_j = y(t_j)$ and $h = (b-a)/n$.

Example 19.13. foo

Definition 19.14. Let $\mathbf{x} \in \mathbb{R}^n$ be the vector whose jth entry is the jth solution obtained by Euler's method to the initial value problem $y' = f(t, y(t))$ with $y(t_0) = y_0$, and let $\mathbf{x}_\mathrm{e} \in \mathbb{R}^n$ be the vector whose jth entry is the exact solution to the initial value problem at t_j, i.e., $(\mathbf{x}_e)_j = y(t_j)$. The *local error* associated with the numerical solution is

$$E_\mathrm{loc} = \|\mathbf{x} - \mathbf{x}_\mathrm{e}\|_\infty,$$

and the *global error* is

$$E_\mathrm{global} = \|\mathbf{x} - \mathbf{x}_\mathrm{e}\|_2.$$

We also define the *relative error* as

$$E_\mathrm{rel} = \frac{\|\mathbf{x} - \mathbf{x}_\mathrm{e}\|_2}{\|\mathbf{x}_\mathrm{e}\|_2}.$$

Definition 19.15. Consider a numerical solution to an initial value problem. A solution with p digits of *precision* if the local error of the solution satisfies

$$E_\mathrm{loc} < \frac{1}{2} \cdot 10^{-p}.$$

CHAPTER 20

Partial differential equations

20.1. Basic PDEs

A number of PDEs can be treated as ODEs. Recall that if the partial derivative of a function f with respect to a variable x is zero, then the function does not depend on x, i.e.,

$$\frac{\partial}{\partial x}F(x,t)=0 \implies F(x,t)=F(t).$$

20.1.1. First order basic PDEs. We begin by solving PDEs that are similar to separable ODEs.

EXAMPLE 20.1. Find the solution $u(x,t)$ of the following PDEs, where u_x denotes the partial derivative of $u(x,t)$ with respect to x.

(1) Suppose that $u_t=0$. It follows from our statement above that

$$\frac{\partial}{\partial t}u(x,t)=0 \implies u(x,t)=F(x),$$

i.e., $u(x,t)$ is a function of x alone.

(2) Suppose that $u_t=x^2-t^2$. Then,

$$\frac{\partial}{\partial t}u(x,t)=x^2-t^2 \implies \int\frac{\partial}{\partial t}u(x,t)\,\mathrm{d}\,t=\int x^2-t^2\,\mathrm{d}\,t \implies u(x,t)=tx^2-\frac{1}{3}t^3+C(x),$$

where $C(x)$ is a function that depends on x alone.

(3) Suppose that $uu_t=2x-t$. Then,

$$u(x,t)\frac{\partial}{\partial t}u(x,t)=2x-t \implies \int u(x,t)\frac{\partial}{\partial t}u(x,t)\,\mathrm{d}\,t=\int 2x-t\,\mathrm{d}\,t.$$

Let $u=u(x,t)$, so that

$$\frac{\mathrm{d}\,u}{\mathrm{d}\,t}=\frac{\mathrm{d}}{\mathrm{d}\,t}u(x,t)=\frac{\partial}{\partial t}u(x,t)\frac{\mathrm{d}}{\mathrm{d}\,t}t \implies \mathrm{d}\,u=\frac{\partial}{\partial t}u(x,t)\,\mathrm{d}\,t.$$

Then, the general solution is given by

$$\int u\,\mathrm{d}\,u=\int 2x-t\,\mathrm{d}\,t \implies \frac{1}{2}u^2=2xt-\frac{1}{2}t^2+C(x) \implies u(x,t)=\sqrt{4xt-t^2+2C(x)}.$$

(4) Suppose that $u^2u_x=xt+2$. Then,

$$(u(x,t))^2\frac{\partial}{\partial x}u(x,t)=xt+2 \implies \int(u(x,t))^2\frac{\partial}{\partial x}u(x,t)\,\mathrm{d}\,x=\int xt+2\,\mathrm{d}\,x.$$

Letting $u=u(x,t)$, it follows by symmetry from our result above that $\mathrm{d}\,u=u_x\,\mathrm{d}\,x$, so that

$$\int u^2\,\mathrm{d}\,u=\int xt+2\,\mathrm{d}\,x \implies \frac{1}{3}u^3=\frac{1}{2}tx^2+2x+C(t) \implies u(x,t)=\sqrt[3]{\frac{3}{2}tx^2+6x+3C(t)}.$$

We will solve the following examples, which resemble linear first order ODEs, by applying the method of integrating factor.

EXAMPLE 20.2. Find the solution $u(x,t)$ of the following PDEs.

(1) Suppose that $u_x + u = 0$. Let $h(x) = 1$, and observe that $H(x) = x$ is an antiderivative of $h(x)$. Multiplying both sides of the equation by $\mathrm{e}^{H(x)}$ gives

$$0 = u_x \mathrm{e}^x + u\mathrm{e}^x = \frac{\partial}{\partial x} u\mathrm{e}^x \implies \int \frac{\partial}{\partial x}\left(u\mathrm{e}^x\right) \mathrm{d}\, x = \int 0 \,\mathrm{d}\, x \implies u\mathrm{e}^x = C(t),$$

which leads to the general solution $u(x,t) = \mathrm{e}^{-x} C(t)$.

(2) Suppose that $u_x + ut^2 = 0$. Let $h(x) = t^2$, and observe that $H(x) = xt^2$ is an antiderivative of $h(x)$, so that

$$0 = u_x \mathrm{e}^{xt^2} + ut^2 \mathrm{e}^{xt^2} = \frac{\partial}{\partial x} u\mathrm{e}^{xt^2} \implies \int \frac{\partial}{\partial x}\left(u\mathrm{e}^{xt^2}\right) \mathrm{d}\, x = \int 0 \,\mathrm{d}\, x \implies u\mathrm{e}^{xt^2} = C(t),$$

which leads to the general solution $u(x,t) = C(t)\,\mathrm{e}^{-xt^2}$.

(3) Suppose that $u_t + u = x$. Let $h(t) = 1$, and observe that $H(t) = t$ is an antiderivative of $h(t)$, so that

$$x\mathrm{e}^t = u_t \mathrm{e}^t + u\mathrm{e}^t = \frac{\partial}{\partial t} u\mathrm{e}^t \implies \int \left(u\mathrm{e}^t\right) \mathrm{d}\, t = \int x\mathrm{e}^t \,\mathrm{d}\, t \implies u\mathrm{e}^t = x\mathrm{e}^t + C(x),$$

which leads to the general solution $u(x,t) = x + C(x)\,\mathrm{e}^{-t}$.

(4) Suppose that $u_t + 2(t-x)u = t - x$. Let $h(t) = 2t - 2x$, and observe that $H(t) = t^2 - 2xt$ is an antiderivative of $h(t)$, so that

$$(t-x)\,\mathrm{e}^{t^2-2xt} = u_t \mathrm{e}^{t^2-2xt} + 2(t-x)\,u\mathrm{e}^{t^2-2xt} = \frac{\partial}{\partial t} u\mathrm{e}^{t^2-2xt}.$$

Integrating both sides with respect to t gives

$$\int \frac{\partial}{\partial t}\left(u\mathrm{e}^{t^2-2xt}\right) \mathrm{d}\, t = \int (t-x)\,\mathrm{e}^{t^2-2xt} \,\mathrm{d}\, t \implies u\mathrm{e}^{t^2-2xt} = \frac{1}{2}\mathrm{e}^{t^2-2xt} + C(x),$$

which leads to the general solution

$$u(x,t) = \frac{1}{2} + C(x)\,\mathrm{e}^{2xt-t^2}.$$

20.1.2. Second order basic PDEs. As with first order basic PDEs, we can apply techniques from second order ODEs to solve certain second order PDEs. We can solve each of the examples below as if it were a nonhomogeneous second order ODE.

EXAMPLE 20.3. Find the solution $u(x,t)$ for the following PDEs, where u_{xx} denotes the second partial derivative of $u(x,t)$ with respect to x.

(1) Suppose that $u_{xx} + u = 2t$. We have $b = 0$ and $c = 1$, so that the characteristic equation is

$$r^2 + br + c = r^2 + 1 = 0 \implies r^2 = -1 \implies r = \pm\sqrt{-1} = \pm\iota,$$

i.e., the roots are $r_1 = \iota$ and $r_2 = -\iota$. Then, the homogeneous solution is

$$\begin{aligned} u_h(x,t) &= C_1(t)\,\mathrm{e}^{\iota x} + C_2(t)\,\mathrm{e}^{-\iota x} \\ &= C_1(t)\left(\cos(x) + \iota\sin(x)\right) + C_2(t)\left(\cos(-x) + \iota\sin(-x)\right) \quad \text{(Euler's Formula)} \\ &= C_1(t)\cos(x) + \iota C_1(t)\sin(x) + C_2(t)\cos(x) - \iota C_2(t)\sin(x) \\ &= \left(C_1(t) + C_2(t)\right)\cos(x) + \iota\left(C_1(t) - C_2(t)\right)\sin(x) \\ &= C_1(t)\cos(x) + C_2(t)\sin(x), \end{aligned}$$

where we have redefined the arbitrary functions $C_1(t)$ and $C_2(t)$. We have $m = 0$, so that the particular solution is $u_p(x,t) = Ax^s\mathrm{e}^{0\cdot x} = Ax^s$. If $s = 0$, then $u_p = A$, so that $\partial_x u_p = \partial_{xx} u_p = 0$. Then, the homogeneous equation becomes $u_{xx} + u = 0 + A \neq 0$, i.e., $u_p = A$ is not a solution of the homogeneous equation. Hence, we will take $s = 0$, so that the nonhomogeneous equation becomes

$$u_{xx} + u = A = 2t \implies A = 2t,$$

so that the general solution is given by

$$u(x,t) = u_h(x,t) + u_p(x,t) = C_1(t)\cos(x) + C_2(t)\sin(x) + 2t.$$

(2) Suppose that $u_{tt}+4u=\sin(t)$. We have $b=0$ and $c=4$, so that the characteristic equation is

$$r^2+br+c=r^2+4=0 \implies r^2=-4 \implies r=\pm\sqrt{4\cdot -1}=\pm 2\iota,$$

i.e., the roots are $r_1=2\iota$ and $r_2=-2\iota$. Then, the homogeneous solution is

$$\begin{aligned}u_h(x,t)&=C_1(x)\,\mathrm{e}^{2\iota t}+C_2(x)\,\mathrm{e}^{-2\iota t}\\&=C_1(x)(\cos(2t)+\iota\sin(2t))+C_2(x)(\cos(-2t)+\iota\sin(-2t))\\&=C_1(x)\cos(2t)+\iota C_1(x)\sin(2t)+C_2(x)\cos(2t)-\iota C_2(x)\sin(2t)\\&=(C_1(x)+C_2(x))\cos(2t)+\iota(C_1(x)-C_2(x))\sin(2t)\\&=C_1(x)\cos(2t)+C_2(x)\sin(2t),\end{aligned}$$

where we have again redefined the arbitrary functions $C_1(x)$ and $C_2(x)$. We have $m=0$ and $\gamma=1$, so that the particular solution will be of the form $u_p(x,t)=t^s(A\cos(t)+B\sin(t))$. If $s=0$, then $u_p=A\cos(t)+B\sin(t)$, so that

$$\partial_t u_p=-A\sin(t)+B\cos(t) \quad \text{and} \quad \partial_{tt}u_p=-A\cos(t)-B\sin(t).$$

We consider each term of the particular solution separately. For $A\cos(t)$, the homogeneous equation is

$$u_{tt}+4u=-A\cos(t)-B\sin(t)+4A\cos(t)=3A\cos(t)-B\sin(t),$$

which will not in general be equal to zero. For $B\sin(t)$, we have

$$u_{tt}+4u=-A\cos(t)-B\sin(t)+4B\sin(t)=-A\cos(t)+3B\sin(t),$$

which will also not in general be equal to zero. Hence, we will take $s=0$, so that the nonhomogeneous equation becomes

$$u_{tt}+4u=-A\cos(t)-B\sin(t)+4(A\cos(t)+B\sin(t))=3A\cos(t)+3B\sin(t)=\sin(t),$$

which implies that $A=0$ and $B=1/3$. Then, the general solution is given by

$$u(x,t)=u_h(x,t)+u_p(x,t)=C_1(x)\cos(2t)+C_2(x)\sin(2t)+\frac{1}{3}\sin(t).$$

(3) Suppose that $u_{tt}+2u_t-15u=xt^2$. We have $b=2$ and $c=-15$, so that the characteristic equation is

$$r^2+br+c=r^2+2r-15=(r+5)(r-3)=0,$$

i.e., the roots are $r_1=-5$ and $r_2=3$. Then, the homogeneous solution is

$$u_h(x,t)=C_1(x)\,\mathrm{e}^{-5t}+C_2(x)\,\mathrm{e}^{3t}.$$

To find the particular solution, we will apply the method of undetermined coefficients. The forcing term is xt^2, which has derivatives (with respect to t) of xt and x (ignoring multiplicative constants). We will seek $u_p(x,t)$ as a linear combination of these terms, i.e., $u_p=Axt^2+Bxt+Cx$, so that $\partial_t u_p=2Axt+Bx$ and $\partial_{tt}u_p=2Ax$. We see that none of the terms in u_p is included in u_h, so that the nonhomogeneous equation becomes

$$\begin{aligned}u_{tt}+2u_t-15u&=2Ax+2(2Axt+Bx)-15\left(Axt^2+Bxt+Cx\right)\\&=-15Axt^2+(4A-15B)xt+(2A+2B-15C)x\\&=xt^2,\end{aligned}$$

which leads to

$$-15A=1 \implies A=-\frac{1}{15}, \quad 4A-15B=0 \implies B=\frac{4}{15}A=-\frac{4}{15^2},$$

and

$$0=2A+2B-15C=-\frac{2}{15}-\frac{8}{15^2}-15C \implies C=-\frac{2}{15^2}-\frac{8}{15^3}=-\frac{38}{15^3}.$$

Then, the general solution is given by

$$\begin{aligned}u(x,t)&=u_h(x,t)+u_p(x,t)\\&=C_1(x)\,\mathrm{e}^{-5t}+C_2(x)\,\mathrm{e}^{3t}-\frac{1}{15}xt^2-\frac{4}{15^2}xt-\frac{38}{15^3}x\end{aligned}$$

$$= C_1(x)\,\mathrm{e}^{-5t} + C_2(x)\,\mathrm{e}^{3t} - x\left(\frac{225}{15^3}t^2 + \frac{60}{15^3}t + \frac{38}{15^3}\right)$$
$$= C_1(x)\,\mathrm{e}^{-5t} + C_2(x)\,\mathrm{e}^{3t} - \frac{\left(225t^2 + 60t + 38\right)x}{3375}.$$

(4) Suppose that $u_{xx} + 2u_x - 15u = \sin(t)$. By symmetry, the homogeneous solution is

$$u_h(x,t) = C_1(t)\,\mathrm{e}^{-5x} + C_2(t)\,\mathrm{e}^{3x}.$$

The forcing term is $\sin(t)$, which has derivative $\cos(t)$. We will seek $u_p(x,t)$ as a linear combination of these terms, i.e., $u_p(x,t) = A\cos(t) + B\sin(t)$, so that $\partial_x u_p = \partial_{xx} u_p = 0$. We see that none of the terms in u_p is included in u_h, so that the nonhomogeneous equation becomes

$$u_{xx} + 2u_x - 15u = 0 + 2\cdot 0 - 15\left(A\cos(t) + B\sin(t)\right) = -15A\cos(t) - 15B\sin(t) = \sin(t),$$

which implies that $A = 0$ and $B = -1/15$. Then, the general solution is given by

$$u(x,t) = u_h(x,t) + u_p(x,t) = C_1(t)\,\mathrm{e}^{-5x} + C_2(t)\,\mathrm{e}^{3x} - \frac{1}{15}\sin(t).$$

20.2. The method of characteristics

20.2.1. Second order PDE with boundary conditions. We will find the solution $u(x,t)$ to the PDE $u_{tx} = f(x,t)$ for $x,t > 0$ that satisfies the auxiliary conditions $u(x,0) = g(x)$ for $x > 0$ and $u(0,t) = h(t)$ for $t > 0$, where f, g, and h are well-behaved functions with $g(0) = h(0)$ and $g'(0) = h'(0)$. We begin by finding the homogeneous solution, i.e., the solution to

$$0 = \frac{\partial^2 u_1}{\partial x \partial t} = \frac{\partial}{\partial x}\left(\frac{\partial u_1}{\partial t}\right),$$

which implies that $u_1^{(0,1)}$ must be a function of t alone, i.e., $u_1^{(0,1)} = f_1(t)$, where $u_1^{(n,k)}$ denotes the nth partial derivative of u_1 with respect to x and the kth partial derivative with respect to t. Integrating both sides with respect to t gives

$$\int u_1^{(0,1)}\,\mathrm{d}\,t = \int f_1(t)\,\mathrm{d}\,t \implies u_1 = F_1(t) + c_1(x),$$

where F_1 is an antiderivative of f_1. Assuming that $u_1(x,t)$ has a continuous second derivative, it follows by symmetry that

$$u_1 = G_1(x) + c_2(t),$$

so that the homogeneous solution is

$$u_1(x,t) = F_1(t) + G_1(x), \tag{20.2.1}$$

where the functions $c_1(x)$ and $c_2(t)$ have been "absorbed" by the (arbitrary) antiderivatives G_1 and F_1, respectively. We have $x,t > 0$, so we obtain the particular solution by integrating $f(x,t)$ over $(0,t]$ and $(0,x]$, i.e.,

$$\int_0^x \int_0^t \frac{\partial^2 u_2}{\partial x \partial t}\,\mathrm{d}\,\tau\,\mathrm{d}\,\xi = u_2(x,t) = \int_0^x \int_0^t f(\tau,\xi)\,\mathrm{d}\,\tau\,\mathrm{d}\,\xi.$$

Then, the general solution is

$$u(x,t) = u_1(x,t) + u_2(x,t) = F(t) + G(x) + \int_0^x \int_0^t f(\tau,\xi)\,\mathrm{d}\,\tau\,\mathrm{d}\,\xi,$$

where we have dropped the subscripts on F and G for clarity. Now, the auxiliary condition $u(x,0) = g(x)$ implies that

$$u(x,0) = F(0) + G(x) + \int_0^x \int_0^0 f(\tau,\xi)\,\mathrm{d}\,\tau\,\mathrm{d}\,\xi$$
$$= F(0) + G(x) + \int_0^x 0\,\mathrm{d}\,\xi$$
$$= F(0) + G(x)$$
$$= g(x),$$

and by symmetry, the condition $u(0,t) = h(t)$ implies that

$$u(0,t) = F(t) + G(0) = h(t).$$

Observing that

$$G(x) = g(x) - F(0) \quad \text{and} \quad F(t) = h(t) - G(0),$$

it follows that

$$G(0) = g(0) - F(0) \implies g(0) = F(0) + G(0)$$

and

$$F(0) = h(0) - G(0) \implies h(0) = F(0) + G(0),$$

so that $g(0) = h(0)$. Finally, noting that

$$F(t) + G(x) = h(t) - G(0) + g(x) - F(0) = h(t) + g(x) - (F(0) + G(0)) = h(t) + g(x) - h(0),$$

we can write the general solution as

$$u(x,t) = h(t) + g(x) - h(0) + \int_0^x \int_0^t f(\tau,\xi)\,\mathrm{d}\tau\,\mathrm{d}\xi. \tag{20.2.2}$$

EXAMPLE 20.4. Find the solution $u(x,t)$ to the PDE $u_{tx} = xt$ for $x,t > 0$ that satisfies $u(x,0) = \sin(x)$ and $u(0,t) = \cos(t) - 1$.

Equation (20.2.2) implies that the general solution is given by

$$\begin{aligned}
u(x,t) &= \cos(t) - 1 + \sin(x) - (\cos(0) - 1) + \int_0^x \int_0^t \xi\tau\,\mathrm{d}\tau\xi \\
&= \cos(t) - 1 + \sin(x) - (1-1) + \int_0^x \xi \left[\frac{1}{2}\tau^2\right|_0^t \mathrm{d}\xi \\
&= \cos(t) - 1 + \sin(x) + \int_0^x \xi\left(\frac{1}{2}t^2 - 0\right)\mathrm{d}\xi \\
&= \cos(t) - 1 + \sin(x) + \frac{1}{2}t^2\left[\frac{1}{2}\xi^2\right|_0^x \\
&= \cos(t) - 1 + \sin(x) + \frac{1}{4}x^2t^2.
\end{aligned}$$

20.2.2. Chain rule. If we have the coordinates (x,y), then we can always express a new set of coordinate (η,ξ) by the substitution $x = x(\eta,\xi)$ and $y = y(\eta,\xi)$, or equivalently $\eta = \eta(x,y)$ and $\xi = \xi(x,y)$. Given a function $f = f(x,y)$, it is sometimes more convenient to represent the function in the new coordinates, i.e., $f(x(\eta,\xi), y(\eta,\xi)) = f(\eta,\xi)$, which necessitates computing the partial derivatives of f with respect to the new coordinates. These derivatives are obtained by the chain rule, i.e.,

$$\frac{\partial}{\partial x} f(\eta,\xi) = \frac{\partial}{\partial \eta(x,y)} f(\eta(x,y), \xi(x,y)) \frac{\partial}{\partial x}\eta(x,y) + \frac{\partial}{\partial \xi(x,y)} f(\eta(x,y), \xi(x,y)) \frac{\partial}{\partial \xi}\xi(x,y),$$

or simplifying notation,

$$\frac{\partial f}{\partial x} = \frac{\partial f}{\partial \eta}\frac{\partial \eta}{\partial x} + \frac{\partial f}{\partial \xi}\frac{\partial \xi}{\partial x}.$$

By symmetry, we have

$$\frac{\partial f}{\partial y} = \frac{\partial f}{\partial \eta}\frac{\partial \eta}{\partial y} + \frac{\partial f}{\partial \xi}\frac{\partial \xi}{\partial y}.$$

EXAMPLE 20.5. Let

$$f(x,y) = \frac{1}{\sqrt{x^2+y^2}},$$

and convert to polar coordinates, i.e., set $x = r\cos(\theta)$ and set $y = r\sin(\theta)$ for $r \geq 0$ and $\theta \in [0, 2\pi]$. Then,

$$f(r,\theta) = \frac{1}{\sqrt{(r\cos(\theta))^2 + (r\sin(\theta))^2}} = \frac{1}{\sqrt{r^2\left(\cos^2(\theta) + \sin^2(\theta)\right)}} = \frac{1}{\sqrt{r^2}} = \frac{1}{r}.$$

The first derivatives in terms of the new coordinates (r,θ) are

$$\frac{\partial f}{\partial x}=\frac{\partial f}{\partial r}\frac{\partial r}{\partial x}+\frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial x}\quad\text{and}\quad\frac{\partial f}{\partial y}=\frac{\partial f}{\partial r}\frac{\partial r}{\partial y}+\frac{\partial f}{\partial \theta}\frac{\partial \theta}{\partial y}. \tag{20.2.3}$$

Noting that

$$x=r\cos(\theta)\implies x^2=r^2\cos^2(\theta)\quad\text{and}\quad y=r\sin(\theta)\implies y^2=r^2\sin^2(\theta),$$

we have

$$x^2+y^2=r^2\cos^2(\theta)+r^2\sin^2(\theta)=r^2\left(\cos^2(\theta)+\sin^2(\theta)\right)\implies r^2=x^2+y^2\implies r=\sqrt{x^2+y^2}.$$

Similarly, we have

$$\frac{y}{x}=\frac{r\sin(\theta)}{r\cos(\theta)}=\tan(\theta)\implies\theta=\arctan\left(\frac{y}{x}\right).$$

Then, the necessary partial derivatives are

$$\frac{\partial f}{\partial r}=-\frac{1}{r^2},\quad\frac{\partial f}{\partial\theta}=0,\quad\frac{\partial r}{\partial x}=\frac{1}{2\sqrt{x^2+y^2}}\cdot 2x=\frac{x}{\sqrt{x^2+y^2}},\quad\text{and}\quad\frac{\partial r}{\partial y}=\frac{y}{\sqrt{x^2+y^2}},$$

so that

$$\frac{\partial f}{\partial x}=-\frac{1}{r^2}\cdot\frac{x}{\sqrt{x^2+y^2}}+0\cdot\frac{\partial\theta}{\partial x}=-\frac{r\cos(\theta)}{r^3}=-\frac{\cos(\theta)}{r^2}$$

and

$$\frac{\partial f}{\partial y}=-\frac{1}{r^2}\cdot\frac{y}{\sqrt{x^2+y^2}}+0\cdot\frac{\partial\theta}{\partial y}=-\frac{r\sin(\theta)}{r^3}=-\frac{\sin(\theta)}{r^2}.$$

EXAMPLE 20.6. Let F be a function of (x,t), let $\eta=x+ct$, and let $\xi=x-ct$ for some $c\in\mathbb{R}$. Find expressions for F_x and F_t with respect to the partial derivatives of the new coordinates (η,ξ).

From (20.2.3), we have

$$\frac{\partial F}{\partial x}=\frac{\partial F}{\partial\eta}\frac{\partial\eta}{\partial x}+\frac{\partial F}{\partial\xi}\frac{\partial\xi}{\partial x}=F_\eta+F_\xi\quad\text{and}\quad\frac{\partial F}{\partial t}=\frac{\partial F}{\partial\eta}\frac{\partial\eta}{\partial t}+\frac{\partial F}{\partial\xi}\frac{\partial\xi}{\partial t}=cF_\eta-cF_\xi=c\left(F_\eta-F_\xi\right).$$

20.2.3. The advection equation. The advection equation, a special case of the wave equation, models the behavior of acoustic waves and is defined as $u_t+cu_x=0$, where c represents wave speed. The general solution to this equation is obtained from the chain rule. Define the new variables $\eta=x+ct$ and $\xi=x-ct$, and observe that

$$0=u_t+cu_x=c\left(u_\eta-u_\xi\right)+c\left(u_\eta+u_\xi\right)=2cu_\eta.$$

Now, if the partial derivative of $u(\eta,\xi)$ with respect to η is zero, then it follows that $u(\eta,\xi)$ is a function of ξ only, i.e., $u(\eta,\xi)=F(\xi)$, so that the general solution is $u(x,t)=F(x-ct)$. The lines $x-ct=k$ for some $k\in\mathbb{R}$ are called *characteristic curves*.

EXAMPLE 20.7. Show that the following are solutions to the advection equation.

(1) Let $u(x,t)=\cos(x-ct)$. Then,

$$u_x=-\sin(x-ct)\text{ and }u_t=c\sin(x-ct)\implies u_t+cu_x=c\sin(x-ct)-c\sin(x-ct)=0.$$

(2) Let $u(x,t)=\sin(x-ct)$. Then,

$$u_x=\cos(x-ct)\text{ and }u_t=-c\cos(x-ct)\implies u_t+cu_x=-c\cos(x-ct)+c\cos(x-ct)=0.$$

(3) Let $u(x,t)=\mathrm{e}^{\iota(x-ct)}$. Then,

$$u_x=\iota\mathrm{e}^{\iota(x-ct)}\text{ and }u_t=-c\iota\mathrm{e}^{\iota(x-ct)}\implies u_t+cu_x=-c\iota\mathrm{e}^{\iota(x-ct)}+c\iota\mathrm{e}^{\iota(x-ct)}=0.$$

The typical wave behavior is given by the solution $u(x,t)$. We will now see how to use the general solution to set boundary conditions like $u(x,0)=f(x)$.

EXAMPLE 20.8. Find the solution $u(x,t)$ to the advection equation with boundary condition $u(x,0)=f(x)$ for the following functions $f(x)$.

(1) Let $f(x)=\cos(x)$. The general solution is $u(x,t)=F(x-ct)$, so that

$$F(x)=F(x-c\cdot 0)=u(x,0)=\cos(x)\implies u(x,t)=\cos(x-ct).$$

(2) Let $f(x) = x + 3$. Then,

$$F(x) = u(x,0) = x + 3 \implies u(x,t) = x - ct + 3.$$

(3) Let $f(x) = 1/\left(1 + x^2\right)$. Then,

$$F(x) = u(x,0) = \frac{1}{1+x^2} \implies u(x,t) = \frac{1}{1+(x-ct)^2}.$$

In the previous examples, and in our general solution to the advection equation, we have assumed that the wave speed c is constant. Suppose now that x is a function of t, i.e., $x = x(t)$. Then, it follows from the chain rule that

$$\frac{\mathrm{d}\,u}{\mathrm{d}\,t} = \frac{\partial}{\partial t}u(x(t),t) = \frac{\partial u}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial u}{\partial t}\frac{\partial t}{\partial t} = \frac{\partial x}{\partial t}u_x + u_t,$$

which implies that

$$\frac{\partial x}{\partial t} = c(x,t).$$

In this case, although we cannot find a general solution using a change of variable, the equation for the characteristic curves $x_t = c(x,t)$ holds.

EXAMPLE 20.9. Solve the advection equation with $u_t + t^2 u_x = 0$ and boundary condition $u(x,0) = \cos(x)$.

We have

$$\frac{\partial x}{\partial t} = c(x,t) = t^2 \implies \int \frac{\partial x}{\partial t}\,\mathrm{d}\,t = \int t^2\,\mathrm{d}\,t \implies x(t) = \frac{1}{3}t^3 + K \implies K = x(t) - \frac{1}{3}t^3$$

for some $K \in \mathbb{R}$. Noting that the solution $u(x(t),t)$ at the characteristic is constant, applying the boundary condition gives

$$u(x(0),0) = \cos(x(0)) = \cos\left(\frac{1}{3}\cdot 0 + K\right) = \cos(K) = \cos\left(x(t) - \frac{1}{3}t^3\right),$$

so that the general solution is

$$u(x,t) = \cos\left(x - \frac{1}{3}t^3\right).$$

EXAMPLE 20.10. Solve the advection equation with $u_t - 2tx^2 u_x = 0$ and boundary condition $u(x,0) = \cos(x)$.

We have

$$\frac{\partial x}{\partial t} = -2tx^2 \implies \frac{1}{x^2}\frac{\partial x}{\partial t} = -2t \implies \int \frac{1}{x^2}\frac{\partial x}{\partial t}\,\mathrm{d}\,t = \int -2t\,\mathrm{d}\,t \implies -\frac{1}{x} = -t^2 + K,$$

which leads to

$$x(t)\left(t^2 - K\right) = 1 \implies x(t) = \frac{1}{t^2 - K} \quad \text{and} \quad K = t^2 - \frac{1}{x(t)}.$$

Applying the boundary condition gives

$$u(x(0),0) = \cos(x(0)) = \cos\left(\frac{1}{0-K}\right) = \cos\left(\frac{1}{K}\right) = \cos\left(\frac{1}{t^2 - 1/x(t)}\right) = \cos\left(\frac{x(t)}{x(t)\,t^2 - 1}\right)$$

so that the general solution is

$$u(x,t) = \cos\left(\frac{x}{xt^2 - 1}\right).$$

20.3. The wave equation

The one-dimensional wave equation is defined as $u_{tt} - c^2 u_{xx} = 0$, and is the most general way of modeling the physical behavior of waves. Define the variables $\eta = x + ct$ and $\xi = x - ct$, and observe that for some function $F(x,t)$, we have

$$\begin{aligned}
\frac{\partial^2 F}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial x}\right) \\
&= \frac{\partial}{\partial x}\left(\frac{\partial}{\partial x}F(\eta,\xi)\right) \\
\text{(chain rule)} \qquad &= \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \eta}\frac{\partial \eta}{\partial x} + \frac{\partial F}{\partial \xi}\frac{\partial \xi}{\partial x}\right) \\
&= \frac{\partial}{\partial x}\left(\frac{\partial F}{\partial \eta}\cdot 1 + \frac{\partial F}{\partial \xi}\cdot 1\right) \\
\text{(chain rule)} \qquad &= \frac{\partial}{\partial \eta}\left(\frac{\partial F}{\partial \eta} + \frac{\partial F}{\partial \xi}\right)\frac{\partial \eta}{\partial x} + \frac{\partial}{\partial \xi}\left(\frac{\partial F}{\partial \eta} + \frac{\partial F}{\partial \xi}\right)\frac{\partial \xi}{\partial x} \\
&= \frac{\partial}{\partial \eta}\left(\frac{\partial F}{\partial \eta} + \frac{\partial F}{\partial \xi}\right)\cdot 1 + \frac{\partial}{\partial \xi}\left(\frac{\partial F}{\partial \eta} + \frac{\partial F}{\partial \xi}\right)\cdot 1 \\
&= \frac{\partial^2 F}{\partial \eta^2} + \frac{\partial^2 F}{\partial \eta \partial \xi} + \frac{\partial^2 F}{\partial \xi \partial \eta} + \frac{\partial^2 F}{\partial \xi^2} \\
&= \frac{\partial^2 F}{\partial \eta^2} + 2\frac{\partial^2 F}{\partial \eta \partial \xi} + \frac{\partial^2 F}{\partial \xi^2},
\end{aligned}$$

where the final equality follows from our assumption that F has continuous second partial derivatives. Similarly, we have

$$\begin{aligned}
\frac{\partial^2 F}{\partial t^2} &= \frac{\partial}{\partial t}\left(\frac{\partial F}{\partial \eta}\frac{\partial \eta}{\partial t} + \frac{\partial F}{\partial \xi}\frac{\partial \xi}{\partial t}\right) \\
&= \frac{\partial}{\partial t}\left(\frac{\partial F}{\partial \eta}\cdot c - \frac{\partial F}{\partial \xi}\cdot c\right) \\
&= c\frac{\partial}{\partial t}\left(\frac{\partial F}{\partial \eta} - \frac{\partial F}{\partial \xi}\right) \\
&= c\left[\frac{\partial}{\partial \eta}\left(\frac{\partial F}{\partial \eta} - \frac{\partial F}{\partial \xi}\right)\frac{\partial \eta}{\partial t} + \frac{\partial}{\partial \xi}\left(\frac{\partial F}{\partial \eta} - \frac{\partial F}{\partial \xi}\right)\frac{\partial \xi}{\partial t}\right] \\
&= c\left[\frac{\partial}{\partial \eta}\left(\frac{\partial F}{\partial \eta} - \frac{\partial F}{\partial \xi}\right)\cdot c - \frac{\partial}{\partial \xi}\left(\frac{\partial F}{\partial \eta} - \frac{\partial F}{\partial \xi}\right)c\right] \\
&= c^2\left(\frac{\partial^2 F}{\partial \eta^2} - \frac{\partial^2 F}{\partial \eta \partial \xi} - \frac{\partial^2 F}{\partial \xi \partial \eta} + \frac{\partial^2 F}{\partial \xi^2}\right) \\
&= c^2\left(\frac{\partial^2 F}{\partial \eta^2} - 2\frac{\partial^2 F}{\partial \eta \partial \xi} + \frac{\partial^2 F}{\partial \xi^2}\right).
\end{aligned}$$

For $F = u(x,t)$, we have

$$0 = \frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = c^2\left(\frac{\partial^2 u}{\partial \eta^2} - 2\frac{\partial^2 u}{\partial \eta \partial \xi} + \frac{\partial^2 u}{\partial \xi^2}\right) - c^2\left(\frac{\partial^2 u}{\partial \eta^2} + 2\frac{\partial^2 u}{\partial \eta \partial \xi} + \frac{\partial^2 u}{\partial \xi^2}\right) = -4c^2\frac{\partial^2 u}{\partial \eta \partial \xi},$$

which implies that the general solution of the wave equation satisfies

$$\frac{\partial^2 u}{\partial \eta \partial \xi} = 0.$$

This is a second-order homogeneous PDE, so it follows from (20.2.1) that $u(\eta,\xi) = F(\xi) + G(\eta)$, which leads to the general solution

$$u(x,t) = F(x - ct) + G(x + ct).$$

EXAMPLE 20.11. Show that the following are solutions to the wave equation.

(1) Let $u(x,t) = \cos(x-ct) + \cos(x+ct)$. Then,

$$u_{tt} = \frac{\partial}{\partial t}\left(c\sin(x-ct) - c\sin(x+ct)\right) = -c^2\cos(x-ct) - c^2\cos(x+ct)$$

and

$$u_{xx} = \frac{\partial}{\partial x}\left(-\sin(x-ct) - \sin(x+ct)\right) = -\cos(x-ct) - \cos(x+ct),$$

so that

$$u_{tt} - c^2 u_{xx} = -c^2\cos(x-ct) - c^2\cos(x+ct) + c^2\cos(x-ct) + c^2\cos(x+ct) = 0.$$

(2) Let $u(x,t) = \sin(x-ct) + \sin(x+ct)$. Then,

$$u_{tt} = \frac{\partial}{\partial t}\left(-c\cos(x-ct) + c\cos(x+ct)\right) = -c^2\sin(x-ct) - c^2\sin(x+ct)$$

and

$$u_{xx} = \frac{\partial}{\partial x}\left(\cos(x-ct) + \cos(x+ct)\right) = -\sin(x-ct) - \sin(x+ct),$$

so that

$$u_{tt} - c^2 u_{xx} = -c^2\sin(x-ct) - c^2\sin(x+ct) + c^2\sin(x-ct) + c^2\sin(x+ct) = 0.$$

20.3.1. Boundary conditions. The boundary conditions for the wave equation as given by $u(x,0) = f(x)$ and $u_t(x,0) = g(x)$, which correspond to the position and speed of the wave at time $t = 0$. From the general solution, we have

$$u(x,0) = F(x - c\cdot 0) + G(x + c\cdot 0) = F(x) + G(x) = f(x)$$

and

$$u_t(x,t) = -cF'(x-ct) + cG'(x+ct) \implies u_t(x,0) = -cF'(x) + cG'(x) = g(x).$$

Integrating this last expression, we obtain

$$\begin{aligned}
\int_0^x g(y)\,\mathrm{d}y &= \int_0^x -cF'(y) + cG'(y)\,\mathrm{d}y \\
&= \left[-cF(y) + cG(y)\right|_0^x \\
&= -cF(x) + cG(x) + cF(0) - cG(0) \\
\implies cF(x) - cG(x) &= cF(0) - cG(0) - \int_0^x g(y)\,\mathrm{d}y \\
\implies F(x) - G(x) &= F(0) - G(0) - \frac{1}{c}\int_0^x g(y)\,\mathrm{d}y.
\end{aligned}$$

Adding this to our expression for $u(x,0)$ gives

$$\begin{aligned}
F(x) + G(x) + F(x) - G(x) &= f(x) + F(0) - G(0) - \frac{1}{c}\int_0^x g(y)\,\mathrm{d}y \\
\implies 2F(x) &= f(x) + F(0) - G(0) - \frac{1}{c}\int_0^x g(y)\,\mathrm{d}y \\
\implies F(x) &= \frac{1}{2}f(x) - \frac{1}{2c}\int_0^x g(y)\,\mathrm{d}y + \frac{1}{2}\left(F(0) - G(0)\right),
\end{aligned}$$

and taking the difference gives

$$\begin{aligned}
F(x) + G(x) - F(x) + G(x) &= f(x) - F(0) + G(0) + \frac{1}{c}\int_0^x g(y)\,\mathrm{d}y \\
\implies 2G(x) &= f(x) - F(0) + G(0) + \frac{1}{c}\int_0^x g(y)\,\mathrm{d}y \\
\implies G(x) &= \frac{1}{2}f(x) + \frac{1}{2c}\int_0^x g(y)\,\mathrm{d}y - \frac{1}{2}\left(F(0) - G(0)\right).
\end{aligned}$$

Substituting these results into the general solution, we have

$$\begin{aligned}
u(x,t) &= F(x-ct) + G(x+ct) \\
&= \frac{1}{2}f(x-ct) - \frac{1}{2c}\int_0^{x-ct} g(y)\,\mathrm{d}\,y + \frac{1}{2}f(x+ct) + \frac{1}{2c}\int_0^{x+ct} g(y)\,\mathrm{d}\,y \\
&= \frac{1}{2}\left(f(x-ct)+f(x+ct)\right) + \frac{1}{2c}\left(\int_0^{x+ct} g(y)\,\mathrm{d}\,y - \int_0^{x-ct} g(y)\,\mathrm{d}\,y\right) \\
&= \frac{1}{2}\left(f(x-ct)+f(x+ct)\right) + \frac{1}{2c}\left(\int_0^{x-ct} g(y)\,\mathrm{d}\,y + \int_{x-ct}^{x+ct} g(y)\,\mathrm{d}\,y - \int_0^{x-ct} g(y)\,\mathrm{d}\,y\right) \\
&= \frac{1}{2}\left(f(x-ct)+f(x+ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} g(y)\,\mathrm{d}\,y,
\end{aligned}$$

where the final equality is referred to as *D'Alembert's solution.*

EXAMPLE 20.12. Find the solution to the wave equation with the given boundary conditions $u(x,0) = f(x)$ and $u_t(x,0) = g(x)$.

(1) Let $f(x) = \cos(x)$ and let $g(x) = 0$. D'Alembert's solution gives

$$u(x,t) = \frac{1}{2}\left(\cos(x-ct)+\cos(x+ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} 0\,\mathrm{d}\,y = \frac{1}{2}\left(\cos(x-ct)+\cos(x+ct)\right).$$

(2) Let $f(x) = \cos(x)$ and let $g(x) = \sin(x)$. D'Alembert's solution gives

$$\begin{aligned}
u(x,t) &= \frac{1}{2}\left(\cos(x-ct)+\cos(x+ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} \sin(y)\,\mathrm{d}\,y \\
&= \frac{1}{2}\left(\cos(x-ct)+\cos(x+ct)\right) + \frac{1}{2c}\left[-\cos(y)\right|_{x-ct}^{x+ct} \\
&= \frac{1}{2}\left(\cos(x-ct)+\cos(x+ct)\right) + \frac{1}{2c}\left(-\cos(x+ct)+\cos(x-ct)\right) \\
&= \frac{1}{2c}\left(c\cos(x-ct)+\cos(x-ct)+c\cos(x+ct)-\cos(x+ct)\right) \\
&= \frac{1}{2c}\left[(c+1)\cos(x-ct)+(c-1)\cos(x+ct)\right].
\end{aligned}$$

(3) Let $g(x) = 0$ and let

$$f(x) = \begin{cases} 1, & -1 \le x \le 1 \\ 0, & \text{otherwise} \end{cases}.$$

Noting that

$$-1 \le x+ct \le 1 \implies -1-ct \le x \le 1-ct \text{ and } -1 \le x-ct \le 1 \implies -1+ct \le x \le 1+ct,$$

we can set

$$f(x-ct) = \begin{cases} 1, & -1+ct \le x \le 1+ct \\ 0, & \text{otherwise} \end{cases}$$

and

$$f(x+ct) = \begin{cases} 1, & -1-ct \le x \le 1-ct \\ 0, & \text{otherwise} \end{cases},$$

which leads to the solution

$$u(x,t) = \frac{1}{2}\left(f(x-ct)+f(x+ct)\right) + \frac{1}{2c}\int_{x-ct}^{x+ct} 0\,\mathrm{d}\,y = \frac{1}{2}\left(f(x-ct)+f(x+ct)\right).$$

The region between $x+ct=-1$ and $x-ct=1$ is called the *region of influence.*

CHAPTER 21

Dimensionless equations and scaling

Many physical problems can be expressed in clearer and more unified form if the variables corresponding to physical units are replaced by dimensionless equivalents. It often happens that this process of rescaling sheds light on the physical process that the equations describe. It is frequently the case that essential characteristics of physical systems can be inferred from dimensionless analysis, which does not involve actually solving the given system.

THEOREM 21.1 (Buckingham Pi Theorem). *Let $f(q_1, q_2, \ldots, q_m) = 0$ be a unit-free law where $[q_i] = L_1^{\alpha_{1_i}} L_2^{\alpha_{2_i}} \cdots L_n^{\alpha_{n_i}}$. Suppose the fundamental matrix $\mathbf{A}$ has rank r. Then there exist $m-r$ independent dimensionless quantities $\pi_1, \ldots \pi_{m-r}$ which can be formed from $q_1, \ldots, q_m$ and a physical law $F(\pi_1, \ldots, \pi_{m-r}) = 0$ that is equivalent to $f(q_1, q_2, \ldots, q_m) = 0$.*

EXAMPLE 21.2 (Bead falling in dense liquid). A small sphere of radius r and density ρ falls with constant velocity in a liquid of density ρ_ℓ and viscosity μ (units mass/ (length $\cdot$ time)) under the influence of gravity g. An empirical formula is

$$v = \frac{2r^2 \rho g}{9\mu}\left(1 - \frac{\rho_\ell}{\rho}\right).$$

Can this be deduced from dimensional analysis?

The physical dimensions in this problem are length L, mass M, and time T. We can express the variables in terms of these units as

$$[r] = L, \quad [v] = \frac{L}{T}, \quad [\rho] = \frac{M}{L^3}, \quad [g] = \frac{L}{T^2}, \quad [\rho_\ell] = \frac{M}{L^3}, \quad \text{and} \quad [\mu] = \frac{M}{LT}.$$

The postulated physical law is of the form $f(v, r, \rho, g, \rho_\ell, \mu) = 0$. Let π be a dimensionless combination of all variables in the problem. The possible forms of π can be written as

$$\pi = r^{\alpha_1} \rho^{\alpha_2} v^{\alpha_3} g^{\alpha_4} \rho_\ell^{\alpha_5} \mu^{\alpha_6}.$$

Setting the units equal gives

$$\pi = L^{\alpha_1}\left(\frac{M}{L^3}\right)^{\alpha_2}\left(\frac{L}{T}\right)^{\alpha_3}\left(\frac{L}{T^2}\right)^{\alpha_4}\left(\frac{M}{L^3}\right)^{\alpha_5}\left(\frac{M}{LT}\right)^{\alpha_6},$$

which leads to the system

$$\begin{aligned} L: &\quad \alpha_1 - 3\alpha_2 + \alpha_3 + \alpha_4 - 3\alpha_5 - \alpha_6 = 0 \\ M: &\quad \alpha_2 + \alpha_5 + \alpha_6 = 0 \\ T: &\quad -\alpha_3 - 2\alpha_4 - \alpha_6 = 0. \end{aligned}$$

Equivalently, we have

$$\begin{bmatrix} 1 & -3 & 1 & 1 & -3 & -1 \\ 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & -1 & -2 & 0 & -1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} = \mathbf{0},$$

which we solve as

$$\left[\begin{array}{cccccc|c} 1 & -3 & 1 & 1 & -3 & -1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & -1 & -2 & 0 & -1 & 0 \end{array}\right] \begin{array}{l} \\ +3R_2 \\ \times -1 \end{array} \sim \left[\begin{array}{cccccc|c} 1 & 0 & 1 & 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 & 1 & 0 \end{array}\right] \begin{array}{l} \\ \\ -R_3 \end{array}$$

$$\sim \left[\begin{array}{cccccc|c} 1 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 2 & 0 & 1 & 0 \end{array}\right],$$

so that $\alpha_1 = \alpha_4 - \alpha_6$, $\alpha_2 = -\alpha_5 - \alpha_6$, and $\alpha_3 = -2\alpha_4 - \alpha_6$. We can express the solution as

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{bmatrix} = \alpha_4 \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix} + \alpha_5 \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} + \alpha_6 \begin{bmatrix} -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix},$$

so that the independent vectors

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ -2 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \\ -1 \\ 0 \\ 0 \\ 1 \end{bmatrix} \right\}$$

span the solution set. These vectors give the dimensionless combinations

$$\pi_1 = rv^{-2}g = \frac{rg}{v^2}, \quad \pi_2 = \rho^{-1}\rho_\ell = \frac{\rho_\ell}{\rho}, \quad \text{and} \quad \pi_3 = r^{-1}\rho^{-1}v^{-1}\mu = \frac{\mu}{r\rho v}.$$

Observe that the empirical law can be written as

$$v = \frac{2r^2\rho g}{9\mu}\left(1 - \frac{\rho_\ell}{\rho}\right) \implies 9\mu v = 2r^2\rho g\,(1 - \pi_2) \implies \frac{\mu}{r\rho v} = \frac{2rg}{9v^2}\,(1 - \pi_2) \implies \pi_3 = \frac{2}{9}\pi_1\,(1 - \pi_2)\,.$$

The final expression involves only dimensionless quantities, so it follows that the empirical law can be deduced from dimensionless analysis.

DEFINITION 21.3. A law $f(q_1, q_2, \ldots, q_m) = 0$ will be called *unit-free* if for all choices of $\bar{L}_i = \lambda_i L_i$, the law is equivalent to $f(\bar{q}_1, \bar{q}_2, \ldots, \bar{q}_m) = 0$.

EXAMPLE 21.4 (Unit-free assumption). Show that the physical law $f(x, t, g) = x - gt^2/2 = 0$ is unit-free.

The physical dimensions in this problem are length L and time T. We can express the variables in terms of these units as

$$[x] = L, \quad [t] = T, \quad \text{and} \quad [g] = \frac{L}{T^2}.$$

Let $\bar{L} = \lambda_1 L$, and let $\bar{T} = \lambda_2 T$, so that

$$\bar{x} = \lambda_1 x, \quad \bar{t} = \lambda_2 t, \quad \text{and} \quad g = \frac{\lambda_1}{\lambda_2^2} g.$$

Then,

$$f(\bar{x}, \bar{t}, \bar{g}) = \bar{x} - \frac{1}{2}\bar{g}\bar{t}^2 = \lambda_1 x - \frac{1}{2}\left(\frac{\lambda_1}{\lambda_2^2} g \cdot \lambda_2^2 t^2\right) = \lambda_1 x - \frac{\lambda_1}{2} g t^2 = \lambda_1 \left(x - \frac{1}{2} g t^2\right) = 0,$$

i.e., this relation is unit-free.