LECTURE NOTES
STATISTICS 424
MATHEMATICAL STATISTICS
SPRING 2003

Robert J. Boik
Department of Mathematical Sciences
Montana State University — Bozeman

Revised August 30, 2004

Contents

Chapter 0

COURSE INFORMATION & SYLLABUS

0.1 Course Information

- Prerequisite: Stat 420
- Required Texts
  - Berry, D. A. & Lindgren, B. W. (1996). *Statistics: Theory and Methods*, 2nd edition. Belmont CA: Wadsworth Publishing Co.
  - Lecture Notes
- Instructor
  - Robert J. Boik, 2–260 Wilson, 994-5339, rjboik@math.montana.edu.
  - Office Hours: Monday & Wednesday 11:00–11:50 & 2:10–3:00; Friday 11:00–11:50.
- Holidays: Monday Jan 20, MLK Day; Monday Feb 17 (Presidents Day); March 10–14 (Spring Break); Apr 18 (University Day)
- Drop dates: Wednesday Feb 5 is the last day to drop without a W grade; Friday April 25 is the last day to drop.
- Grading: 600 Points Total; Grade cutoffs (percentages) — 90, 80, 70, 60; All exams are closed book. Tables and equations will be provided.
  - Homework: 200 points
  - Exam-1, Wed Feb 19, 6:10–?: 100 points, Wilson 1-139
  - Exam-2, Monday March 31, 6:10-?: 100 points, Wilson 1-141
  - Comprehensive Final, Monday, May 5, 8:00–9:50 AM: 200 points

- Homepage: http://www.math.montana.edu/∼rjboik/classes/424/stat.424.html
  Homework assignments and revised lecture notes will be posted on the Stat 424 home page.

0.2 Syllabus

1. Continuous Random Variables: Remainder of Chapter 5
2. Families of Continuous Distributions: Chapter 6
3. Data: Chapter 7
4. Samples, Statistics, and Sampling Distributions: Chapter 8
5. Estimation: Chapter 9
6. Significance Testing: Chapter 10
7. Tests as Decision Rules: Chapter 11

0.3 Study Suggestions

1. How to do poorly in Stat 424.
   (a) Come to class.
   (b) Do home work.
   (c) When home work is returned, pay most attention to your score.
   (d) Skim the text to see if it matches the lecture.
   (e) Read notes to prepare for exams.

   Except for item (c), there is nothing in the above list that hurts performance in 424. On the contrary, if you do not come to class etc, then you will likely do worse. The problem with the above list is that the suggestions are not active enough. The way to learn probability and mathematical statistics is to do probability and mathematical statistics. Watching me do a problem or a proof helps, but it is not enough. You must do the problem or proof yourself.

2. How to do well in Stat 424.
   (a) In class
      - Make a note of all new terms.
      - Make a note of new results or theorems.
      - Make a note of which results or theorems were proven.

- For each proof, make a note of the main steps, especially of how the proof begins.

(b) Re-write notes at home

- Organize ideas and carefuly define all new terms. Use the text and the bound lecture notes for help.
- Write-up each proof, filling in details. Use the text and the bound lecture notes for help.
- Review weak areas. For example, a proof may use certain mathematical tools and/or background material in probability and statistics. If your knowledge of the tools and/or background is weak, then review the material. Summarize your review in your notes. Use the text and the bound lecture notes for help.

(c) Prepare for exams.

- Practice—re-work homework problems from scratch. Be careful to not make the same mistakes as were made on the original home work. Learn from the mistakes on the graded home work.
- Practice—take the practice exam with notes, text, and bound lecture notes as aids.
- Practice—re-take the practice exam without notes, text, and bound lecture notes as aids.
- Practice—re-work proofs without notes, text, and bound lecture notes as aids.

0.4 Types of Proofs

1. Proof by Construction: To prove the claim "If A then B." start by assuming that A is true. Ask yourself what are the sufficient conditions for B to be true. Show that if A is true, then one or more of the sufficient conditions for B are true. A proof by construction essentially constructs B from A.

2. Proof by Contradiction: To prove the claim "If A then B." start by assuming that A is true and that B is false. Work forward from both of these assumptions and show that they imply a statement that obviously is false. For example, if A is true and B is false, then $\text{Var}(X) < 0$. This false statement contradicts the possibility that A can be true and B can be false. Therefore, if A is true, B also must be true.

3. Proof by Contrapositive: This is similar to the proof by contradiction. To prove the claim "If A then B." start by assuming that A is true and that B is false. Work forward only from the assumption that B is false and show that it implies that A also must be false. This contradicts the assumptions that A is true and B is false. Therefore, if A is true, B also must be true.

4. Proof by Induction: To prove the claim that "If A then $B_1, B_2, B_3, \ldots, B_\infty$ all are true," start by proving the claim "If A then B_1. Any of the above types of proof may be used to prove this claim. Then prove the claim "If A and B_k then B_{k+1}. The proofs of the two claims imply, by induction, that A implies that $B_1, B_2, \ldots, B_\infty$ all are true.

Chapter 5

CONTINUOUS RANDOM VARIABLES

5.1 Cumulative Distribution Function (CDF)

1. Definition: Let X be a random variable. Then the cdf of X is denoted by $F_X(x)$ and defined by
$$F_X(x) = P(X \leq x).$$
If X is the only random variable under consideration, then $F_X(x)$ can be written as $F(x)$.

2. Example: Discrete Distribution. Suppose that $X \sim \text{Bin}(3, 0.5)$. Then $F(x)$ is a step function and can be written as
$$F(x) = \begin{cases} 0 & x \in (-\infty, 0); \\ \frac{1}{8} & x \in [0, 1); \\ \frac{1}{2} & x \in [1, 2); \\ \frac{7}{8} & x \in [2, 3); \\ 1 & x \in [3, \infty). \end{cases}$$

3. Example: Continuous Distribution. Consider modeling the probability of vehicle accidents on I-94 in the Gallatin Valley by a Poisson process with rate λ per year. Let T be the time until the first accident. Then
$$\begin{aligned} P(T \leq t) &= P(\text{at least one accident in time } t) \\ &= 1 - P(\text{no accidents in time } t) = 1 - \frac{e^{-\lambda}\lambda^0}{0!} = 1 - e^{-\lambda t}. \end{aligned}$$
Therefore,
$$F(t) = \begin{cases} 0 & t < 0; \\ 1 - e^{-\lambda t} & t \geq 0. \end{cases}$$

4. Example: Uniform Distribution. Suppose that X is a random variable with support $\mathcal{S} = [a, b]$, where $b > a$. Further, suppose that the probability that X falls in an interval in $\mathcal{S}$ is proportional to the length of the interval. That is, $P(x_1 \leq X \leq x_2) = \lambda(x_2 - x_1)$ for $a \leq x_1 \leq x_2 \leq b$. To solve for λ, let $x_1 = a$ and $x_2 = b$. then

$$P(a \leq X \leq b) = 1 = \lambda(b - a) \Longrightarrow \lambda = \frac{1}{b - a}.$$

Accordingly, the cdf is

$$F(x) = P(X \leq x) = P(a \leq X \leq x) = \begin{cases} 0 & x < a; \\ \dfrac{x - a}{b - a} & x \in [a, b]; \\ 1 & x > b. \end{cases}$$

In this case, X is said to have a uniform distribution: $X \sim \text{Unif}(a, b)$.

5. Properties of a cdf

   (a) $F(-\infty) = 0$ and $F(\infty) = 1$. Your text tries (without success) to motivate this result by using equation 1 on page 157. Ignore the discussion on the bottom of page 160 and the top of page 161.

   (b) F is non-decreasing; i.e., $F(a) \leq F(b)$ whenever $a \leq b$.

   (c) $F(x)$ is right continuous. That is, $\lim_{\epsilon \to 0^+} F(x + \epsilon) = F(x)$.

6. Let X be a rv with cdf $F(x)$.

   (a) If $b \geq a$, then $P(a < X \leq b) = F(b) - F(a)$.

   (b) For any x, $P(X = x) = \lim\limits_{\epsilon \to 0^+} P(x - \epsilon < X \leq x) = F(x) - F(x-)$, where $F(x-)$ is F evaluated as $x - \epsilon$ and ϵ is an infinitesimally small positive number. If the cdf of X is continuous from the left, then $F(x-) = F(x)$ and $P(X = x) = 0$. If the cdf of X has a jump at x, then $F(x) - F(x-)$ is the size of the jump.

   (c) Example: Problem 5-8.

7. Definition of Continuous Distribution: The distribution of the rv X is said to be continuous if the cdf is continuous at each x and the cdf is differentiable (except, possibly, at a countable number of points).

8. Monotonic transformations of a continuous rv: Let X be a continuous rv with cdf $F_X(x)$.

   (a) Suppose that $g(X)$ is a continuous one-to-one increasing function. Then for y in the counter-domain (range) of g, the inverse function $x = g^{-1}(y)$ exists. Let $Y = g(X)$. Find the cdf of Y. Solution:

$$P(Y \leq y) = P[g(X) \leq y] = P(X \leq g^{-1}(y)] = F_X[g^{-1}(y)].$$

(b) Suppose that $g(X)$ is a continuous one-to-one decreasing function. Then for y in the counter-domain (range) of g, the inverse function $x = g^{-1}(y)$ exists. Let $Y = g(X)$. Find the cdf of Y. Solution:

$$P(Y \leq y) = P[g(X) \leq y] = P(X > g^{-1}(y)] = 1 - F_X[g^{-1}(y)].$$

(c) Example: Suppose that $X \sim \text{Unif}(0,1)$, and $Y = g(X) = hX + k$ where $h < 0$. Then, $X = g^{-1}(Y) = (Y - k)/h$;

$$F_X(x) = \begin{cases} 0 & x < 0; \\ x & x \in [0,1]; \\ 1 & x > 1, \end{cases}$$

and

$$F_Y(y) = 1 - F_x[(y-k)/h] = \begin{cases} 0 & y < h + k; \\ \dfrac{y - (h+k)}{-k} & y \in [h+k, k]; \\ 1 & y > k, \end{cases}$$

That is, $Y \sim \text{Unif}(h + k, k)$.

(d) Inverse CDF Transformation.

i. Suppose that X is a continuous rv having a strictly increasing cdf $F_X(x)$. Recall that a strictly monotone function has an inverse. Denote the inverse of the cdf by F_X^{-1}. That is, if $F_X(x) = y$, then $F_X^{-1}(y) = x$. Let $Y = F_X(X)$. Then the distribution of Y is $\text{Unif}(0,1)$.

Proof: If $W \sim \text{Unif}(0,1)$, then the cdf of W is $F_W(w) = w$. The cdf of Y is

$$\begin{aligned} F_Y(y) &= P(Y \leq y) = P(F_X(X) \leq y) = P[X \leq F_X^{-1}(y)] \\ &= F_X\left[F_X^{-1}(y)\right] = y. \end{aligned}$$

If Y has support $[0,1]$ and $F_Y(y) = y$, then it must be true that $Y \sim \text{Unif}(0,1)$.

ii. Let U be a rv with distribution $U \sim \text{Unif}(0,1)$. Suppose that $F_X(x)$ is a strictly increasing cdf for a continuous random variable X. Then the cdf of the rv $F_X^{-1}(U)$ is $F_X(x)$.

Proof:

$$P\left[F_X^{-1}(U) \leq x\right] = P\left[U \leq F_X(x)\right] = F_U\left[F_X(x)\right] = F_X(x)$$

because $F_U(u) = u$.

(e) Application of inverse cdf transformation: Given $U_1, U_2, \ldots, U_n$, a random sample from Unif$(0,1)$, generate a random sample from $F_X(x)$. Solution: Let $X_i = F_X^{-1}(U_i)$ for $i = 1, \ldots, n$.

i. Example 1: Suppose that $F_X(x) = 1 - e^{-\lambda x}$ for $x > 0$, where $\lambda > 0$. Then $X_i = -\ln(1 - U_i)/\lambda$ for $i = 1, \ldots, n$ is a random sample from F_X.

ii. Example 2: Suppose that

$$F_X(x) = \left[1 - \left(\frac{a}{x}\right)^b\right] I_{(a,\infty)}(x),$$

where $a > 0$ and $b > 0$ are constants. Then $X_i = a(1 - U_i)^{-b}$ for $i = 1, \ldots, n$ is a random sample from F_X.

9. Non-monotonic transformations of a continuous rv. Let X be a continuous rv with cdf $F_X(x)$. Suppose that $Y = g(X)$ is a continuous but non-monotonic function. As in the case of monotonic functions, $F_Y(y) = P(Y \le y) = P[g(X) \le y]$, but in this case each inverse solution $x = g^{-1}(y)$ must be used to find an expression for $F_Y(y)$ in terms of $F_X[g^{-1}(y)]$. For example, suppose that $X \sim$ Unif$(-1, 2)$ and $g(X) = Y = X^2$. Note that $x = \pm\sqrt{y}$ for $y \in [0, 1]$ and $x = +\sqrt{y}$ for $y \in (1, 4]$. The cdf of Y is

$$\begin{aligned} F_Y(y) &= P(X^2 \le y) = \begin{cases} P(-\sqrt{y} \le X \le \sqrt{y}) & y \in [0,1]; \\ P(X \le \sqrt{y}) & y \in (1,4] \end{cases} \\ &= \begin{cases} F_X(\sqrt{y}) - F_X(-\sqrt{y}) & y \in [0,1]; \\ F_X(\sqrt{y}) & y \in (1,4] \end{cases} \\ &= \begin{cases} 0 & y < 0; \\ 2\sqrt{y}/3 & y \in [0,1]; \\ (\sqrt{y}+1)/3 & y \in (1,4]; \\ 1 & y > 4; \end{cases} \end{aligned}$$

Plot the function $g(x)$ over $x \in \mathcal{S}_X$ as an aid to finding the inverse solutions $x = g^{-1}(y)$.

5.2 Density and the Probability Element

1. Mathematical Result: Assume that $F(x)$ is a continuous cdf. Let $g(x + m)$ be a differentiable function and let $y = x + m$. Then

$$\frac{d}{dm} g(x+m)\Big|_{m=0} = \frac{d}{d\,y} g(y) \times \frac{d}{d\,m} y\Big|_{m=0} = \frac{d}{d\,y} g(y)\Big|_{m=0} = \frac{d}{dx} g(x)$$

by the chain rule.

2. Probability Element: Suppose that X is a continuous rv. Let Δx be a small positive number. Define $h(a,b)$ as

$$h(a,b) \stackrel{\text{def}}{=} P(a \le X \le a+b) = F_X(a+b) - F_X(a).$$

Expand $h(x,\Delta x) = P(x \le X \le x+\Delta x)$ in a Taylor series around $\Delta x = 0$:

$$\begin{aligned} h(x,\Delta x) &= F(x+\Delta x) - F(x) \\ &= h(x,0) + \frac{d}{d\Delta x} h(x,\Delta x)\Big|_{\Delta x=0} \Delta x + o(\Delta x) \\ &= 0 + \frac{d}{d\Delta x} F(x+\Delta x)\Big|_{\Delta x=0} \Delta x + o(\Delta x) \\ &= \left[\frac{d}{dx}F(x)\right]\Delta x + o(\Delta x), \text{ where} \\ &\quad \lim_{\Delta x \to 0} \frac{o(\Delta x)}{\Delta x} = 0. \end{aligned}$$

The function

$$d\,F(x) = \left[\frac{d}{dx}F(x)\right]\Delta x$$

is called the differential. In the field of statistics, the differential of a cdf is called the probability element. The probability element is an approximation to $h(x,\Delta x)$. Note that the probability element is a linear function of the derivative $\frac{d}{dx}F(x)$.

3. Example; Suppose that

$$F(x) = \begin{cases} 0 & x < 0; \\ 1 - e^{-3x} & \text{otherwise.} \end{cases}$$

Note that $F(x)$ is a cdf. Find the probability element at $x = 2$ and approximate the probability $P(2 \le X \le 2.01)$. Solution: $\frac{d}{dx}F(x) = 3e^{-3x}$ so the probability element is $3e^{-6}\Delta x$ and
$P(2 \le X \le 2.01) \approx 3e^{-6} \times 0.01 = 0.00007436$. The exact probability is $F(2.01) - F(2) = 0.00007326$.

4. The average density in the interval $(x, x+\Delta x)$ is defined as

$$\text{Average density} \stackrel{\text{def}}{=} \frac{P(x < X < x+\Delta x)}{\Delta x}.$$

5. Density: The probability density function (pdf) at x is the limit of the average density as $\Delta x \to 0$:

$$\text{pdf} \quad = \quad f(x) \stackrel{\text{def}}{=} \lim_{\Delta x \to 0} \frac{P(x \le X \le x+\Delta x)}{\Delta x}$$

$$\begin{aligned} &= \lim_{\Delta x \to 0} \frac{F_X(x+\Delta x) - F_X(x)}{\Delta x} \\ &= \frac{\left[\frac{d}{dx}F(x)\right]\Delta x + o(\Delta x)}{\Delta x} \\ &= \frac{d}{dx}F(x). \end{aligned}$$

Note that the probability element can be written as $d\,F(x) = f(x)\Delta x$.

6. Example: Suppose that λ is a positive real number. If

$$F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0; \\ 0 & \text{otherwise.} \end{cases} \text{ then } f(x) = \frac{d}{dx}F(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0; \\ 0 & \text{otherwise.} \end{cases}$$

7. Example: If $X \sim \text{Unif}(a, b)$, then

$$F(x) = \begin{cases} 0 & x < a; \\ \frac{x-a}{b-a} & x \in [a, b]; \\ 1 & x > b \end{cases} \text{ and } f(x) = \frac{d}{dx}F(x) = \begin{cases} 0 & x < a; \\ \frac{1}{b-a} & x \in [a, b]; \\ 0 & x > b. \end{cases}$$

8. Properties of a pdf

   i $f(x) \geq 0$ for all x.

   ii $\displaystyle\int_{-\infty}^{\infty} f(x) = 1$

9. Relationship between pdf and cdf: If X is a continuous rv with pdf $f(x)$ and cdf $F(x)$, then

$$f(x) = \frac{d}{dx}F(x)$$
$$F(x) = \int_{-\infty}^{x} f(u)du \text{ and}$$
$$P(a < X < b) = P(a \leq X \leq b) = P(a < X \leq b)$$
$$= P(a \leq X < b) = F(b) - F(a) = \int_a^b f(x)dx.$$

10. PDF example - Cauchy distribution. Let $f(x) = c/(1 + x^2)$ for $-\infty < x < \infty$ and where c is a constant. Note that $f(x)$ is nonnegative and

$$\int_{-\infty}^{\infty} \frac{1}{1+x^2}dx = \arctan(x)\Big|_{-\infty}^{\infty} = \frac{\pi}{2} - \frac{-\pi}{2} = \pi.$$

Accordingly, if we let $c = 1/\pi$, then

$$f(x) = \frac{1}{\pi(1+x^2)}$$

is a pdf. It is called the Cauchy pdf. The corresponding cdf is

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} \frac{1}{\pi(1+u^2)} du = \frac{\arctan(u)}{\pi}\Big|_{-\infty}^{x} \\ &= \frac{1}{\pi}\left[\arctan(x) + \frac{\pi}{2}\right] = \frac{\arctan(x)}{\pi} + \frac{1}{2}. \end{aligned}$$

11. PDF example - Gamma distribution: A more general waiting time distribution: Let T be the time of arrival of the r^{th} event in a Poisson process with rate parameter λ. Find the pdf of T. Solution: $T \in (t, t+\Delta t)$ if and only if (a) $r-1$ events occur before time t and (b) one event occurs in the interval $(t, t+\Delta t)$. The probability that two or more events occur in $(t, t+\Delta t)$ is $o(\Delta t)$ and can be ignored. By the Poisson assumptions, outcomes (a) and (b) are independent and the probability of outcome (b) is $\lambda\Delta t + o(\Delta t)$. Accordingly,

$$\begin{aligned} P(t < T < t+\Delta t) \approx f(t)\Delta t &= \frac{e^{-\lambda t}(\lambda t)^{r-1}}{(r-1)!} \times \lambda \Delta t \\ &= \left[\frac{e^{-\lambda t}\lambda^r t^{r-1}}{(r-1)!}\right] \Delta t \end{aligned}$$

and the pdf is

$$f(t) = \begin{cases} 0 & t < 0; \\ \dfrac{e^{-\lambda t}\lambda^r t^{r-1}}{(r-1)!} & t \geq 0 \end{cases} \quad = \quad \frac{e^{-\lambda t}\lambda^r t^{r-1}}{\Gamma(r)} I_{[0,\infty)}(t).$$

12. Transformations with Single-Valued Inverses: If X is a continuous random variable with pdf $f_X(x)$ and $Y = g(X)$ is a single-valued differentiable function of X, then the pdf of Y is

$$f_Y(y) = f_X\left[g^{-1}(y)\right] \left|\frac{d}{d\,y} g^{-1}(y)\right|$$

for $y \in \mathcal{S}_{g(x)}$ (i.e., support of $Y = g(X)$). The term

$$J(y) = \frac{d}{d\,y} g^{-1}(y)$$

is called the Jacobian of the transformation.

(a) Justification 1: Suppose that $Y = g(X)$ is strictly increasing. Then $F_Y(y) = F_X\left[g^{-1}(y)\right]$ and

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}F_Y(y) = f_X\left[g^{-1}(y)\right]\frac{d}{dy}g^{-1}(y) \\ &= f_X\left[g^{-1}(y)\right]\left|\frac{d}{dy}g^{-1}(y)\right| \end{aligned}$$

because the Jacobian is positive. Suppose that $Y = g(X)$ is strictly decreasing. Then $F_Y(y) = 1 - F_X\left[g^{-1}(y)\right]$ and

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\left[1 - F_Y(y)\right] = -f_X\left[g^{-1}(y)\right]\frac{d}{dy}g^{-1}(y) \\ &= f_X\left[g^{-1}(y)\right]\left|\frac{d}{dy}g^{-1}(y)\right| \end{aligned}$$

because the Jacobian is negative.

(b) Justification 2: Suppose that $g(x)$ is strictly increasing. Recall that

$$P(x \le X \le x + \Delta x) = f_X(x)\Delta x + o(\Delta x).$$

Note that

$$x \le X \le x + \Delta x \Longleftrightarrow g(x) \le g(X) \le g(x + \Delta x).$$

Accordingly,

$$\begin{aligned} P(x \le X \le x + \Delta x) = P(y \le Y \le y + \Delta y) &= f_Y(y)\Delta y + o(\Delta y) \\ &= f_X(x)\Delta x + o(\Delta x) \end{aligned}$$

where $y + \Delta y = g(x + \Delta x)$. Expanding $g(x + \Delta x)$ around $\Delta x = 0$ reveals that

$$y + \Delta y = g(x + \Delta x) = g(x) + \frac{d\,g(x)}{d\,x}\Delta x + o(\Delta x).$$

Also,

$$\begin{aligned} y = g(x) &\Longrightarrow g^{-1}(y) = x \\ &\Longrightarrow \frac{d\,g^{-1}(y)}{d\,y} = \frac{d\,x}{d\,y} \\ &\Longrightarrow \frac{d\,y}{d\,x} = \frac{d\,g(x)}{d\,x} = \left[\frac{d\,g^{-1}(y)}{d\,y}\right]^{-1} \\ &\Longrightarrow y + \Delta y = g(x) + \left[\frac{d\,g^{-1}(y)}{d\,y}\right]^{-1}\Delta x \\ &\Longrightarrow \Delta y = \left[\frac{d\,g^{-1}(y)}{d\,y}\right]^{-1}\Delta x \end{aligned}$$

$$\Longrightarrow \Delta x = \frac{d\, g^{-1}(y)}{d\, y}\Delta y.$$

Lastly, equating $f_X(x)\Delta x$ to $f_Y(y)\Delta y$ reveals that

$$\begin{aligned} f_Y(y)\Delta y = f_X(x)\Delta x &= f_X\left[g^{-1}(y)\right]\Delta x \\ &= f_X\left[g^{-1}(y)\right]\frac{d\, g^{-1}(y)}{d\, y}\Delta y \\ \Longrightarrow f_Y(y) &= f_X\left[g^{-1}(y)\right]\frac{d\, g^{-1}(y)}{d\, y}. \end{aligned}$$

The Jacobian $\frac{dg^{-1}(y)}{dy}$ is positive for an increasing function, so the absolute value operation is not necessary. A similar argument can be made for the case when $g(x)$ is strictly decreasing.

13. Transformations with Multiple-Valued Inverses: If $g(x)$ has more than one inverse function, then a separate probability element must be calculated for each of the inverses. For example, suppose that $X \sim \text{Unif}(-1, 2)$ and $Y = g(X) = X^2$. There are two inverse functions for $y \in [0, 1]$, namely $x = -\sqrt{y}$ and $x = +\sqrt{y}$. There is a single inverse function for $y \in (1, 4]$. The pdf of Y is found as

$$\begin{aligned} f(y) &= \begin{cases} 0 & y < 0; \\ f(-\sqrt{y})\left|\frac{-d\sqrt{y}}{dy}\right| + f(\sqrt{y})\left|\frac{d\sqrt{y}}{dy}\right| & y \in [0, 1]; \\ f(\sqrt{y})\left|\frac{d\sqrt{y}}{dy}\right| & y \in (1, 4]; \\ 0 & y > 4. \end{cases} \\ &= \begin{cases} 0 & y < 0; \\ \frac{1}{3\sqrt{y}} & y \in [0, 1]; \\ \frac{1}{6\sqrt{y}} & y \in (1, 4]; \\ 0 & y > 4. \end{cases} \end{aligned}$$

5.3 The Median and Other Percentiles

1. Definition: The number x_p is said to be the $100p^{\text{th}}$ percentile of the distribution of X if x_p satisfies

$$F_X(x_p) = P(X \leq x_p) = p.$$

2. If the cdf $F_X(x)$ is strictly increasing, then $x_p = F_X^{-1}(p)$ and x_p is unique.

3. If $F_X(x)$ is not strictly increasing, then x_p may not be unique.

4. Median: The median is the 50^{th} percentile (i.e., $p = 0.5$).

5. Quartiles: The first and third quartiles are $x_{0.25}$ and $x_{0.75}$ respectively.

6. Example: If $X \sim \text{Unif}(a, b)$, then $(x_p - a)/(b - a) = p$; $x_p = a + p(b - a)$; and $x_{0.5} = (a + b)/2$.

7. Example: If $F_X(x) = 1 - e^{-\lambda x}$ (i.e., waiting time distribution), then $1 - e^{-\lambda x_p} = p$; $x_p = -\ln(1 - p)/\lambda$; and $x_{0.5} = \ln(2)/\lambda$.

8. Example—Cauchy: Suppose that X is a random variable with pdf

$$f(x) = \frac{1}{\sigma\pi\left[1 + \dfrac{(x-\mu)^2}{\sigma^2}\right]},$$

where $-\infty < x < \infty$; $\sigma > 0$; and μ is a finite number. Then

$$\begin{aligned} F(x) &= \int_{-\infty}^{x} f(u)du = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\pi(1+z^2)}dz \\ &\qquad \left(\text{make the change of variable from } x \text{ to } z = \frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\pi}\arctan\left(\frac{x-\mu}{\sigma}\right) + \frac{1}{2}. \end{aligned}$$

Accordingly,

$$\begin{aligned} & F(x_p) = p \Longrightarrow \\ x_p &= \mu + \sigma\tan\left[\pi(p - 0.5)\right]; \\ x_{0.25} &= \mu + \sigma\tan\left[\pi(0.25 - 0.5)\right] = \mu - \sigma; \\ x_{0.5} &= \mu + \sigma\tan(0) = \mu; \text{ and} \\ x_{0.75} &= \mu + \sigma\tan\left[\pi(0.75 - 0.5)\right] = \mu + \sigma. \end{aligned}$$

9. Definition Symmetric Distribution: A distribution is said to be symmetric around c if $F_X(c - \delta) = 1 - F_X(c + \delta)$ for all δ.

10. Definition Symmetric Distribution: A distribution is said to be symmetric around c if $f_X(c - \delta) = f_X(c + \delta)$ for all δ.

11. Median of a symmetric distribution. Suppose that the distribution of X is symmetric around c. Then, set δ to $c - x_{0.5}$ to obtain

$$F_X(x_{0.5}) = \frac{1}{2} = 1 - F_X(2c - x_{0.5}) \Longrightarrow F_X(2c - x_{0.5}) = \frac{1}{2} \Longrightarrow c = x_{0.5}.$$

That is, if the distribution of X is symmetric around c, then the median of the distribution is c.

5.4 Expected Value

1. Definition: Let X be a rv with pdf $f(x)$. Then the expected value (or mean) of X, if it exists, is
$$\mathrm{E}(X) = \mu_X = \int_{-\infty}^{\infty} x f(x) dx.$$

2. The expectation is said to exist if the integral of the positive part of the function is finite and the integral of the negative part of the function is finite.

5.5 Expected Value of a Function

1. Let X be a rv with pdf $f(x)$. Then the expected value of $g(X)$, if it exists, is
$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

2. Linear Functions. The integral operator is linear. If $g_1(X)$ and $g_2(X)$ are functions whose expectation exists and a, b, c are constants, then
$$\mathrm{E}\left[a g_1(X) + b g_2(X) + c\right] = a\mathrm{E}\left[g_1(X)\right] + b\mathrm{E}\left[g_2(X)\right] + c.$$

3. Symmetric Distributions: If the distribution of X is symmetric around c and the expectation exists, then $\mathrm{E}(X) = c$.

   Proof: Assume that the mean exists. First, show that $\mathrm{E}(X - c) = 0$:
$$\begin{aligned}
&\mathrm{E}(X-c) = \int_{-\infty}^{\infty} (x-c) f(x) dx \\
&= \int_{-\infty}^{c} (x-c) f(x) dx + \int_{c}^{\infty} (x-c) f(x) dx \\
&(\text{ let } x = c-u \text{ in integral 1 and let } x = c+u \text{ in integral 2}) \\
&= -\int_{0}^{\infty} u f(c-u) du + \int_{0}^{\infty} u f(c+u) du \\
&= \int_{0}^{\infty} u \left[f(c+u) - f(c-u)\right] du = 0
\end{aligned}$$
   by symmetry of the pdf around c. Now use $\mathrm{E}(X-c) = 0 \iff \mathrm{E}(X) = c$.

4. Example: Suppose that $X \sim \mathrm{Unif}(a, b)$. That is,
$$f(x) = \begin{cases} \dfrac{1}{b-a} & x \in [a, b]; \\ 0 & \text{otherwise.} \end{cases}$$

A sketch of the pdf shows that the distribution is symmetric around $(a+b)/2$. More formally,

$$f\left(\frac{a+b}{2}-\delta\right) = f\left(\frac{a+b}{2}+\delta\right) = \begin{cases} \dfrac{1}{b-a} & \delta \in \left[-\frac{b-a}{2}, \frac{b-a}{2}\right]; \\ 0 & \text{otherwise.} \end{cases}$$

Accordingly, $\mathrm{E}(X) = (a+b)/2$. Alternatively, the expectation can be found by integrating $xf(x)$:

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{\infty} xf(x)\,dx = \int_a^b \frac{x}{b-a}dx \\ &= \frac{x^2}{2(b-a)}\bigg|_a^b = \frac{b^2-a^2}{2(b-a)} \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{a+b}{2}. \end{aligned}$$

5. Example: Suppose that X has a Cauchy distribution. The pdf is

$$f(x) = \frac{1}{\sigma\pi\left[1+\dfrac{(x-\mu)^2}{\sigma^2}\right]},$$

where μ and σ are constants that satisfy $|\mu| < \infty$ and $\sigma \in (0,\infty)$. By inspection, it is apparent that the pdf is symmetric around μ. Nonetheless, the expectation is not μ, because the expectation does not exist. That is,

$$\begin{aligned} &\int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} \frac{x}{\sigma\pi\left[1+\dfrac{(x-\mu)^2}{\sigma^2}\right]}dx \\ &= \mu + \sigma\int_{-\infty}^{\infty} \frac{z}{\pi(1+z^2)}dz \text{ where } z = \frac{x-\mu}{\sigma} \\ &= \mu + \sigma\int_{-\infty}^{0} \frac{z}{\pi(1+z^2)}dz + \sigma\int_0^{\infty} \frac{z}{\pi(1+z^2)}dz \\ &= \mu + \sigma\frac{\ln(1+z^2)}{2\pi}\bigg|_{-\infty}^{0} + \sigma\frac{\ln(1+z^2)}{2\pi}\bigg|_0^{\infty} \end{aligned}$$

and neither the positive nor the negative part is finite.

6. Example: Waiting time distribution. Suppose that X is a rv with pdf $\lambda e^{-\lambda x}$ for $x > 0$ and where $\lambda > 0$. Then, using integration by parts,

$$\begin{aligned} \mathrm{E}(X) &= \int_0^{\infty} x\lambda e^{-\lambda x}dx = -xe^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} e^{-\lambda x}dx \\ &= 0 - \frac{1}{\lambda}e^{-\lambda x}\Big|_0^{\infty} = \frac{1}{\lambda}. \end{aligned}$$

5.6 Average Deviations

1. Variance

   (a) Definition:

   $$\mathrm{Var}(X) \stackrel{\mathrm{def}}{=} \mathrm{E}(X-\mu_X)^2 = \int_{-\infty}^{\infty}(x-\mu_X)^2 f(x)dx$$

   if the expectation exists. It is conventional to denote the variance of X by σ_X^2.

   (b) Computational formula: Be able to verify that $\mathrm{Var}(X) = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2$.

   (c) Example: Suppose that $X \sim \mathrm{Unif}(a,b)$. Then

   $$\mathrm{E}(X^r) = \int_a^b \frac{x^r}{b-a}dx = \frac{x^{r+1}}{(r+1)(b-a)}\bigg|_a^b = \frac{b^{r+1}-a^{r+1}}{(r+1)(b-a)}.$$

   Accordingly, $\mu_X = (a+b)/2$,

   $$\begin{aligned}\mathrm{E}(X^2) &= \frac{b^3-a^3}{3(b-a)} = \frac{(b-a)(b^2+ab+a^2)}{3(b-a)} = \frac{b^2+ab+a^2}{3} \text{ and}\\ \mathrm{Var}(X) &= \frac{b^2+ab+a^2}{3} - \frac{(b+a)^2}{4} = \frac{b^2-2ab+a^2}{12} = \frac{(b-a)^2}{12}.\end{aligned}$$

   (d) Example: Suppose that $f(x) = \lambda e^{-\lambda x}$ for $x > 0$ and where $\lambda > 0$. Then $\mathrm{E}(X) = 1/\lambda$,

   $$\begin{aligned}\mathrm{E}(X^2) &= \int_0^{\infty} x^2\lambda e^{-\lambda x}dx\\ &= -x^2e^{-\lambda x}\Big|_0^{\infty} + \int_0^{\infty} 2xe^{-\lambda x}dx = 0 + \frac{2}{\lambda^2} \text{ and}\\ \mathrm{Var}(X) &= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}.\end{aligned}$$

2. MAD

   (a) Definition:

   $$\mathrm{Mad}(X) \stackrel{\mathrm{def}}{=} \mathrm{E}(|X-\mu_X|) = \int_{-\infty}^{\infty}|x-\mu_X|f(x)dx.$$

   (b) Alternative expression: First, note that

   $$\mathrm{E}(|X-c|) = \int_{-\infty}^{c}(c-x)f(x)dx + \int_c^{\infty}(x-c)f(x)dx$$

$$= \quad c\left[2F_X(c)-1\right] - \int_{-\infty}^{c} xf(x)dx + \int_{c}^{\infty} xf(x)dx.$$

Accordingly,

$$\text{Mad}(X) = \mu_X\left[2F_X(\mu_X)-1\right] - \int_{-\infty}^{\mu_X} xf(x)dx + \int_{\mu_X}^{\infty} xf(x)dx.$$

(c) Leibnitz's Rule: Suppose that $a(\theta)$, $b(\theta)$, and $g(x,\theta)$ are differentiable functions of θ. Then

$$\begin{aligned}\frac{d}{d\theta}\int_{a(\theta)}^{b(\theta)} g(x,\theta)dx &= g\left[b(\theta),\theta\right]\frac{d}{d\theta}b(\theta) - g\left[a(\theta),\theta\right]\frac{d}{d\theta}a(\theta) \\ &+ \int_{a(\theta)}^{b(\theta)} \frac{d}{d\theta}g(x,\theta)dx.\end{aligned}$$

(d) Result: If the expectation $\text{E}(|X-c|)$ exists, then the minimizer of $\text{E}(|X-c|)$ with respect to c is $c = F_X^{-1}(0.5) =$ median of X.

Proof: Set the derivative of $\text{E}(|X-c|)$ to zero and solve for c:

$$\begin{aligned}&\frac{d}{dc}\text{E}(|X-c|) \\ &= \frac{d}{dc}\left\{c\left[2F_X(c)-1\right] - \int_{-\infty}^{c} xf_X(x)dx + \int_{c}^{\infty} xf_X(x)dx\right\} \\ &= 2F_X(c) - 1 + 2cf_X(c) - cf_X(c) - cf_X(c) \\ &= 2F_X(c) - 1.\end{aligned}$$

Equating the derivative to zero and solving for c reveals that c is a solution to $F_X(c) = 0.5$. That is, c is the median of X. Use the second derivative test to verify that the solution is a minimizer:

$$\begin{aligned}\frac{d^2}{dc^2}\text{E}(|X-c|) &= \frac{d}{dc}\left[2F_X(c)-1\right] = 2f_X(c) > 0 \\ &\Longrightarrow c \text{ is a minimizer.}\end{aligned}$$

(e) Example: Suppose that $X \sim \text{Unif}(a,b)$. Then $F_X(\frac{a+b}{2}) = 0.5$ and

$$\text{Mad}(X) = -\int_{a}^{\frac{a+b}{2}} \frac{x}{b-a}dx + \int_{\frac{a+b}{2}}^{b} \frac{x}{b-a}dx = \frac{b-a}{4}.$$

(f) Example: Suppose that $f_X(x) = \lambda e^{-\lambda x}$ for $x > 0$ and where $\lambda > 0$. Then $\text{E}(X) = 1/\lambda$, $\text{Median}(X) = \ln(2)/\lambda$, $F_X(x) = 1 - e^{-\lambda x}$, and

$$\text{Mad}(X) = \frac{1}{\lambda}\left[2 - 2e^{-1} - 1\right]$$

$$-\int_0^{\lambda^{-1}} x\lambda e^{-\lambda x}dx + \int_{\lambda^{-1}}^{\infty} x\lambda e^{-\lambda x}dx = \frac{2}{\lambda e},$$

where $\int x\lambda e^{-\lambda x}dx = -xe^{-\lambda x} - \lambda^{-1}e^{-\lambda x}$ has been used. The mean absolute deviation from the median is

$$\mathrm{E}\left|X - \frac{\ln(2)}{\lambda}\right| = -\int_0^{\ln(2)\lambda^{-1}} x\lambda e^{-\lambda x}dx + \int_{\ln(2)\lambda^{-1}}^{\infty} x\lambda e^{-\lambda x}dx$$
$$= \frac{\ln(2)}{\lambda}.$$

3. Standard Scores
   (a) Let $Z = \dfrac{X - \mu_X}{\sigma_X}$.
   (b) Moments: $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = 1$.
   (c) Interpretation: Z scores are scaled in standard deviation units.
   (d) Inverse Transformation: $X = \mu_X + \sigma_X Z$.

5.7 Bivariate Distributions

1. Definition: A function $f_{X,Y}(x, y)$ is a bivariate pdf if
   (i) $f_{X,Y}(x, y) \geq 0$ for all x, y and
   (ii) $\displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1.$

2. Bivariate CDF: If $f_{X,Y}(x, y)$ is a bivariate pdf, then

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u, v)dvdu.$$

3. Properties of a bivariate cdf:
   (i) $F_{X,Y}(x, \infty) = F_X(x)$
   (ii) $F_{X,Y}(\infty, y) = F_Y(y)$
   (iii) $F_{X,Y}(\infty, \infty) = 1$
   (iv) $F_{X,Y}(-\infty, y) = F_{X,Y}(x, -\infty) = F_{X,Y}(-\infty, -\infty) = 0$
   (v) $f_{X,Y}(x, y) = \dfrac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$

4. Joint pdfs and joint cdfs for three or more random variables are obtained as straightforward generalizations of the above definitions and conditions.

5. Probability Element: $f_{X,Y}(x,y)\Delta x\Delta y$ is the joint probability element. That is,

$$P(x \le X \le x+\Delta x, y \le Y \le y+\Delta y) = f_{X,Y}(x,y)\Delta x\Delta y + o(\Delta x\Delta y).$$

6. Example: Bivariate Uniform. If $(X,Y) \sim \text{Unif}(a,b,c,d)$, then

$$f_{X,Y}(x,y) = \begin{cases} \dfrac{1}{(b-a)(d-c)} & x \in (a,b), \quad y \in (c,d); \\ 0 & \text{otherwise.} \end{cases}$$

For this density, the probability $P(x_1 \le X \le x_2, y_1 \le Y \le y_2)$ is the volume of the rectangle. For example, if $(X,Y) \sim \text{Unif}(0,4,0,6)$, then $P(2.5 \le X \le 3.5, 1 \le Y \le 4) = (3.5-2.5)(4-1)/(4\times 6) = 3/24$. Another example is $P(X^2+Y^2 > 16) = 1 - P(X^2+Y^2 \le 16) = 1 - 4\pi/24 = 1-\pi/6$ because the area of a circle is πr^2 and therefore, the area of a circle with radius 4 is 16π and the area of the quarter circle in the support set is 4π.

7. Example: $f_{X,Y}(x,y) = \frac{6}{5}(x+y^2)$ for $x \in (0,1)$ and $y \in (0,1)$. Find $P(X+Y<1)$. Solution: First sketch the region of integration, then use calculus:

$$\begin{aligned} P(X+Y<1) &= P(X<1-Y) = \int_0^1 \int_0^{1-y} \frac{6}{5}(x+y^2)dxdy \\ &= \frac{6}{5}\int_0^1 \left(\frac{x^2}{2} + xy^2\right)\Big|_0^{1-y} dy \\ &= \frac{6}{5}\int_0^1 \frac{(1-y)^2}{2} + (1-y)y^2 dy \\ &= \frac{6}{5}\left(\frac{y}{2} - \frac{y^2}{2} + \frac{y^3}{6} + \frac{y^3}{3} - \frac{y^4}{4}\right)\Big|_0^1 = \frac{3}{10}. \end{aligned}$$

8. Example: Bivariate standard normal

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{e^{-\frac{1}{2}(x^2+y^2)}}{2\pi} = \frac{e^{-\frac{1}{2}x^2}}{\sqrt{2\pi}}\frac{e^{-\frac{1}{2}y^2}}{\sqrt{2\pi}} \\ &= f_X(x)f_Y(y). \end{aligned}$$

Using numerical integration, $P(X+Y<1) = 0.7602$. The matlab code is

```
g = inline('normpdf(y).*normcdf(1-y)','y');
Prob=quadl(g,-5,5)
```

where $\pm\infty$ has been approximated by ± 5.

9. Marginal Densities:

(a) Integrate out unwanted variables to obtain marginal densities. For example,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy; \quad f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx;$$
$$\text{and } f_{X,Y}(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{W,X,Y,Z}(w,x,y,z)dwdz.$$

(b) Example: If $f_{X,Y}(x,y) = \frac{6}{5}(x+y^2)$ for $x \in (0,1)$ and $y \in (0,1)$, then

$$f_X(x) = \frac{6}{5}\int_0^1 (x+y^2)dy = \frac{6x+2}{5} \text{ for } x \in (0,1) \text{ and}$$
$$f_Y(y) = \frac{6}{5}\int_0^1 (x+y^2)dx = \frac{6y^2+3}{5} \text{ for } y \in (0,1).$$

10. Expected Values

(a) The expected value of a function $g(X,Y)$ is

$$\mathrm{E}\left[g(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f_{X,Y}(x,y)dxdy.$$

(b) Example: If $f_{X,Y}(x,y) = \frac{6}{5}(x+y^2)$ for $x \in (0,1)$ and $y \in (0,1)$, then

$$\mathrm{E}(X) = \int_0^1\int_0^1 x\frac{6}{5}(x+y^2)dxdy = \int_0^1 \frac{3y^2+2}{5}dy = \frac{3}{5}.$$

5.8 Several Variables

1. The joint pdf of n continuous random variables, $X_1, \ldots, X_n$ is a function that satisfies

(i) $f(x_1, \ldots, x_n) \geq 0$, and

(ii) $\displaystyle\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_1 \cdots dx_n = 1.$

2. Expectations are linear regardless of the number of variables:

$$\mathrm{E}\left[\sum_{i=1}^{k} a_i g_i(X_1, X_2, \ldots, X_n)\right] = \sum_{i=1}^{k} a_i \mathrm{E}\left[g_i(X_1, X_2, \ldots, X_n)\right]$$

if the expectations exist.

3. Exchangeable Random variables

(a) Let $x_1^*, \ldots, x_n^*$ be a permutation of $x_1, \ldots, x_n$. Then, the joint density of $X_1, \ldots, X_n$ is said to be exchangeable if

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = f_{X_1,\ldots,X_n}(x_1^*, \ldots, x_n^*)$$

for all $x_1, \ldots, x_n$ and for all permutations $x_1^*, \ldots, x_n^*$.

(b) Result: If the joint density is exchangeable, then all marginal densities are identical. For example,

$$\begin{aligned} f_{X_1,X_2}(x_1, x_2) &= \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1, x_2, x_3)\, dx_3 \\ &= \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_3, x_2, x_1)\, dx_3 \text{ by exchangeability} \\ &= \int_{-\infty}^{\infty} f_{X_1,X_2,X_3}(x_1, x_2, x_3)\, dx_1 \text{ by relabeling variables} \\ &= f_{X_2,X_3}(x_2, x_3). \end{aligned}$$

(c) Result: If the joint density is exchangeable, then all bivariate marginal densities are identical, and so forth.

(d) Result: If the joint density is exchangeable, then the moments of X_i (if they exist) are identical for all i.

(e) Example Suppose that $f_{X,Y}(x, y) = 2$ for $x \geq 0$, $y \geq 0$, and $x + y \leq 1$. Then

$$\begin{aligned} f_X(x) &= \int_0^{1-x} 2dy = 2(1-x) \text{ for } x \in (0,1) \\ f_Y(y) &= \int_0^{1-y} 2dx = 2(1-y) \text{ for } y \in (0,1) \text{ and} \\ \mathrm{E}(X) &= \mathrm{E}(Y) = \frac{1}{3}. \end{aligned}$$

5.9 Covariance and Correlation

1. Review covariance and correlation results for discrete random variables (Section 3.4) because they also hold for continuous random variables. Below are lists of the most important definitions and results.

   (a) Definitions

   - $\mathrm{Cov}(X, Y) \stackrel{\text{def}}{=} \mathrm{E}\left[(X - \mu_X)(Y - \mu_Y)\right]$.
   - $\mathrm{Cov}(X, Y)$ is denoted by $\sigma_{X,Y}$.
   - $\mathrm{Var}(X) = \mathrm{Cov}(X, X)$.
   - $\mathrm{Corr}(X, Y) \stackrel{\text{def}}{=} \mathrm{Cov}(X, Y)/\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}$.
   - $\mathrm{Corr}(X, Y)$ is denoted by $\rho_{X,Y}$.

(b) Covariance and Correlation Results (be able to prove any of these).

- $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y)$.
- Cauchy-Schwartz Inequality: $[\mathrm{E}(XY)]^2 \leq \mathrm{E}(X^2)\mathrm{E}(Y^2)$.
- $\rho_{X,Y} \in [-1,1]$ To proof, use the Cauchy-Schwartz inequality.
- $\mathrm{Cov}(a+bX, c+dY) = bd\,\mathrm{Cov}(X,Y)$.
- $\mathrm{Cov}\left(\sum_i a_i X_i, \sum_i b_i Y_i\right) = \sum_i \sum_j a_i b_j\,\mathrm{Cov}(X_i, Y_j)$. For example, $\mathrm{Cov}(aW+bX, cY+dZ) =$ $ac\,\mathrm{Cov}(W,Y) + ad\,\mathrm{Cov}(W,Z) + bc\,\mathrm{Cov}(X,Y) + bd\,\mathrm{Cov}(X,Z)$.
- $\mathrm{Corr}(a+bX, c+dY) = \mathrm{sign}(bd)\,\mathrm{Corr}(X,Y)$.
- $\mathrm{Var}\left(\sum_i X_i\right) = \sum_i \sum_j \mathrm{Cov}(X_i, X_j) =$ $\sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$.
- Parallel axis theorem: $\mathrm{E}(X-c)^2 = \mathrm{Var}(X) + (\mu_X - c)^2$. Hint on proof: first add zero $X - c = (X - \mu_X) + (\mu_X - c)$, then take expectation.

2. Example (simple linear regression with correlated observations): Suppose that $Y_i = \alpha + \beta x_i + \varepsilon_i$ for $i = 1, \ldots, n$ and where $\varepsilon_1, \ldots, \varepsilon_n$ have an exchangeable distribution with $\mathrm{E}(\varepsilon_1) = 0$, $\mathrm{Var}(\varepsilon_1) = \sigma^2$ and $\mathrm{Cov}(\varepsilon_1, \varepsilon_2) = \rho\sigma^2$. The ordinary least squares estimator of β is

$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Then,

$$\mathrm{E}(\widehat{\beta}) = \beta \text{ and } \mathrm{Var}(\widehat{\beta}) = \frac{\sigma^2(1-\rho)}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Proof: First examine the numerator of $\widehat{\beta}$:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \overline{Y}) &= \sum_{i=1}^{n}(x_i - \bar{x})Y_i - \sum_{i=1}^{n}(x_i - \bar{x})\overline{Y} \\ &= \sum_{i=1}^{n}(x_i - \bar{x})Y_i - \overline{Y}\sum_{i=1}^{n}(x_i - \bar{x}) \\ &= \sum_{i=1}^{n}(x_i - \bar{x})Y_i \text{ because } = \sum_{i=1}^{n}(x_i - \bar{x}) = 0.\end{aligned}$$

In the same manner, it can be shown that

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})=\sum_{i=1}^{n}(x_i-\bar{x})x_i. \qquad (*)$$

Denote the denominator of $\widehat{\beta}$ by V_x. That is,

$$V_x=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}(x_i-\bar{x})x_i.$$

The least squares estimator can therefore be written as

$$\widehat{\beta}=\frac{1}{V_x}\sum_{i=1}^{n}(x_i-\bar{x})Y_i \text{ or as}$$

$$\widehat{\beta}=\sum_{i=1}^{n}w_iY_i, \text{ where } w_i=\frac{x_i-\bar{x}}{V_v}.$$

Note that

$$\sum_{i=1}^{n}w_i=\frac{1}{V_x}\sum_{i=1}^{n}(x_i-\bar{x})=0.$$

The expectation of $\widehat{\beta}$ is

$$\mathrm{E}(\widehat{\beta})=\sum_{i=1}^{n}w_i\mathrm{E}(Y_i)$$

$$=\sum_{i=1}^{n}w_i(\alpha+\beta x_i) \text{ because } \mathrm{E}(Y_i)=\mathrm{E}(\alpha+\beta x_i+\varepsilon_i)=\alpha+\beta x_i$$

$$=\alpha\sum_{i=1}^{n}w_i+\beta\sum_{i=1}^{n}w_ix_i$$

$$=0+\frac{\beta}{V_x}\sum_{i=1}^{n}(x_i-\bar{x})x_i \text{ because } \sum_{i=1}^{n}w_i=0$$

$$=\frac{\beta}{V_x}\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x}) \text{ by } (*)$$

$$=\frac{\beta}{V_x}V_x=\beta.$$

The variance of $\widehat{\beta}$ is

$$\mathrm{Var}(\widehat{\beta})=\mathrm{Var}\left(\sum_{i=1}^{n}w_iY_i\right)$$

$$=\sum_{i=1}^{n}w_i^2\,\mathrm{Var}(Y_i)+\sum_{i\neq j}w_iw_j\,\mathrm{Cov}(Y_i,Y_j) \quad \begin{array}{c}\text{using results on variances}\\ \text{of linear combinations}\end{array}$$

$$= \sigma^2 \sum_{i=1}^{n} w_i^2 + \rho\sigma^2 \sum_{i \neq j} w_i w_j. \qquad (**)$$

To complete the proof, first note that

$$\sum_{i=1}^{n} w_i^2 = \frac{1}{V_x^2} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{V_x^2} V_x = \frac{1}{V_x}.$$

Second, note that

$$\left(\sum_{i=1}^{n} w_i \right)^2 = 0 \text{ because } \sum_{i=1}^{n} w_i = 0 \text{ and}$$

$$0 = \left(\sum_{i=1}^{n} w_i \right)^2 = \left(\sum_{i=1}^{n} w_i \right) \left(\sum_{j=1}^{n} w_j \right)$$

$$= \sum_{i=1}^{n} w_i \sum_{j=1}^{n} w_j = \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j$$

$$= \sum_{i=1}^{n} w_i^2 + \sum_{i \neq j}^{n} w_i w_j = \frac{1}{V_x} + \sum_{i \neq j}^{n} w_i w_j$$

$$\Longrightarrow \sum_{i \neq j}^{n} w_i w_j = -\frac{1}{V_x}.$$

Lastly, use the above two results in equation $(**)$ to obtain

$$\text{Var}(\widehat{\beta}) = \sigma^2 \sum_{i=1}^{n} w_i^2 + \rho\sigma^2 \sum_{i \neq j} w_i w_j$$

$$= \frac{\sigma^2}{V_x} - \frac{\rho\sigma^2}{V_x} = \frac{\sigma^2(1-\rho)}{V_x} = \frac{\sigma^2(1-\rho)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

5.10 Independence

1. Definition: Continuous random variables X and Y are said to be independent if their joint pdf factors into a product of the marginal pdfs. That is,

$$X \perp\!\!\!\perp Y \Longleftrightarrow f_{X,Y}(x,y) = f_X(x) f_Y(y) \ \forall (x,y).$$

2. Example: if $f_{X,Y}(x,y) = 2$ for $x \in (0, 0.5)$ and $y \in (0,1)$ then $X \perp\!\!\!\perp Y$. Note, the joint pdf can be written as

$$f_{X,Y}(x,y) \quad = \quad 2I_{(0,0.5)}(x) I_{(0,1)}(y) = 2I_{(0,0.5)}(x) \times I_{(0,1)}(y)$$

$$= \quad f_X(x) \times f_Y(y)$$

where

$$I_A(x) = \begin{cases} 1 & x \in A; \\ 0 & \text{otherwise.} \end{cases}$$

3. Example: if $f_{X,Y}(x,y) = 8xy$ for $0 \le x \le y \le 1$, then X and Y are not independent. Note

$$f_{X,Y}(x,y) = 8xy I_{(0,1)}(y) I_{(0,y)}(x),$$

but

$$f_X(x) = \int_x^1 f_{X,Y}(x,y)\,dy = 4x(1-x^2) I_{(0,1)}(x),$$
$$f_Y(y) = \int_0^y f_{X,Y}(x,y)\,dx = 4y^3 I_{(0,1)}(y), \text{ and}$$
$$f_X(x) f_Y(y) = 16xy^3(1-x^2) I_{(0,1)}(x) I_{(0,1)}(y) \neq f_{X,Y}(x,y).$$

4. Note: $\text{Cov}(X,Y) = 0 \not\Longrightarrow X \perp\!\!\!\perp Y$. For example, if

$$f_{X,Y}(x,y) = \frac{1}{3} I_{(1,2)}(x) I_{(-x,x)}(y),$$

then

$$\begin{aligned} \text{E}(X) &= \int_1^2 \int_{-x}^{x} \frac{x}{3} dy dx = \int_1^2 = \frac{2x^2}{3} dx = \frac{14}{9}, \\ \text{E}(Y) &= \int_1^2 \int_{-x}^{x} \frac{y}{3} dy dx = \int_1^2 \frac{x^2 - x^2}{6} dx = 0, \text{ and} \\ \text{E}(XY) &= \int_1^2 \int_{-x}^{x} \frac{xy}{3} dy dx = \int_1^2 \frac{x(x^2 - x^2)}{6} dx = 0. \end{aligned}$$

Accordingly, X and Y have correlation 0, but they are not independent.

5. Result: Let A and B be subsets of the real line. Then random variables X and Y are independent if and only if

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

for all choices of sets A and B.

Proof: First assume that $X \perp\!\!\!\perp Y$. Let A and B be arbitrary sets on the real line. Then

$$P(X \in A, Y \in B) = \int_A \int_B f_{X,Y}(x,y)\,dy\,dx$$
$$= \int_A \int_B f_X(x) f_Y(y)\,dy\,dx \text{ by independence}$$

$$= \int_A f_X(x) \int_B f_Y(y)\, dy\, dx = P(X \in A)P(Y \in B)$$

Therefore,

$$X \perp\!\!\!\perp Y \Longrightarrow P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any choice of sets. Second, assume that
$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all choices of sets A and B.
Choose $A = (-\infty, x]$ and choose $B = (-\infty, y]$. Then

$$P(X \in A, Y \in B) = P(X \leq x, Y \leq y) = F_{X,Y}(x, y)$$
$$= P(X \in A)P(Y \in B) = P(X \leq x)P(Y \leq y) = F_X(x)F_Y(y).$$

Accordingly,

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x\, \partial y} F_{X,Y}(x, y)$$
$$= \frac{\partial^2}{\partial x\, \partial y} F_X(x)F_Y(y)$$
$$= \frac{\partial}{\partial x} F_X(x)\, \frac{\partial}{\partial y} F_Y(y) = f_X(x)f_Y(y).$$

Therefore

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B) \Longrightarrow f_{X,Y}(x, y) = f_X(x)f_Y(y).$$

6. Result: If X and Y are independent, then so are $g(X)$ and $h(Y)$ for any g and h.

   Proof: Let A be any set of intervals in the range of $g(x)$ and let B be any set of intervals in the range of $h(y)$. Denote by $g^{-1}(A)$ the set of all intervals in the support of X that satisfy $x \in g^{-1}(A) \Longleftrightarrow g(x) \in A$. Similarly, denote by $h^{-1}(B)$ the set of all intervals in the support of Y that satisfy $y \in h^{-1}(B) \Longleftrightarrow h(y) \in B$. If $X \perp\!\!\!\perp Y$, then,

$$P[g(X) \in A, h(Y) \in B] = P\left[X \in g^{-1}(A), Y \in h^{-1}(B)\right]$$
$$= P\left[X \in g^{-1}(A)\right] \times P\left[Y \in h^{-1}(B)\right] = P[g(X) \in A] \times P[h(Y) \in B].$$

   The above equality implies that $g(X) \perp\!\!\!\perp h(Y)$ because the factorization is satisfied for all A and B in the range spaces of $g(X)$ and $h(Y)$. Note that we already proved this result for discrete random variables.

7. The previous two results readily extend to any number of random variables (not just two).

8. Suppose that X_i for $i = 1, \ldots, n$ are independent. Then

(a) $g_1(X_1), \ldots, g_n(X_n)$ are independent,

(b) The Xs in any subset are independent,

(c) $\text{Var}\left(\sum a_i X_i\right) = \sum a_i^2 \text{Var}(X_i)$, and

(d) if the Xs are iid with variance σ^2, then $\text{Var}\left(\sum a_i X_i\right) = \sigma^2 \sum a_i^2$.

5.11 Conditional Distributions

1. Definition: If $f_{X,Y}(x,y)$ is a joint pdf, then the pdf of Y, conditional on $X = x$ is
$$f_{Y|X}(y|x) \stackrel{\text{def}}{=} \frac{f_{X,Y}(x,y)}{f_X(x)}$$
provided that $f_X(x) > 0$.

2. Example: Suppose that X and Y have joint distribution
$$f_{X,Y}(x,y) = 8xy \text{ for } 0 < x < y < 1.$$
Then,
$$\begin{aligned}
f_X(x) &= \int_x^1 f_{X,Y}(x,y)dy = 4x(1-x^2), \quad 0 < x < 1; \\
\text{E}(X^r) &= \int_0^1 4x(1-x^2)x^r dx = \frac{8}{(r+2)(r+4)}; \\
f_Y(y) &= 4y^3, \quad 0 < y < 1; \\
\text{E}(Y^r) &= \int_0^1 4y^3 y^r dy = \frac{4}{r+4} \\
f_{X|Y}(x|y) &= \frac{8xy}{4y^3} = \frac{2x}{y^2}, \quad 0 < x < y; \text{ and} \\
f_{Y|X}(y|x) &= \frac{8xy}{4x(1-x^2)} = \frac{2y}{1-x^2}, \quad x < y < 1.
\end{aligned}$$
Furthermore,
$$\begin{aligned}
\text{E}(X^r|Y=y) &= \int_0^y x^r \frac{2x}{y^2} dx = \frac{2y^r}{r+2} \text{ and} \\
\text{E}(Y^r|X=x) &= \int_x^1 y^r \frac{2y}{1-x^2} dy = \frac{2(1-x^{r+2})}{(r+2)(1-x^2)}.
\end{aligned}$$

3. Regression Function: Let (X,Y) be a pair of random variables with joint pdf $f_{X,Y}(x,y)$. Consider the problem of predicting Y after observing $X = x$. Denote the predictor as $\hat{y}(x)$. The best predictor is defined as the function $\hat{Y}(X)$ that minimizes
$$SSE = \text{E}\left[Y - \hat{Y}(X)\right]^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} [y - \hat{y}(x)]^2 f_{X,Y}(x,y)dydx.$$

(a) Result: The best predictor is $\hat{y}(x) = \mathrm{E}(Y|X = x)$.

Proof: Write $f_{X,Y}(x, y)$ as $f_{Y|X}(y|x)f_X(x)$. Accordingly,

$$SSE = \int_{-\infty}^{\infty} \left\{ \int_{-\infty}^{\infty} [y - \hat{y}(x)]^2 f_{Y,|X}(y, x)dy \right\} f_X(x)dx.$$

To minimize *SSE*, minimize the quantity in { } for each value of x. Note that $\hat{y}(x)$ is a constant with respect to the conditional distribution of Y given $X = x$. By the parallel axis theorem, the quantity in { } is minimized by $\hat{y}(x) = \mathrm{E}(Y|X = x)$.

(b) Example: Suppose that X and Y have joint distribution

$$f_{X,Y}(x, y) = 8xy \text{ for } 0 < x < y < 1.$$

Then,

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{8xy}{4x(1-x^2)} = \frac{2y}{1-x^2}, \quad x < y < 1 \text{ and} \\ \hat{y}(x) &= \mathrm{E}(Y|X = x) = \int_x^1 y \frac{2y}{1-x^2} dy = \frac{2(1-x^3)}{3(1-x^2)}. \end{aligned}$$

(c) Example; Suppose that (Y, X) has a bivariate normal distribution with moments $\mathrm{E}(Y) = \mu_Y$, $\mathrm{E}(X) = \mu_X$, $\mathrm{Var}(X) = \sigma_X^2$, $\mathrm{Var}(Y) = \sigma_Y^2$, and $\mathrm{Cov}(X, Y) = \rho_{X,Y}\sigma_X\sigma_Y$. Then it can be shown (we will not do so) that the conditional distribution of Y given X is

$$\begin{aligned} (Y|X = x) &\sim \mathrm{N}(\alpha + \beta x, \sigma^2), \text{ where} \\ \beta &= \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{\rho_{X,Y}\sigma_Y}{\sigma_X}; \quad \alpha = \mu_Y - \beta\mu_X \text{ and} \\ \sigma^2 &= \sigma_Y^2 \left(1 - \rho_{X,Y}^2\right). \end{aligned}$$

4. Averaging Conditional pdfs and Moments (be able to prove any of these results)

(a) $\mathrm{E}_X \left[f_{Y|X}(y|X) \right] = f_Y(y)$.

Proof:

$$\begin{aligned} \mathrm{E}_X \left[f_{Y|X}(y|X) \right] &= \int_{-\infty}^{\infty} f_{Y|X}(y|x) f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} \frac{f_{X,Y}(X, Y)}{f_X(x)} f_X(x)\, dx \\ &= \int_{-\infty}^{\infty} f_{X,Y}(X, Y)\, dx = f_Y(y). \end{aligned}$$

(b) $\mathrm{E}_X\left\{\mathrm{E}[h(Y)|X]\right\} = \mathrm{E}[h(Y)]$. This is the rule of iterated expectation. A special case is $\mathrm{E}_X\left[\mathrm{E}(Y|X)\right] = \mathrm{E}(Y)$.

Proof:

$$\begin{aligned}
\mathrm{E}_X\left\{\mathrm{E}[h(Y)|X]\right\} &= \int_{-\infty}^{\infty} \mathrm{E}[h(Y)|x] f_X(x)\,dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y) f_{Y|X}(y|x)\,dy f_X(x)\,dx \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y)\frac{f_{X,Y}(x,y)}{f_X(x)}\,dy f_X(x)\,dx \\
= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y) f_{X,Y}(x,y)\,dy\,dx &= \int_{-\infty}^{\infty} h(y)\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dx\,dy \\
&= \int_{-\infty}^{\infty} h(y) f_Y(y)\,dy = \mathrm{E}[h(Y)].
\end{aligned}$$

(c) $\mathrm{Var}(Y) = \mathrm{E}_X\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left[\mathrm{E}(Y|X)\right]$. That is, the variance of Y is equal to the expectation of the conditional variance plus the variance of the conditional expectation.

Proof:

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{E}(Y^2) - \left[\mathrm{E}(Y)\right]^2 \\
&= \mathrm{E}_X\left[\mathrm{E}(Y^2|X)\right] - \left\{\mathrm{E}_X\left[\mathrm{E}(Y|X)\right]\right\}^2 \\
&\quad \text{by the rule of iterated expectation} \\
&= \mathrm{E}_X\left\{\mathrm{Var}(Y|X) + \left[\mathrm{E}(Y|X)\right]^2\right\} - \left\{\mathrm{E}_X\left[\mathrm{E}(Y|X)\right]\right\}^2 \\
&\quad \text{because } \mathrm{Var}(Y|X) = \mathrm{E}(Y^2|X) - \left[\mathrm{E}(Y|X)\right]^2 \\
&= \mathrm{E}_X\left[\mathrm{Var}(Y|X)\right] + \mathrm{E}_X\left[\mathrm{E}(Y|X)\right]^2 - \left\{\mathrm{E}_X\left[\mathrm{E}(Y|X)\right]\right\}^2 \\
&= \mathrm{E}_X\left[\mathrm{Var}(Y|X)\right] + \mathrm{Var}\left[\mathrm{E}(Y|X)\right] \\
&\quad \text{because } \mathrm{Var}[\mathrm{E}(Y|X)] = \mathrm{E}_X\left[\mathrm{E}(Y|X)\right]^2 - \left\{\mathrm{E}_X\left[\mathrm{E}(Y|X)\right]\right\}^2.
\end{aligned}$$

5. Example: Suppose that X and Y have joint distribution

$$f_{X,Y}(x,y) = \frac{3y^2}{x^3} \text{ for } 0 < y < x < 1.$$

Then,

$$\begin{aligned}
f_Y(y) &= \int_y^1 \frac{3y^2}{x^3}dx = \frac{3}{2}(1-y^2), \text{ for } 0 < y < 1; \\
\mathrm{E}(Y^r) &= \int_0^1 \frac{3}{2}(1-y^2)y^r dy = \frac{3}{(r+1)(r+3)}; \\
&\Longrightarrow \mathrm{E}(Y) = \frac{3}{8} \text{ and } \mathrm{Var}(Y) = \frac{19}{320}; \\
f_X(x) &= \int_0^x \frac{3y^2}{x^2}dy = 1, \text{ for } 0 < x < 1;
\end{aligned}$$

$$\begin{aligned}
f_{Y|X}(y|x) &= \frac{3y^2}{x^3}, \text{ for } 0 < y < x < 1; \\
\mathrm{E}(Y^r|X=x) &= \int_0^x \frac{3y^2}{x^3} y^r dy = \frac{3x^r}{3+r} \\
&\Longrightarrow \mathrm{E}(Y|X=x) = \frac{3x}{4} \text{ and} \\
&\quad \mathrm{Var}(Y|X=x) = \frac{3x^2}{80}; \\
\mathrm{Var}\left[\mathrm{E}(Y|X)\right] &= \mathrm{Var}\left(\frac{3X}{4}\right) = \frac{9}{16} \times \frac{1}{12} = \frac{3}{64}; \\
\mathrm{E}\left[\mathrm{Var}(Y|X)\right] &= \mathrm{E}\left(\frac{3X^2}{80}\right) = \frac{1}{80}; \\
\frac{19}{320} &= \frac{3}{64} + \frac{1}{80}.
\end{aligned}$$

5.12 Moment Generating Functions

1. Definition: If X is a continuous random variable, then the mgf of X is

$$\psi_X(t) = \mathrm{E}\left(e^{tX}\right) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx,$$

provided that the expectation exists for t in a neighborhood of 0. If X is discrete, then replace integration by summation. If all of the moments of X do not exist, then the mgf will not exist. Note that the mgf is related to the pgf by

$$\psi_X(t) = \eta_X(e^t)$$

whenever $\eta_X(t)$ exists for t in a neighborhood of 1. Also note that if $\psi_X(t)$ is a mgf, then $\psi_X(0) = 1$.

2. Example: Exponential Distribution. If $f_X(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x)$, then

$$\begin{aligned}
\psi_X(t) &= \int_0^{\infty} e^{tx} \lambda e^{-\lambda x} dx \\
&= \frac{\lambda}{\lambda - t} \int_0^{\infty} (\lambda - t) e^{-(\lambda - t)x} dx \\
&= \frac{\lambda}{\lambda^*} \int_0^{\infty} \lambda^* e^{-\lambda^* x} dx, \text{ where } \lambda^* = \lambda - t, \\
&= \frac{\lambda}{\lambda^*} = \frac{\lambda}{\lambda - t} \text{ provided that } \lambda > t.
\end{aligned}$$

3. Example: Geometric Distribution. If $X \sim \mathrm{Geom}(p)$, then

$$\psi_X(t) = \sum_{x=1}^{\infty} e^{tx} (1-p)^{x-1} p = p e^t \sum_{x=1}^{\infty} (1-p)^{x-1} e^{t(x-1)}$$

$$= pe^t \sum_{x=0}^{\infty} \left[(1-p)e^t\right]^x = \frac{pe^t}{1-(1-p)e^t}$$

provided that $t < -\ln(1-p)$.

4. MGF of a linear function: If $\psi_X(t)$ exists, then

$$\psi_{a+bX}(t) = \mathrm{E}\left[e^{t(a+bX)}\right] = e^{at}\psi_X(tb).$$

For example, if $Z = (X - \mu_X)/\sigma_X$, then

$$\psi_Z(t) = e^{-t\mu_X/\sigma_X}\psi_X(t/\sigma_X).$$

5. Independent Random Variables: If X_i for $i = 1, \ldots, n$ are independent, $\psi_{X_i}(t)$ exists for each i, and $S = \sum X_i$, then

$$\psi_S(t) = \mathrm{E}\left(e^{t\sum X_i}\right) = \mathrm{E}\left[\prod_{i=1}^{n} e^{tX_i}\right] = \prod_{i=1}^{n} \psi_{X_i}(t).$$

If the Xs are iid random variables, then

$$\psi_S(t) = \left[\psi_X(t)\right]^n.$$

6. Result: Moment generating functions are unique. Each distribution has a unique moment generating function and each moment generating function corresponds to exactly one distribution. Accordingly, if the moment generating function exists, then it uniquely determines the distribution. For example, if the mgf of Y is

$$\psi_Y(t) = \frac{e^t}{2-e^t} = \frac{\frac{1}{2}e^t}{1-\frac{1}{2}e^t},$$

then $Y \sim \mathrm{Geom}(0.5)$.

7. Computing Moments. Consider the derivative of $\psi_X(t)$ with respect to t evaluated at $t = 0$:

$$\begin{aligned} \frac{d}{dt}\psi_X(t)\bigg|_{t=0} &= \int_{-\infty}^{\infty} \frac{d}{dt}e^{tx}\bigg|_{t=0} f_X(x)dx \\ &= \int_{-\infty}^{\infty} x f_X(x)dx = \mathrm{E}(X). \end{aligned}$$

Similarly, higher order moments can be found by taking higher order derivatives:

$$\mathrm{E}(X^r) = \frac{d^r}{(dt)^r}\psi_X(t)\bigg|_{t=0}.$$

Alternatively, expand e^{tx} around $t = 0$ to obtain

$$e^{tx} = \sum_{r=0}^{\infty} \frac{(tx)^r}{r!}.$$

Therefore

$$\begin{aligned} \psi_X(t) &= \mathrm{E}\left(e^{tX}\right) = \mathrm{E}\left[\sum_{r=0}^{\infty} \frac{(tX)^r}{r!}\right] \\ &= \sum_{r=0}^{\infty} \mathrm{E}(X^r)\frac{t^r}{r!}. \end{aligned}$$

Accordingly, $\mathrm{E}(X^r)$ is the coefficient of $t^r/r!$ in the expansion of the mgf.

8. Example: Suppose that $X \sim \mathrm{Geom}(p)$. Then the moments of X are

$$\mathrm{E}(X^r) = \frac{d^r}{(dt)^r}\psi_X(t)\Big|_{t=0} = \frac{d^r}{(dt)^r}\left[\frac{pe^t}{1-(1-p)e^t}\right]\Big|_{t=0}.$$

Specifically,

$$\begin{aligned} &\frac{d}{dt}\psi_X(t) = \frac{d}{dt}\left[\frac{pe^t}{1-(1-p)e^t}\right] = \\ &\quad \psi_X(t) + \frac{1-p}{p}\psi_X(t)^2 \text{ and} \\ &\frac{d^2}{(dt)^2}\psi_X(t) = \frac{d}{dt}\left[\psi_X(t) + \frac{1-p}{p}\psi_X(t)^2\right] = \\ &\quad \psi_X(t) + \frac{1-p}{p}\psi_X(t)^2 + \frac{1-p}{p}2\psi_X(t)\left[\psi_X(t) + \frac{1-p}{p}\psi_X(t)^2\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}(X) &= 1 + \frac{1-p}{p} = \frac{1}{p}; \\ \mathrm{E}(X^2) &= 1 + \frac{1-p}{p} + \frac{1-p}{p}2\left[1 + \frac{1-p}{p}\right] = \frac{2-p}{p^2} \text{ and} \\ \mathrm{Var}(X) &= \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}. \end{aligned}$$

9. Example: Suppose $Y \sim \mathrm{Unif}(a, b)$. Use the mgf to find the central moments $\mathrm{E}[(Y - \mu_Y)^r] = \mathrm{E}[(Y - \frac{a+b}{2})^r]$. Solution:

$$\begin{aligned} \psi_Y(t) &= \frac{1}{b-a}\int_a^b e^{ty}dy = \frac{e^{tb} - e^{ta}}{t(b-a)} \\ \psi_{Y-\mu_Y}(t) &= e^{-t(a+b)/2}\psi_Y(t) \end{aligned}$$

$$
\begin{aligned}
&= \frac{e^{-t(a+b)/2}\left[e^{tb}-e^{ta}\right]}{t(b-a)} \\
&= \left(\frac{2}{t(b-a)}\right)\frac{e^{\frac{t}{2}(b-a)}-e^{-\frac{t}{2}(b-a)}}{2} \\
&= \left(\frac{2}{t(b-a)}\right)\sinh\left(\frac{t(b-a)}{2}\right) \\
&= \frac{2}{t(b-a)}\sum_{i=0}^{\infty}\left(\frac{t(b-a)}{2}\right)^{2i+1}\frac{1}{(2i+1)!} \\
&= \sum_{i=0}^{\infty}\left(\frac{t(b-a)}{2}\right)^{2i}\frac{1}{(2i+1)!} = \sum_{i=0}^{\infty}\left(\frac{t^{2i}}{(2i)!}\right)\frac{(b-a)^{2i}}{2^{2i}(2i+1)}.
\end{aligned}
$$

Therefore, the odd moments are zero, and

$$\mathrm{E}(Y-\mu_Y)^{2i} = \frac{(b-a)^{2i}}{4^i(2i+1)}.$$

For example, $\mathrm{E}(Y-\mu_Y)^2 = (b-a)^2/12$ and $\mathrm{E}(Y-\mu_Y)^4 = (b-a)^4/80$.

Chapter 6

FAMILIES OF CONTINUOUS DISTRIBUTIONS

6.1 Normal Distributions

1. PDF and cdf of the Standard Normal Distribution:

$$\begin{aligned} f_Z(z) &= \frac{e^{-z^2/2}}{\sqrt{2\pi}} I_{(-\infty,\infty)}(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}} \text{ and} \\ F_Z(z) &= P(Z \leq z) = \Phi(z) = \int_{-\infty}^{z} f_Z(u)\, du \end{aligned}$$

2. Result: $\displaystyle\int_{-\infty}^{\infty} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, dx = 1.$

   Proof: To verify that $f_Z(z)$ integrates to one, it is sufficient to show that

$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi}.$$

   Let

$$K = \int_{-\infty}^{\infty} e^{-x^2/2}\, dx.$$

   Then

$$\begin{aligned} K^2 &= \left(\int_{-\infty}^{\infty} e^{-u^2/2}\, du \right)^2 \\ &= \left(\int_{-\infty}^{\infty} e^{-u_1^2/2}\, du_1 \right) \left(\int_{-\infty}^{\infty} e^{-u_2^2/2}\, du_2 \right) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(u_1^2+u_2^2)}\, du_1\, du_2. \end{aligned}$$

   Now transform to polar coordinates:

$$u_1 = r \sin\theta;\ u_2 = r\cos\theta \text{ and}$$

$$\begin{aligned} K^2 &= \int_0^{2\pi}\int_0^{\infty} e^{-\frac{1}{2}(r^2)} r\, dr\, d\theta \\ &= \int_0^{2\pi}\left[-e^{-\frac{1}{2}(r^2)}\Big|_0^{\infty}\right] d\theta \\ &= \int_0^{2\pi} 1\, d\theta = 2\pi. \end{aligned}$$

Therefore $K = \sqrt{2\pi}$ and $f_Z(z)$ integrates to one.

3. Other Normal Distributions: Transform from Z to $X = \mu + \sigma Z$, where μ and σ are constants that satisfy $|\mu| < \infty$ and $0 < \sigma < \infty$. The inverse transformation is $z = (x-\mu)/\sigma$ and the Jacobian of the transformation is

$$|J| = \left|\frac{dz}{dx}\right| = \frac{1}{\sigma}.$$

Accordingly, the pdf of X is

$$f_X(x) = f_Z\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma} = \frac{e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi\sigma^2}}.$$

We will use the notation $X \sim \mathrm{N}(\mu, \sigma^2)$ to mean that X has a normal distribution with parameters μ and σ.

4. Completing a square. Let a and b be constants. Then $x^2 - 2ax + b = (x-a)^2 - a^2 + b$ for all x.

Proof:

$$\begin{aligned} x^2 - ax + b &= x^2 - 2\left(\frac{a}{2}\right)x + \left(\frac{a}{2}\right)^2 - \left(\frac{a}{2}\right)^2 + b \\ &= \left[x - \left(\frac{a}{2}\right)\right]^2 - \left(\frac{a}{2}\right)^2 + b. \end{aligned}$$

5. Moment Generating Function: Suppose that $X \sim \mathrm{N}(\mu, \sigma^2)$. Then $\psi_X(t) = e^{\mu t + t^2\sigma^2/2}$.

Proof:

$$\begin{aligned} \psi_X(t) &= \mathrm{E}(e^{tX}) = \int_{-\infty}^{\infty} e^{tx}\frac{e^{-\frac{1}{2\sigma^2}(x-\mu)^2}}{\sqrt{2\pi\sigma^2}}\, dx \\ &= \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2\sigma^2}\left[-2t\sigma^2 x + (x-\mu)^2\right]}}{\sqrt{2\pi\sigma^2}}\, dx. \end{aligned}$$

Now complete the square in the exponent:

$$-2t\sigma^2 x + (x-\mu)^2 = -2t\sigma^2 x + x^2 - 2x\mu + \mu^2 = x^2 - 2x(\mu + t\sigma^2) + \mu^2$$

$$= \left[x - (\mu + t\sigma^2)\right]^2 - (\mu + t\sigma^2)^2 + \mu^2 = \left[x - (\mu + t\sigma^2)\right]^2 - 2t\mu\sigma^2 - t^2\sigma^4.$$

Therefore,

$$\begin{aligned}
\psi_X(t) &= e^{-\frac{1}{2\sigma^2}(-2t\mu\sigma^2 - t^2\sigma^4)} \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2\sigma^2}\left[x-(\mu+t\sigma^2)\right]^2}}{\sqrt{2\pi\sigma^2}}\, dx \\
&= e^{t\mu + t^2\sigma^2/2} \int_{-\infty}^{\infty} \frac{e^{-\frac{1}{2\sigma^2}(x-\mu^*)^2}}{\sqrt{2\pi\sigma^2}}\, dx \text{ where } \mu^* = \mu + t\sigma^2 \\
&= e^{t\mu + t^2\sigma^2/2}
\end{aligned}$$

because the second term is the integral of the pdf of a random variable with distribution $\mathrm{N}(\mu^*, \sigma^2)$ and this integral is one.

6. Moments of Normal Distributions

   (a) Moments of the standard normal distribution: Let Z be a normal random variable with $\mu = 0$ and $\sigma = 1$. That is, $Z \sim \mathrm{N}(0, 1)$. The moment generating function of Z is $\psi_Z(t) = e^{t^2/2}$. The Taylor series expansion of $\psi_Z(t)$ around $t = 0$ is

$$\begin{aligned}
\psi_Z(t) &= e^{t^2/2} = \sum_{i=0}^{\infty} \frac{1}{i!} \left(\frac{t^2}{2}\right)^i \\
&= \sum_{i=0}^{\infty} \left(\frac{(2i)!}{2^i i!}\right) \left(\frac{t^{2i}}{(2i)!}\right).
\end{aligned}$$

   Note that all odd powers in the expansion are zero. Accordingly,

$$\mathrm{E}(Z^r) = \begin{cases} 0 & \text{if } r \text{ is odd} \\ \dfrac{r!}{2^{r/2}\left(\frac{r}{2}\right)!} & \text{if } r \text{ is even.} \end{cases}$$

   It can be shown by induction that if r is even, then

$$\frac{r!}{2^{r/2}\left(\frac{r}{2}\right)!} = (r-1)(r-3)(r-5)\cdots 1.$$

   In particular, $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = \mathrm{E}(Z^2) = 1$.

   (b) Moments of Other Normal Distributions: Suppose that $X \sim \mathrm{N}(\mu, \sigma^2)$. Then X can be written as $X = \mu + Z\sigma$, where $Z \sim \mathrm{N}(0, 1)$. To obtain the moments of X, one may use the moments of Z or one may differentiate the moment generating function of X. For example, using the moments of Z, the first two moments of X are

$$\begin{aligned}
\mathrm{E}(X) &= \mathrm{E}(\mu + \sigma Z) = \mu + \sigma\mathrm{E}(Z) = \mu \text{ and} \\
\mathrm{E}(X^2) &= \mathrm{E}\left[(\mu + \sigma Z)^2\right] = \mathrm{E}(\mu^2 + 2\mu\sigma Z + \sigma^2 Z^2) = \mu^2 + \sigma^2.
\end{aligned}$$

Note that $\text{Var}(X) = \text{E}(X^2) - [\text{E}(X)]^2 = \sigma^2$. The alternative approach is to use the moment generating function:

$$\begin{aligned}
\text{E}(X) &= \frac{d}{dt}\psi_X(t)\Big|_{t=0} = \frac{d}{dt}e^{t\mu+t^2\sigma^2/2}\Big|_{t=0} \\
&= (\mu + t\sigma^2)e^{t\mu+t^2\sigma^2/2}\Big|_{t=0} = \mu \text{ and} \\
\text{E}(X^2) &= \frac{d^2}{dt^2}\psi_X(t)\Big|_{t=0} = \frac{d}{dt}(\mu + t\sigma^2)e^{t\mu+t^2\sigma^2/2}\Big|_{t=0} \\
&= \sigma^2 e^{t\mu+t^2\sigma^2/2} + (\mu + t\sigma^2)^2 e^{t\mu+t^2\sigma^2/2}\Big|_{t=0} = \mu^2 + \sigma^2.
\end{aligned}$$

7. Box-Muller method for generating standard normal variables. Let Z_1 and Z_2 be iid random variables with distributions $Z_i \sim \text{N}(0,1)$. The joint pdf of Z_1 and Z_2 is

$$f_{Z_1,Z_2}(z_1, z_2) = \frac{e^{-\frac{1}{2}(z_1^2+z_2^2)}}{2\pi}.$$

Transform to polar coordinates: $Z_1 = R\sin(T)$ and $Z_2 = R\cos(T)$. The joint distribution of R and T is

$$\begin{aligned}
f_{R,T}(r,t) &= \frac{re^{-\frac{1}{2}r^2}}{2\pi} I_{(0,\infty)}(r) I_{(0,2\pi)}(T) = f_R(r) \times f_T(t) \text{ where} \\
f_R(r) &= re^{-\frac{1}{2}r^2} I_{(0,\infty)}(r) \text{ and } f_T(t) = \frac{1}{2\pi} I_{(0,2\pi)}(t).
\end{aligned}$$

Factorization of the joint pdf reveals that $R \perp\!\!\!\perp T$. Their respective cdfs are

$$F_R(r) = 1 - e^{-\frac{1}{2}r^2} \text{ and } F_T(t) = \frac{t}{2\pi}.$$

Let $U_1 = F_R(R)$ and $U_2 = F_T(T)$. Recall that $U_i \sim \text{Unif}(0,1)$. Solving the cdf equations for R and T yields

$$R = \sqrt{-2\ln(1-U_1)} \text{ and } T = 2\pi U_2.$$

Lastly, express Z_1 and Z_2 as functions of R and T:

$$\begin{aligned}
Z_1 &= R\sin(T) = \sqrt{-2\ln(1-U_1)}\sin(2\pi U_2) \text{ and} \\
Z_2 &= R\cos(T) = \sqrt{-2\ln(1-U_1)}\cos(2\pi U_2).
\end{aligned}$$

Note that U_1 and $1 - U_1$ have the same distributions. Therefore Z_1 and Z_2 can be generated from U_1 and U_2 by

$$Z_1 = \sqrt{-2\ln(U_1)}\sin(2\pi U_2) \text{ and } Z_2 = \sqrt{-2\ln(U_1)}\cos(2\pi U_2).$$

8. Linear Functions of Normal Random Variables: Suppose that X and Y are independently distributed random variables with distributions $X \sim \text{N}(\mu_X, \sigma_X^2)$ and $Y \sim \text{N}(\mu_Y, \sigma_Y^2)$.

(a) The distribution of $aX + b$ is $\mathrm{N}(a\mu_X + b, a^2\sigma_X^2)$.

Proof: The moment generating function of $aX + b$ is

$$\begin{aligned}\psi_{aX+b}(t) = \mathrm{E}(e^{t(aX+b)}) = e^{tb}\mathrm{E}(e^{taX}) = e^{tb}\psi_X(ta)\\ = e^{tb}e^{ta\mu+t^2a^2\sigma^2/2} = e^{t(a\mu+b)+t^2(a\sigma)^2/2}\end{aligned}$$

and this is the moment generating function of a random variable with distribution $\mathrm{N}(a\mu + b, a^2\sigma^2)$.

(b) Application: Suppose that $X \sim \mathrm{N}(\mu, \sigma^2)$. Let $Z = (X - \mu)/\sigma$. Note, $Z = aX + b$, where $a = 1/\sigma$ and $b = -\mu/\sigma$. Accordingly, $Z \sim \mathrm{N}(0, 1)$.

(c) The distribution of $aX + bY$ is $\mathrm{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$.

Proof: The moment generating function of $aX + bY$ is

$$\begin{aligned}\psi_{aX+bY}(t) = \mathrm{E}(e^{t(aX+bY)}) = \mathrm{E}(e^{taX})\mathrm{E}(e^{tbY}) \text{ by independence}\\ = \psi_X(ta)\psi_Y(tb) = e^{ta\mu_X+t^2a^2\sigma_X^2/2}e^{tb\mu_Y+t^2b^2\sigma_Y^2/2}\\ = e^{t(a\mu_X+b\mu_Y)+t^2(a^2\sigma_X^2+b^2\sigma_Y^2)/2}.\end{aligned}$$

and this is the moment generating function of a random variable with distribution $\mathrm{N}(a\mu_X + b\mu_Y, a^2\sigma_X^2 + b^2\sigma_Y^2)$.

(d) The above result is readily generalized. Suppose that X_i for $i = 1, \ldots, n$ are independently distributed as $X_i \sim \mathrm{N}(\mu_i, \sigma_i^2)$. If $T = \sum_{i=1}^n a_i X_i$, then $T \sim \mathrm{N}(\mu_T, \sigma_T^2)$, where $\mu_T = \sum_{i=1}^n a_i\mu_i$ and $\sigma_T^2 = \sum_{i=1}^n a_i^2\sigma_i^2$.

9. Probabilities and Percentiles

(a) If $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$, then the probability of an interval is

$$\begin{aligned}P(a \le X \le b) &= P\left[\frac{a-\mu_X}{\sigma_X} \le Z \le \frac{b-\mu_X}{\sigma_X}\right]\\ &= \Phi\left(\frac{b-\mu_X}{\sigma_X}\right) - \Phi\left(\frac{a-\mu_X}{\sigma_X}\right).\end{aligned}$$

(b) If $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$, then the $100p^{\text{th}}$ percentile of X is

$$x_p = \mu_X + \sigma_X z_p,$$

where z_p is the $100p^{\text{th}}$ percentile of the standard normal distribution.

Proof:

$$P(X \le \mu_X + \sigma_X z_p) = P\left(\frac{X-\mu_X}{\sigma_X} \le z_p\right) = P(Z \le z_p) = p$$

because $Z = (X - \mu_X)/\sigma_X \sim \mathrm{N}(0, 1)$.

10. Log Normal Distribution

(a) Definition: If $\ln(X) \sim \mathrm{N}(\mu, \sigma^2)$, then X is said to have a log normal distribution. That is

$$\ln(X) \sim \mathrm{N}(\mu, \sigma^2) \Longleftrightarrow X \sim \mathrm{LogN}(\mu, \sigma^2).$$

Note: μ and σ^2 are the mean and variance of $\ln(X)$, not of X.

(b) PDF: Let $Y = \ln(X)$, and assume that $Y \sim \mathrm{N}(\mu, \sigma^2)$. Note that $x = g(y)$ and $y = g^{-1}(x)$, where $g(y) = e^y$ and $g^{-1}(x) = \ln(x)$. The Jacobian of the transformation is

$$|J| = \left|\frac{d}{dx} y\right| = \left|\frac{d}{dx} \ln(x)\right| = \frac{1}{x}.$$

Accordingly, the pdf of X is

$$f_X(x) = f_Y\left[g^{-1}(x)\right] \frac{1}{x} = \frac{e^{-\frac{1}{2\sigma^2}[\ln(x)-\mu]^2}}{x\sigma\sqrt{2\pi}} I_{(0,\infty)}(x).$$

(c) CDF: If $Y \sim \mathrm{LogN}(\mu, \sigma^2)$, then

$$P(Y \leq y) = P[\ln(Y) \leq \ln(y)] = \Phi\left(\frac{\ln(y) - \mu}{\sigma}\right).$$

(d) Moments of a log normal random variable. Suppose that $X \sim \mathrm{LogN}(\mu, \sigma^2)$. Then $\mathrm{E}(X^r) = e^{\mu r + r^2\sigma^2/2}$.

Proof: Let $Y = \ln(X)$. Then $X = e^Y$ and $Y \sim \mathrm{N}(\mu, \sigma^2)$ and

$$\mathrm{E}(X^r) = \mathrm{E}\left(e^{rY}\right) = e^{r\mu + r^2\sigma^2/2},$$

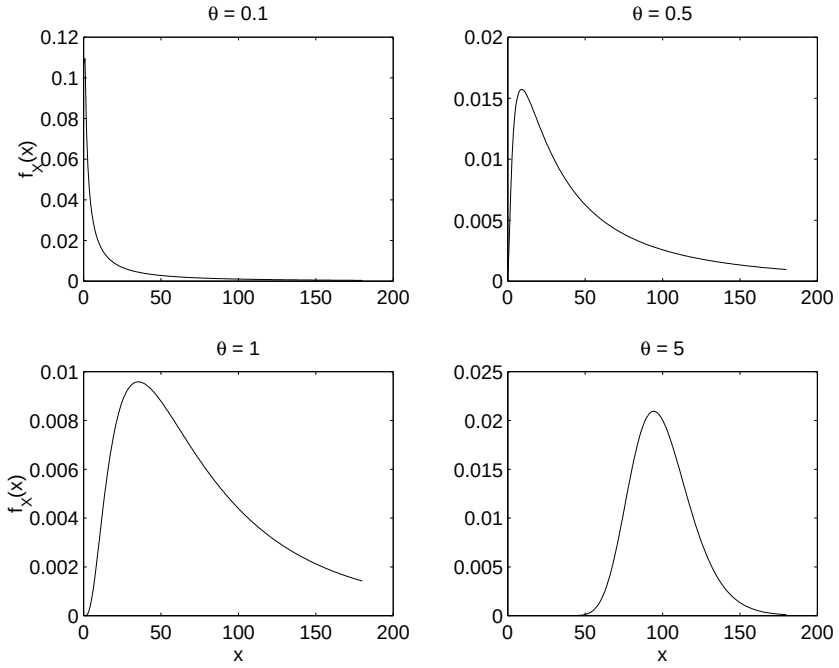
where the result is obtained by using the mgf of a normal random variable. To obtain the mean and variance, set r to 1 and 2:

$$\mathrm{E}(X) = e^{\mu+\sigma^2/2} \text{ and } \mathrm{Var}(X) = e^{2\mu+2\sigma^2} - e^{2\mu+\sigma^2} = e^{2\mu+\sigma^2}\left[e^{\sigma^2} - 1\right].$$

(e) Displays of various log normal distributions. The figure below displays four log normal distributions. The parameters of the distribution are summarized in the following table:

Plot	$\mu =$ $\mathrm{E}[\ln(X)]$	$\sigma^2 =$ $\mathrm{Var}[\ln(X)]$	$\tau =$ $\mathrm{E}(X)$	$\delta =$ $\sqrt{\mathrm{Var}(X)}$	$\theta =$ τ/δ
1	3.2976	4.6151	100	1000	0.1
2	3.8005	1.6094	100	200	0.5
3	4.2586	0.6931	100	100	1
4	4.5856	0.0392	100	20	5

Note that each distribution has mean equal to 100. The distributions differ in terms of θ, which is the coefficient of variation.



If the coefficient of variation is small, then the log normal distribution resembles an exponential distribution, As the coefficient of variation increases, the log normal distribution converges to a normal distribution.

6.2 Exponential Distributions

1. PDF and cdf

$$\begin{aligned} f_X(x) &= \lambda e^{-\lambda x} I_{[0,\infty)}(x) \text{ where } \lambda \text{ is a positive parameter, and} \\ F_X(x) &= 1 - e^{-\lambda x} \end{aligned}$$

provided that $x \geq 0$. We will use the notation $X \sim \text{Expon}(\lambda)$ to mean that X has an exponential distribution with parameter λ. Note that the $100p^{\text{th}}$ percentile is $x_p = -\ln(1-p)/\lambda$. The median, for example, is $x_{0.5} = \ln(2)/\lambda$.

2. Moment Generating Function. If $X \sim \text{Expon}(\lambda)$, then $\psi_X(t) = \lambda/(\lambda - t)$ for $t < \lambda$.

Proof:

$$\begin{aligned} \psi_X(t) &= \text{E}(e^{tX}) = \int_0^\infty e^{tx} \lambda e^{-\lambda x}\, dx \\ &= \int_0^\infty \lambda e^{(\lambda - t)x}\, dx = \frac{\lambda}{\lambda - t} \int_0^\infty (\lambda - t) e^{(\lambda - t)x}\, dx = \frac{\lambda}{\lambda - t} \end{aligned}$$

because the last integral is the integral of the pdf of a random variable with distribution $\text{Expon}(\lambda - t)$, provided that $\lambda - t > 0$.

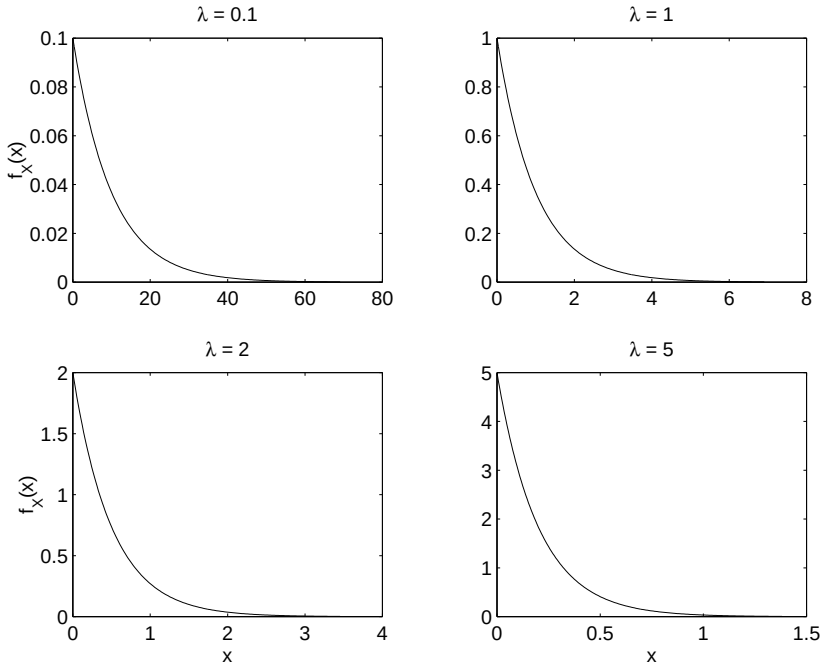
3. Moments: If $X \sim \text{Expon}(\lambda)$, then $\text{E}(X^r) = r!/\lambda^r$.

   Proof:

$$\begin{aligned}\psi_X(t) &= \frac{\lambda}{\lambda - t} = \frac{1}{1 - t/\lambda} = \sum_{r=0}^{\infty} \left(\frac{t}{\lambda}\right)^r \\ &= \sum_{r=0}^{\infty} \left(\frac{t^r}{r!}\right)\left(\frac{r!}{\lambda^r}\right)\end{aligned}$$

   provided that $-\lambda < t < \lambda$. Note that $\text{E}(X) = 1/\lambda$, $\text{E}(X^2) = 2/\lambda^2$ and $\text{Var}(X) = 1/\lambda^2$.

4. Displays of exponential distributions. Below are plots of four exponential distributions. Note that the shapes of the distributions are identical.



5. Memoryless Property: Suppose that $X \sim \text{Expon}(\lambda)$. The random variable can be thought of as the waiting time for an event to occur. Given that an event has not occurred in the interval $[0, w)$, find the probability that the additional waiting time is at least t. That is, find $P(X > t + w | X > w)$. Note: $P(X > t)$ is sometimes called the reliability function. It is denoted as $R(t)$ and is related to $F_X(t)$ by

$$R(t) = P(X > t) = 1 - P(X \leq t) = 1 - F_X(t).$$

   The reliability function represents the probability that the lifetime of a product (i.e., waiting for failure) is at least t units. For the exponential

distribution, the reliability function is $R(t) = e^{-\lambda t}$. We are interested in the conditional reliability function $R(t+w|X>w)$. Solution:

$$\begin{aligned} R(t+w|X>w) &= P(X>t+w|X>w) = \frac{P(X>t+w)}{P(X>w)} \\ &= \frac{e^{-\lambda(t+w)}}{e^{-\lambda w}} = e^{-\lambda t}. \end{aligned}$$

Also,

$$R(t+w|X>w) = 1 - F_X(t+w|X>w) \Longrightarrow F_X(t+w|X>w) = 1 - e^{-\lambda t}.$$

That is, no matter how long one has been waiting, the conditional distribution of the remaining life time is still Expon(λ). It is as though the distribution does not remember that we have already been waiting w time units.

6. Poison Inter-arrival Times: Suppose that events occur according to a Poisson process with rate parameter λ. Assume that the process begins at time 0. Let T_1 be the arrival time of the first event and let T_r be the time interval from the $(r-1)^{\text{st}}$ arrival to the r^{th} arrival. That is, $T_1, \ldots, T_n$ are inter-arrival times. Then T_i for $i = 1, \ldots, n$ are iid Expon(λ).

Proof: Consider the joint pdf or $T_1, T_2, \ldots, T_n$:

$$\begin{aligned} &f_{T_1,T_2,\ldots,T_n}(t_1, t_2, \ldots, t_n) = \\ &= f_{T_1}(t_1) \times f_{T_2|T_1}(t_2|t_1) \times f_{T_3|T_1,T_2}(t_3|t_1, t_2) \\ &\quad \times \cdots \times f_{T_n|T_1,\ldots,T_{n-1}}(t_n|t_1, \ldots, t_{n-1}) \end{aligned}$$

by the multiplication rule. To obtain the first term, first find the cdf of T_1:

$$\begin{aligned} F_{T_1}(t_1) &= P(T_1 \le t_1) = P\left[\text{one or more events in } (0, t_1)\right] \\ &= 1 - P\left[\text{no events in } (0, t_1)\right] = 1 - \frac{e^{-\lambda t_1}(\lambda t_1)^0}{0!} = 1 - e^{-\lambda t_1}. \end{aligned}$$

Differentiating the cdf yields

$$f_{T_1}(t_1) = \frac{d}{dt_1}(1 - e^{-\lambda t_1}) = \lambda e^{-\lambda t_1} I_{(0,\infty)}(t_1).$$

The second term is the conditional pdf of T_2 given $T_1 = t_1$. Recall that in a Poisson process, events in non-overlapping intervals are independent. Therefore,

$$f_{T_2|T_1}(t_2|t_1) = f_{T_2}(t_2) = \lambda e^{-\lambda t_2}.$$

Each of the remaining conditional pdfs also is just an exponential pdf. Therefore,

$$f_{T_1,T_2,\ldots,T_n}(t_1, t_2, \ldots, t_n) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} I_{[0,\infty)}(t_i).$$

This joint pdf is the product of n marginal exponential pdfs. Therefore, the inter-arrival times are iid exponential random variables. That is,
$T_i \sim$ iid Expon(λ).

6.3 Gamma Distributions

1. Erlang Distributions:

   (a) Consider a Poisson process with rate parameter λ. Assume that the process begins at time 0. Let Y be the time of the r^{th} arrival. Using the differential method, the pdf of Y can be obtained as follows:

$$\begin{aligned} P(y < Y < y + dy) &\approx P(r-1 \text{ arrivals before time } y) \\ &\quad \times P[\text{one arrival in } (y, y+dy)] \\ &= \frac{e^{-\lambda y}(\lambda y)^{r-1}}{(r-1)!} \times \lambda dy. \end{aligned}$$

   Accordingly,

$$f_Y(y) = \frac{e^{-\lambda y}\lambda^r y^{r-1}}{(r-1)!} I_{[0,\infty)}(y).$$

   The above pdf is called the Erlang pdf.

   (b) Note that Y is the sum of r iid Expon(λ) random variables (see page 49 of these notes). Accordingly, $\mathrm{E}(Y) = r/\lambda$ and $\mathrm{Var}(Y) = r/\lambda^2$.

   (c) CDF of an Erlang random variable: $F_Y(y) = 1 - P(Y > y)$ and $P(Y > y)$ is the probability that fewer than r events occur in $[0, y)$. Accordingly,

$$F_Y(y) = 1 - P(Y > y) = 1 - \sum_{i=0}^{r-1} \frac{e^{-\lambda y}(\lambda y)^i}{i!}.$$

2. Gamma Function

   (a) Definition: $\Gamma(\alpha) = \displaystyle\int_0^\infty u^{\alpha-1}e^{-u}\,du$, where $\alpha > 0$.

   (b) Alternative expression: Let $z = \sqrt{2u}$ so that $u = z^2/2$; $du = z\,dz$; and $\Gamma(\alpha) = \displaystyle\int_0^\infty \frac{z^{2\alpha-1}e^{-z^2/2}}{2^{\alpha-1}}\,dz$.

   (c) Properties of $\Gamma(\alpha)$

      i. $\Gamma(1) = 1$.

      Proof:

$$\Gamma(1) = \int_0^\infty e^{-w}dw = -e^{-w}\Big|_0^\infty = -0 + 1 = 1.$$

      ii. $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$.

      Proof:

$$\Gamma(\alpha+1) = \int_0^\infty w^\alpha e^{-w}dw.$$

Let $u = w^\alpha$, let $dv = e^{-w}dw$ and use integration by parts to obtain $du = \alpha w^{\alpha-1}$, $v = -e^{-w}$ and

$$\begin{aligned}\Gamma(\alpha+1) &= -w^\alpha e^{-w}\Big|_0^\infty + \alpha\int_0^\infty w^{\alpha-1}e^{-w}dw \\ &= 0 + \alpha\Gamma(\alpha).\end{aligned}$$

iii. If n is a positive integer, then $\Gamma(n) = (n-1)!$.
Proof: $\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2)$ etc.

iv. $\Gamma(\frac{1}{2}) = \sqrt{\pi}$.
Proof:

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty \frac{e^{-z^2/2}}{2^{-\frac{1}{2}}}dz = \int_{-\infty}^\infty \frac{e^{-z^2/2}}{\sqrt{2}}dz = \sqrt{\pi}$$

because the integral of the standard normal distribution is one.

3. Gamma Distribution

(a) PDF and cdf: If $Y \sim \text{Gam}(\alpha, \lambda)$, then

$$f_Y(y) = \frac{y^{\alpha-1}\lambda^\alpha e^{-\lambda y}}{\Gamma(\alpha)}I_{(0,\infty)}(y) \text{ and } F_Y(y) = \int_0^y \frac{u^{\alpha-1}\lambda^\alpha e^{-\lambda u}}{\Gamma(\alpha)}du.$$

(b) Note: α is called the shape parameter and λ is called the scale parameter.

(c) Moment Generating Function: If $Y \sim \text{Gam}(\alpha, \lambda)$, then

$$\begin{aligned}\psi_Y(t) &= \int_0^\infty e^{ty}\frac{y^{\alpha-1}\lambda^\alpha e^{-\lambda y}}{\Gamma(\alpha)}dy \\ &= \int_0^\infty \frac{y^{\alpha-1}\lambda^\alpha e^{-(\lambda-t)y}}{\Gamma(\alpha)}dy \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha}\int_0^\infty \frac{y^{\alpha-1}(\lambda-t)^\alpha e^{-(\lambda-t)y}}{\Gamma(\alpha)}dy \\ &= \frac{\lambda^\alpha}{(\lambda-t)^\alpha}\end{aligned}$$

because the last integral is the integral of a random variable with distribution $\text{Gam}(\alpha, \lambda - t)$ provided that $\lambda - t > 0$.

(d) Moments: If $Y \sim \text{Gam}(\alpha, \lambda)$, then

$$\begin{aligned}\text{E}(Y) &= \frac{d}{dt}\psi_Y(t)\Big|_{t=0} = \frac{\lambda^\alpha\alpha}{(\lambda-t)^{\alpha+1}}\Big|_{t=0} = \frac{\alpha}{\lambda}; \\ \text{E}(Y^2) &= \frac{d^2}{(dt)^2}\psi_Y(t)\Big|_{t=0} = \frac{\lambda^\alpha\alpha(\alpha+1)}{(\lambda-t)^{\alpha+2}}\Big|_{t=0} \\ &= \frac{\alpha(\alpha+1)}{\lambda^2}; \text{ and} \\ \text{Var}(Y) &= \text{E}(Y^2) - [\text{E}(Y)]^2 = \frac{\alpha(\alpha+1)}{\lambda^2} - \frac{\alpha^2}{\lambda^2} = \frac{\alpha}{\lambda^2}.\end{aligned}$$

(e) General expression for moments (including fractional moments). If $Y \sim \text{Gam}(\alpha, \lambda)$, then

$$\text{E}(Y^r) = \frac{\Gamma(\alpha + r)}{\lambda^r \, \Gamma(\alpha)} \text{ provided that } \alpha + r > 0.$$

Proof:

$$\begin{aligned} \text{E}(Y^r) &= \int_0^\infty \frac{y^r y^{\alpha-1} \lambda^\alpha e^{-\lambda y}}{\Gamma(\alpha)} dy = \int_0^\infty \frac{y^{\alpha+r-1} \lambda^\alpha e^{-\lambda y}}{\Gamma(\alpha)} dy \\ &= \frac{\Gamma(\alpha + r)}{\lambda^r \Gamma(\alpha)} \int_0^\infty \frac{y^{\alpha+r-1} \lambda^{\alpha+r} e^{-\lambda y}}{\Gamma(\alpha + r)} dy = \frac{\Gamma(\alpha + r)}{\lambda^r \Gamma(\alpha)} \end{aligned}$$

because the last integral is the integral of a random variable with distribution $\text{Gam}(\alpha + r, \lambda)$, provided that $\alpha + r > 0$.

(f) Distribution of the sum of iid exponential random variables. Suppose that $Y_1, Y_2, \ldots, Y_k$ are iid $\text{Expon}(\lambda)$ random variables. Then $T = \sum_{i=1}^k Y_i \sim \text{Gam}(k, \lambda)$.

Proof: $\psi_{Y_i}(t) = \lambda/(\lambda - t) \Longrightarrow \psi_T(t) = \lambda^k/(\lambda - t)^k$.

(g) Note that the Erlang distribution is a gamma distribution with shape parameter α equal to an integer.

6.4 Chi Squared Distributions

1. Definition: Let Z_i for $i = 1, \ldots k$ be iid $\text{N}(0, 1)$ random variables. Then $Y = \sum_{i=1}^k Z_i^2$ is said to have a χ^2 distribution with k degrees of freedom. That is, $Y \sim \chi^2_k$.

2. MGF: $\psi_Y(t) = (1 - 2t)^{-\frac{k}{2}}$ for $t < 0.5$.

   Proof: First find the mgf of Z_i^2:

$$\begin{aligned} \psi_{Z_i^2}(t) &= \text{E}(e^{tZ^2}) = \int_{-\infty}^\infty e^{tz^2} \frac{e^{-\frac{1}{2}z^2}}{\sqrt{2\pi}} dz \\ &= \int_{-\infty}^\infty \frac{e^{-\frac{1}{2(1-2t)^{-1}}z^2}}{\sqrt{2\pi}} dz = (1 - 2t)^{-\frac{1}{2}} \int_{-\infty}^\infty \frac{e^{-\frac{1}{2(1-2t)^{-1}}z^2}}{(1 - 2t)^{-\frac{1}{2}}\sqrt{2\pi}} dz \\ &= (1 - 2t)^{-\frac{1}{2}} \end{aligned}$$

   because the last integral is the integral of a $\text{N}[0, (1 - 2t)^{-1}]$ random variable. It follows that the mgf of Y is $(1 - 2t)^{-\frac{k}{2}}$. Note that this is the mgf of a Gamma random variable with parameters $\lambda = 0.5$ and $\alpha = k/2$. Accordingly,

$$Y \sim \chi^2_k \Longleftrightarrow Y \sim \text{ Gamma} \left(\frac{k}{2}, \frac{1}{2} \right) \text{ and}$$

$$f_Y(y) = \frac{y^{\frac{k}{2}-1}e^{-\frac{1}{2}y}}{\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}} I_{(0,\infty)}(y).$$

3. Properties of χ^2 Random variables

   (a) If $Y \sim \chi^2_k$, then $\mathrm{E}(Y^r) = \dfrac{2^r\Gamma(k/2+r)}{\Gamma(k/2)}$ provided that $k/2 + r > 0$.

   Proof: Use the moment result for Gamma random variables.

   (b) Using $\Gamma(\alpha+1) = \alpha\Gamma(\alpha)$, it is easy to show that $\mathrm{E}(Y) = k$, $\mathrm{E}(Y^2) = k(k+2)$, and $\mathrm{Var}(X) = 2k$.

   (c) $Y \dot{\sim} \mathrm{N}(k, 2k)$ for large k. This is an application of the central limit theorem. A better approximation (again for large k) is $\sqrt{2Y} - \sqrt{2k-1} \dot{\sim} \mathrm{N}(0,1)$.

   (d) If $Y_1, Y_2, \ldots, Y_n$ are independently distributed as $Y_i \sim \chi^2_{k_i}$, then $\sum_{i=1}^n Y_i \sim \chi^2_k$, where $k = \sum_{i=1}^n k_i$.

   Proof: use the mgf.

   (e) If $X \sim \chi^2_k$, $X + Y \sim \chi^2_n$, and $X \perp\!\!\!\perp Y$, then $Y \sim \chi^2_{n-k}$.

   Proof: See page 248 in the text. Note that by independence $\psi_{X+Y}(t) = \psi_X(t)\psi_Y(t)$.

6.5 Distributions for Reliability

1. Definition: Suppose that L is a nonnegative continuous rv. In particular, suppose that L is the lifetime (time to failure) of a component. The reliability function is the probability that the lifetime exceeds x. That is,

$$\text{Reliability Function of } L = R_L(x) \stackrel{\text{def}}{=} P(L > x) = 1 - F_L(x).$$

2. Result: If L is a nonnegative continuous rv whose expectation exists, then

$$\mathrm{E}(L) = \int_0^\infty R_L(x)\,dx = \int_0^\infty \left[1 - F_L(x)\right]\,dx.$$

   Proof: Use integration by parts with $u = R_L(x) \Longrightarrow du = -f_(x)$ and $dv = dx \Longrightarrow v = x$. Making these substitutions,

$$\begin{aligned}\int_0^\infty R_L(u)\,du &= \int_0^\infty u\,dv = uv\Big|_0^\infty - \int_0^\infty v\,du \\ &= x\left[1 - F_L(x)\right]\Big|_0^\infty + \int_0^\infty x f_L(x)\,dx \\ &= \int_0^\infty x f_L(x)\,dx = \mathrm{E}(L)\end{aligned}$$

$$\text{provided that } \lim_{x\to\infty} x\left[1 - F_L(x)\right] = 0.$$

3. Definition: the hazard function is the instantaneous rate of failure at time x, given that the component lifetime is at least x. That is,

$$\text{Hazard Function of } L = h_L(x) \stackrel{\text{def}}{=} \lim_{dx \to 0} \frac{P(x < L < x + dx | L > x)}{dx}$$
$$= \lim_{dx \to 0} \left[\frac{F_L(x + dx) - F_L(x)}{dx} \right] \frac{1}{R_L(x)} = \frac{f_L(x)}{R_L(x)}.$$

4. Result:

$$h_L(x) = -\frac{d}{dx} \ln[R_L(x)] = -\frac{1}{R_L(x)} \times \frac{d}{dx} R_L(x)$$
$$= -\frac{1}{R_L(x)} \times -f_L(x) = \frac{f_L(x)}{R_L(x)}.$$

5. Result: If $R_L(0) = 1$, then

$$R_L(x) = \exp\left\{ -\int_0^x h_L(u)\, du \right\}.$$

Proof:

$$h_L(x) = -\frac{d}{dx} \left\{ \ln\left[R_L(x)\right] - \ln\left[R_L(0)\right] \right\}$$
$$\Longrightarrow -h_L(x) = \frac{d}{dx} \left\{ \ln\left[R_L(u)\right] \Big|_0^x \right\} \Longrightarrow -\int_0^x h_L(u)\, du = \ln\left[R_L(x)\right]$$
$$\Longrightarrow \exp\left\{ -\int_0^x h_L(u)\, du \right\} = R_L(x).$$

6. Result: the hazard function is constant if and only if time to failure has an exponential distribution. Proof: First, suppose that time to failure has an exponential distribution. Then,

$$f_L(x) = \lambda e^{-\lambda x} I_{(0,\infty)}(x) \Longrightarrow R_L(x) = e^{-\lambda x} \Longrightarrow h_L(x) = \frac{\lambda e^{-\lambda x}}{e^{-\lambda x}} = \lambda.$$

Second, suppose that the hazard function is a constant, λ. Then,

$$h_L(x) = \lambda \Longrightarrow R_L(x) = \exp\left\{ -\int_0^x \lambda\, du \right\}$$
$$= e^{-\lambda x} \Longrightarrow f_L(x) = \frac{d}{dx} \left[1 - e^{-\lambda x} \right] = \lambda e^{-\lambda x}.$$

7. Weibull Distribution: Increasing hazard function. The hazard function for the Weibull distribution is

$$h_L(x) = \frac{\alpha x^{\alpha - 1}}{\beta^\alpha},$$

where α and β are positive constants. The corresponding reliability function is

$$R_L(x) = \exp\left\{-\int_0^x h_L(u)\,du\right\} = \exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\},$$

and the pdf is

$$f_L(x) = \frac{d}{dx}F_L(x) = \frac{\alpha x^{\alpha-1}}{\beta^{\alpha}}\exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\} I_{(0,\infty)}(x).$$

8. Gompertz Distribution: exponential hazzard function. The hazzard function for the Gompertz distribution is

$$h_L(x) = \alpha e^{\beta x},$$

where α and β are positive constants. The corresponding reliability function is

$$R_L(x) = \exp\left\{-\frac{\alpha}{\beta}\left[e^{\beta x} - 1\right]\right\},$$

and the pdf is

$$f_L(x) = \frac{d}{dx}F_L(x) = \alpha e^{\beta x}\exp\left\{-\frac{\alpha}{\beta}\left[e^{\beta x} - 1\right]\right\} I_{(0,\infty)}(x).$$

9. Series Combinations: If a system fails whenever any single component fails, then the components are said to be in series. The time to failure of the system is the minimum time to failure of the components. If the failure times of the components are statistically independent, then the reliability function of the system is

$$\begin{aligned} R(x) &= P(\text{system life } > x) = P(\text{all components survive to } x) \\ &= \prod R_i(x), \end{aligned}$$

where $R_i(x)$ is the reliability function of the i^{th} component.

10. Parallel Combinations: If a system fails only if all components fail, then the components are said to be in parallel. The time to failure of the system is the maximum time to failure of the components. If the failure times of the components are statistically independent, then the reliability function of the system is

$$\begin{aligned} R(x) &= P(\text{all components fail by time } x) = 1 - P(\text{no component fails by time } x) \\ &= 1 - \prod F_i(x) = 1 - \prod \left[1 - R_i(x)\right], \end{aligned}$$

where $F_i(x)$ is the cdf of the i^{th} component.

6.6 t, F, and Beta Distributions

1. t distributions: Let Z and X be independently distributed as $Z \sim \mathrm{N}(0,1)$ and $X \sim \chi^2_k$. Then
$$T = \frac{Z}{\sqrt{X/k}}$$
has a central t distribution with k degrees of freedom. The pdf is
$$f_T(t) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}\left(1+\frac{t^2}{k}\right)^{(k+1)/2}}.$$

   - If $k = 1$, then the pdf of T is
   $$f_T(t) = \frac{1}{\pi(1+t^2)}.$$
   which is the pdf of a standard Cauchy random variable.
   - Moments of a t random variable. Suppose that $T \sim t_k$. Then
   $$\begin{aligned} \mathrm{E}(T^r) &= \mathrm{E}\left(\frac{k^{r/2}Z^r}{X^{r/2}}\right) = k^{r/2}\mathrm{E}(Z^r)\mathrm{E}(X^{-r/2}), \text{ where} \\ & Z \sim \mathrm{N}(0,1), \quad X \sim \chi^2_k, \text{ and } Z \perp\!\!\!\perp X. \end{aligned}$$
   Recall that odd moments of Z are zero. Even moments of Z and moments of X are
   $$\mathrm{E}(Z^{2i}) = \frac{(2i)!}{i!2^i} \text{ and } \mathrm{E}(X^a) = \frac{2^a\Gamma(k/2+a)}{\Gamma(k/2)}$$
   provided that $a < k/2$. Therefore, if r is a non-negative integer, then
   $$\mathrm{E}(T^r) = \begin{cases} \text{does not exist} & \text{if } r > k; \\ 0 & \text{if } r \text{ is odd and } r < k; \\ k^{r/2}\dfrac{r!\,\Gamma\left(\frac{k-r}{2}\right)}{\left(\frac{r}{2}\right)!\,2^r\,\Gamma\left(\frac{k}{2}\right)} & \text{if } r \text{ is even and } r < k. \end{cases}$$
   Using the above expression, it is easy to show that $\mathrm{E}(T) = 0$ if $k > 1$ and that $\mathrm{Var}(T) = k/(k-2)$ if $k > 2$.

2. F Distributions: Let U_1 and U_2 be independent χ^2 random variables with degrees of freedom k_1 and k_2, respectively. Then
$$Y = \frac{U_1/k_1}{U_2/k_2}$$

has a central F distribution with k_1 and k_2 degrees of freedom. That is, $Y \sim F_{k_1,k_2}$. The pdf is

$$f_Y(y) = \left(\frac{k_1}{k_2}\right)^{k_1/2} \frac{\Gamma\left(\frac{k_1+k_2}{2}\right) y^{(k_1-2)/2}}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)\left(1+\frac{yk_1}{k_2}\right)^{(k_1+k_2)/2}} I_{(0,\infty)}(y).$$

- If $T \sim t_k$, then $T^2 \sim F_{1,k}$.
- Moments of an F random variable. Suppose that $Y \sim F_{k_1,k_2}$. Then

$$\begin{aligned} \mathrm{E}(Y^r) &= \mathrm{E}\left(\frac{(k_2U_1)^r}{(k_1U_2)^r}\right) = \left(\frac{k_2}{k_1}\right)^r \mathrm{E}(U_1^r)\mathrm{E}(U_2^{-r}), \text{ where} \\ & U_1 \sim \chi^2_{k_1}, \quad U_2 \sim \chi^2_{k_2}, \text{ and } U_1 \perp\!\!\!\perp U_2. \end{aligned}$$

  Using the general expression for the moments of a χ^2 random variable, it can be shown that for any real valued r,

$$\mathrm{E}(Y^r) = \begin{cases} \text{does not exist} & \text{if } r > k_2/2; \\ \left(\frac{k_2}{k_1}\right)^r \frac{\Gamma\left(\frac{k_1}{2}+r\right)\Gamma\left(\frac{k_2}{2}-r\right)}{\Gamma\left(\frac{k_1}{2}\right)\Gamma\left(\frac{k_2}{2}\right)} & \text{if } r < k_2/2. \end{cases}$$

  Using the above expression, it is easy to show that

$$\mathrm{E}(Y) = \frac{k_2}{k_2-2} \text{ if } k > 2 \text{ and that } \mathrm{Var}(Y) = \frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)} \text{ if } k_2 > 4.$$

3. Beta Distributions: Let U_1 and U_2 be independent χ^2 random variables with degrees of freedom k_1 and k_2, respectively. Then

$$Y = \frac{U_1}{U_1+U_2}$$

has a beta distribution with parameters $k_1/2$ and $k_2/2$. That is, $Y \sim$ Beta $\left(\frac{k_1}{2}, \frac{k_2}{2}\right)$. More generally, if $U_1 \sim \mathrm{Gam}(\alpha_1)$, $U_2 \sim \mathrm{Gam}(\alpha_2)$, and $U_1 \perp\!\!\!\perp U_2$, then

$$Y = \frac{U_1}{U_1+U_2} \sim \mathrm{Beta}(\alpha_1, \alpha_2).$$

If $Y \sim \mathrm{Beta}(\alpha_1, \alpha_2)$, then the pdf of Y is

$$f_Y(y) = \frac{y^{\alpha_1-1}(1-y)^{\alpha_2-1}}{\beta(\alpha_1,\alpha_2)} I_{(0,1)}(y),$$

where $\beta(\alpha_1, \alpha_2)$ is the beta function and is defined as

$$\beta(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}.$$

- If $B \sim \text{Beta}(\alpha_1, \alpha_2)$, then

$$\frac{\alpha_2 B}{\alpha_1(1-B)} \sim F_{2\alpha_1, 2\alpha_2}.$$

- If $B \sim \text{Beta}(\alpha_1, \alpha_2)$, where $\alpha_1 = \alpha_2 = 1$, then $B \sim \text{Unif}(0, 1)$.
- If $X \sim \text{Beta}(\alpha_1, \alpha_2)$, then

$$\text{E}(X^r) = \frac{\Gamma(\alpha_1 + r)\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1 + \alpha_2 + r)\Gamma(\alpha_1)} \text{ provided that } \alpha_1 + r > 0.$$

Proof:

$$\begin{aligned}
\text{E}(X^r) &= \int_0^1 \frac{x^r x^{\alpha_1 - 1}(1-x)^{\alpha - 1}}{\beta(\alpha_1, \alpha_2)} dx = \int_0^1 \frac{x^{\alpha_1 + r - 1}(1-x)^{\alpha - 1}}{\beta(\alpha_1, \alpha_2)} dx \\
&= \frac{\beta(\alpha_1 + r, \alpha_2)}{\beta(\alpha_1, \alpha_2)} \int_0^1 \frac{x^{\alpha_1 + r - 1}(1-x)^{\alpha - 1}}{\beta(\alpha_1 + r, \alpha_2)} dx \\
&= \frac{\beta(\alpha_1 + r, \alpha_2)}{\beta(\alpha_1, \alpha_2)} = \frac{\Gamma(\alpha_1 + r)\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1 + \alpha_2 + r)\Gamma(\alpha_1)},
\end{aligned}$$

provided that $\alpha_1 + r > 0$, because the last integral is the integral of the pdf of a random variable with distribution $\text{Beta}(\alpha_1 + r, \alpha_2)$.

- If $F \sim F_{k_1, k_2}$, then

$$\frac{k_1 F}{k_1 F + k_2} \sim \text{Beta}\left(\frac{k_1}{2}, \frac{k_2}{2}\right).$$

Chapter 7

ORGANIZING & DESCRIBING DATA

The topics in this chapter are covered in Stat 216, 217, and 401. Please read this chapter. With a few exceptions, I will not lecture on these topics. Below is a list of terms and methods that you should be familiar with.

7.1 Frequency Distributions

1. Contingency (frequency) tables for categorical random variables, cell, marginal distributions.
2. Bar graph for categorical and for discrete random variables.

7.2 Data on Continuous Variables

1. Stem & Leaf Displays for continuous random variables.
2. Frequency Distributions & Histograms for continuous random variables. Area should be proportional to frequency regardless of whether bin widths are equal or not.
3. Scatter Plots for paired continuous random variables.
4. Statistic: A numerical characteristic of the sample. A statistic is a random variable.

7.3 Order Statistics

1. Order statistics are the ordered sample values. The conventional notation is to denote the i^{th} order statistic as $X_{(i)}$, where $X_{(1)} \leq X_{(2)} \leq X_{(3)} \leq \cdots \leq X_{(n)}$.
2. Sample median: 50^{th} percentile.

3. Quartiles: $Q_1 = 25^{\text{th}}$, $Q_2 = 50^{\text{th}}$, and $Q_3 = 75^{\text{th}}$ percentiles.

4. Interquartile Range: $Q_3 - Q_1$.

5. Range: $X_{(n)} - X_{(1)}$.

6. Midrange: $(X_{(n)} + X_{(1)})/2$.

7. Midhinge: $(Q_1 + Q_3)/2$.

8. Five Number Summary: $X_{(1)}, Q_1, Q_2, Q_3, X_{(n)}$.

9. Quantiles: For a data set of size n, the quantiles are the order statistics $X_{(1)}, \ldots, X_{(n)}$. The quantiles are special cases of percentiles (the book has this backwards). The i^{th} quantile is the $100p_i^{\text{th}}$ percentile, where $p_i = (i - 3/8)/(n + 1/4)$. Note, the percentile is defined so that $p_i \in (0, 1)$ for all i. For large n, $p_i \approx i/n$.

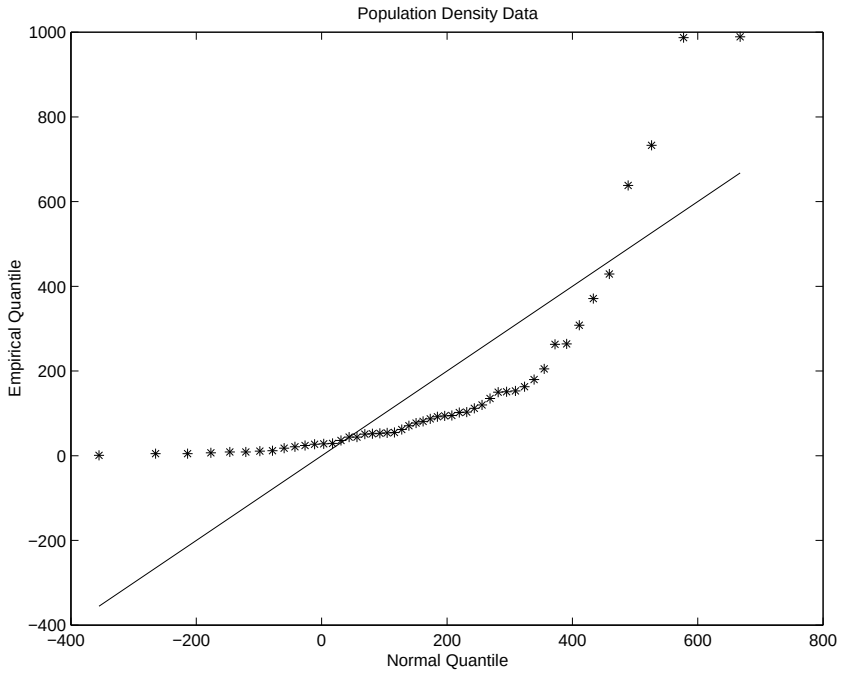
10. Q-Q Plots: These are scatter plots of the quantiles from two distributions. If the distributions are the same, then the scatter plot should show a line of points at a 45 degree angle. One application is to plot the empirical quantiles against the quantiles from a theoretical distribution. This is called a probability plot. Suppose, for example, that it is believed that the data have been sampled from a distribution having cdf F. Then the probability plot is obtained by plotting $F^{-1}(p_i)$ against $X_{(i)}$ for $i = 1, \ldots, n$.

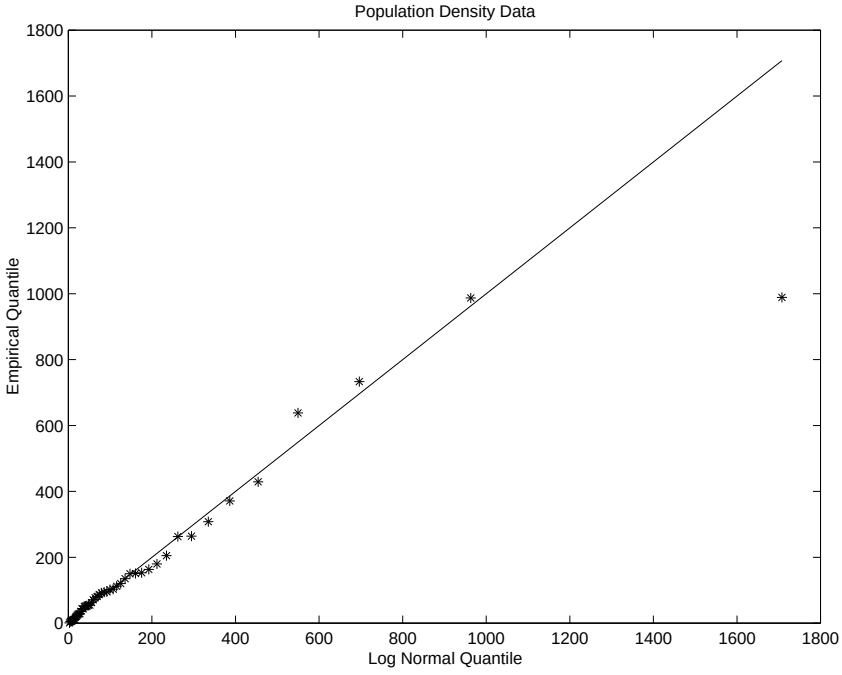
    To visualize whether or not the data could have come from a normal distribution, for example, the empirical quantiles can be plotted against normal quantiles, $\mu + \sigma\Phi^{-1}(p_i)$. For example, problem 7-17 on page 284 gives the population densities per square mile for each of the 50 states. Below are Q-Q plots comparing the quantiles of the data to the quantiles of the normal distribution and to the quantiles of the log normal distribution. The computations to construct the plots are on the following page. In the table, the variable $\ln(y)$ is labeled as w. The quantiles of the normal distribution and the log normal distribution are

$$\bar{y} + s_y z_{p_i} \text{ and } \exp\left\{\bar{w} + s_w z_{p_i}\right\},$$

    respectively, where $z_{p_i} = \Phi^{-1}(p_i)$ is the $100p_i^{\text{th}}$ percentile of the standard normal distribution.

    The smallest three values correspond to Alaska, Montana, and Wyoming. Values 46–50 correspond to Maryland, Connecticut, Massachusetts, New Jersey, and Rhode Island, respectively.


Population Density Data
Empirical Quantile
Normal Quantile
1000
800
600
400
200
0
−200
−400
−400
−200
0
200
400
600
800


Population Density Data
Empirical Quantile
Log Normal Quantile
1800
1600
1400
1200
1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400
1600
1800

i	$y_{(i)}$	p_i	z_{p_i}	$\bar{y} + s_y z_{p_i}$	$\bar{w} + s_w z_{p_i}$	$\exp\{\bar{w} + s_w z_{p_i}\}$
1	1	0.012	−2.24	−355.23	0.95	2.58
2	5	0.032	−1.85	−264.99	1.52	4.58
3	5	0.052	−1.62	−213.93	1.85	6.34
4	7	0.072	−1.46	−176.66	2.08	8.03
5	9	0.092	−1.33	−146.62	2.27	9.72
6	9	0.112	−1.22	−121.08	2.44	11.43
⋮	⋮	⋮	⋮	⋮	⋮	⋮
21	54	0.410	−0.23	104.59	3.87	47.87
22	55	0.430	−0.18	116.19	3.94	51.53
23	62	0.450	−0.13	127.70	4.02	55.43
24	71	0.470	−0.07	139.13	4.09	59.60
25	77	0.490	−0.02	150.51	4.16	64.07
26	81	0.510	0.02	161.89	4.23	68.87
27	87	0.530	0.07	173.27	4.30	74.03
28	92	0.550	0.13	184.70	4.38	79.61
29	94	0.570	0.18	196.21	4.45	85.64
30	95	0.590	0.23	207.81	4.52	92.18
⋮	⋮	⋮	⋮	⋮	⋮	⋮
46	429	0.908	1.33	459.02	6.12	454.22
47	638	0.928	1.46	489.06	6.31	549.64
48	733	0.948	1.62	526.33	6.55	696.36
49	987	0.968	1.85	577.39	6.87	962.96
50	989	0.988	2.24	667.63	7.44	1707.73

7.4 Data Analysis

1. Random variable versus realization: Let $X_1, X_2, \ldots, X_n$ be a random sample from some population. Then X_i is a random variable whose distribution depends on the population at hand. Also, the distribution of $X_1, X_2, \ldots, X_n$ is exchangeable. We will use lower case letters to denote a realization of the random sample. That is, $x_1, x_2, \ldots, x_n$ is a realization of the random sample.

2. Outlier: An observation that is far from the bulk of the data.

3. Random Sample: A simple random sample is a sample taken from the population in a manner such that each possible sample of size n has an equal probability of being selected. Note, this implies that each unit has the same probability of being selected, but a sample taken such that each unit has the same probability of being selected is not necessarily a simple random sample.

4. Transformations of X and/or Y are sometimes useful to change a non-linear relationship into a linear relationship.

7.5 The Sample Mean

1. $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ is a random variable whereas $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ is a realization.

2. $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$ with probability 1 and $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

3. If $X_1, \ldots, X_n$ is a random sample without replacement from a finite population of size N with mean μ and variance σ^2, then

$$\mathrm{E}(\overline{X}) = \mu \text{ and } \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(1 - \frac{(n-1)}{(N-1)}\right).$$

4. If $X_1, \ldots, X_n$ is a random sample with or without replacement from an infinite population or with replacement from a finite population with mean μ and variance σ^2, then

$$\mathrm{E}(\overline{X}) = \mu \text{ and } \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}.$$

7.6 Measures of Dispersion

1. Sample variance: $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ is a random variable whereas $s^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is a realization.

2. If $X_1, \ldots, X_n$ is a random sample with or without replacement from an infinite population or with replacement from a finite population with mean μ_X and variance σ_X^2, then

$$\mathrm{E}(S_X^2) = \sigma_X^2.$$

Proof: First write $(X_i - \overline{X})^2$ as

$$(X_i - \overline{X})^2 = X_i^2 - 2X_i\overline{X} + \overline{X}^2.$$

Accordingly,

$$S_X^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right].$$

Recall that if Y is a random variable with mean μ_Y and variance σ_Y^2, then $\mathrm{E}(Y^2) = \mu_Y^2 + \sigma_Y^2$. In this application, $\mathrm{E}(\overline{X}^2) = \mu_X^2 + \sigma_X^2/n$. Accordingly,

$$\mathrm{E}(S_X^2) = \frac{1}{n-1}\left[n(\mu_X^2 + \sigma_X^2) - n\left(\mu_X^2 + \frac{\sigma_X^2}{n}\right)\right] = \sigma_X^2.$$

3. Let $Y_1, \ldots, Y_n$ be a sample with sample mean $\overline{Y}$ and sample variance S_Y^2. Define X_i by $X_i = a + bY_i$ for $i = 1, \ldots, n$. Then the sample mean and sample variance of $X_1, \ldots, X_n$, are

$$\overline{X} = a + b\overline{Y} \text{ and } S_X^2 = b^2 S_Y^2.$$

Proof:

$$\begin{aligned}\overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}(a + bY_i) \\ &= \frac{1}{n}\left(na + b\sum_{i=1}^{n} Y_i\right) = a + b\overline{Y}.\end{aligned}$$

Also,

$$\begin{aligned}S_X^2 &= \frac{1}{n-1}\sum \left[X_i - \overline{X}\right]^2 = \frac{1}{n-1}\sum \left[a + bY_i - \left(a + b\overline{Y}\right)\right]^2 \\ &= \frac{1}{n-1}\sum \left[bY_i - b\overline{Y}\right]^2 = \frac{1}{n-1}b^2\sum \left[Y_i - \overline{Y}\right]^2 = b^2 S_Y^2.\end{aligned}$$

This result also holds true for realizations $y_1, y_2, \ldots, y_n$.

4. MAD $= n^{-1}\sum_{i=1}^{n} |X_i - \overline{X}|$ or, more commonly, MAD is defined as MAD $= n^{-1}\sum_{i=1}^{n} |X_i - M|$, where M is the sample median.

5. Result: Let $g(a) = \sum_{i=1}^{n} |X_i - a|$. Then, the minimizer of $g(a)$ with respect to a is the sample median.

Proof: The strategy is to take the derivative of $g(a)$ with respect to a; set the derivative to zero; and solve for a. First note that we can ignore any X_i that equals a because it contributes nothing to $g(a)$. If $X_i \neq a$, then

$$\begin{aligned}\frac{d}{da}|X_i - a| &= \frac{d}{da}\sqrt{(X_i - a)^2} \\ &= \frac{1}{2}\left[(X_i - a)^2\right]^{-\frac{1}{2}} 2(X_i - a)(-1) = -\frac{X_i - a}{|X_i - a|} \\ &= \begin{cases} -1 & X_i > a; \\ 1 & X_i < a. \end{cases}\end{aligned}$$

Accordingly,

$$\begin{aligned}\frac{d}{da}g(a) &= \sum_{i=1}^{n}\left[-I_{(-\infty, X_i)}(a) + I_{(X_i, \infty)}(a)\right] \\ &= -\#X\text{s larger than } a + \#X\text{s smaller than } a.\end{aligned}$$

Setting the derivative to zero implies that the number of Xs smaller than a must be equal to the number of Xs larger than a. Thus, a must be the sample median.

7.7 Correlation

1. Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be a random sample of ordered pairs from a population having means (μ_X, μ_Y), variances (σ_X^2, σ_Y^2), and covariance $\sigma_{X,Y}$. The sample covariance between X and Y is

$$S_{X,Y} \stackrel{\text{def}}{=} \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}).$$

2. The equation for $S_{X,Y}$ can be written as

$$S_{X,Y} = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - n\overline{X}\,\overline{Y}\right].$$

Proof: Multiply the X and Y deviations to obtain the following:

$$\begin{aligned} S_{X,Y} &= \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i Y_i - X_i\overline{Y} - \overline{X}Y_i + \overline{X}\,\overline{Y}\right) \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - \overline{Y}\sum_{i=1}^{n} X_i - \overline{X}\sum_{i=1}^{n} Y_i + n\overline{X}\,\overline{Y}\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - n\overline{Y}\,\overline{X} - n\overline{XY} + n\overline{X}\,\overline{Y}\right] \\ &= \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i Y_i - n\overline{Y}\,\overline{X}\right]. \end{aligned}$$

3. If the population is infinite or samples are taken with replacement, then $\mathrm{E}(S_{X,Y}) = \sigma_{X,Y}$.

Proof: First note that $\sigma_{X,Y} = \mathrm{E}(X_i Y_i) - \mu_X\mu_Y$ and, by independence, $\mathrm{E}(X_i Y_j) = \mu_X\mu_Y$ if $i \neq j$. Also

$$\begin{aligned} \overline{X}\,\overline{Y} &= \frac{1}{n^2}\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} Y_j\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} X_i Y_j \\ &= \frac{1}{n^2}\left[\sum_{i=1}^{n} X_i Y_i + \sum_{i\neq j} X_i Y_j\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} \mathrm{E}(S_{X,Y}) &= \frac{1}{n-1}\mathrm{E}\left[\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i\neq j} X_i Y_j\right] \\ &= \frac{1}{n-1}\mathrm{E}\left[\left(1 - \frac{1}{n}\right)\sum_{i=1}^{n} X_i Y_i - \frac{1}{n}\sum_{i\neq j} X_i Y_j\right] \end{aligned}$$

$$\frac{1}{n-1}\left[(n-1)\mathrm{E}(X_iY_i) - (n-1)\mathrm{E}(X_iY_j)\right]$$
$$= \mathrm{E}(X_iY_i) - \mathrm{E}(X_iY_j) = \mathrm{E}(X_iY_i) - \mu_X\mu_Y = \sigma_{X,Y}.$$

4. Sample Correlation Coefficient:

$$r_{X,Y} \stackrel{\text{def}}{=} \frac{S_{X,Y}}{\sqrt{S_X^2 S_Y^2}}.$$

5. If $U_i = a + bX_i$ and $V = c + dY_i$ for $i = 1, \ldots, n$, then the sample covariance between U and V is

$$S_{U,V} = bdS_{X,Y}.$$

Proof: By the definition of sample covariance

$$S_{U,V} = \frac{1}{n-1}\sum_{i=1}^{n}\left(U_i - \overline{U}\right)\left(V_i - \overline{V}\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left[a + bX_i - \left(a + b\overline{X}\right)\right]\left[c + dY_i - \left(c + d\overline{Y}\right)\right]$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(bX_i - b\overline{X}\right)\left(dY_i - d\overline{Y}\right)$$
$$= \frac{1}{n-1}bd\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = bdS_{X,Y}.$$

6. If $U_i = a + bX_i$ and $V = c + dY_i$ for $i = 1, \ldots, n$, then the sample correlation between U and V is

$$r_{U,V} = \mathrm{sign}(bd)\, r_{X,Y}.$$

Proof: By the definition of sample correlation,

$$r_{U,V} = \frac{S_{U,V}}{\sqrt{S_U^2 S_V^2}} = \frac{bdS_{X,Y}}{\sqrt{b^2 S_X^2\, d^2 S_Y^2}}$$
$$= \frac{bd}{|bd|}\frac{S_{X,Y}}{\sqrt{S_X^2\, S_Y^2}}$$
$$= \mathrm{sign}(bd)\, r_{X,Y}.$$

Chapter 8

SAMPLES, STATISTICS, & SAMPLING DISTRIBUTIONS

1. Definition: Parameter—A characteristic of the population.

2. Definition: Statistic—A characteristic of the sample. Specifically, a statistic is a function of the sample;

$$T = g(X_1, X_2, \ldots, X_n) \text{ and } t = g(x_1, x_2, \ldots, x_n).$$

   The function T is a random variable and the function t is a realization of the random variable. For example, $T_1 = \overline{X}$ and $T_2 = S_X^2$ are statistics.

3. Definition: Sampling Distribution—A sampling distribution is the distribution of a statistic. For example, the sampling distribution of $\overline{X}$ is the distribution of $\overline{X}$.

8.1 Random Sampling

1. Some non-random samples

   - Voluntary response sample: the respondent controls whether or not s/he is in the sample.
   - Sample of convenience: the investigator obtains a set of units from the population by using units that are available or can be obtained inexpensively.

2. Random sampling from a finite population

   - Procedure: select units from the population at random, one at a time. Sampling can be done with or without replacement.
   - Properties of random sampling
     - The distribution of the sample is exchangeable

  - All possible samples of size n are equally likely (this is the definition of a simple random sample).
  - Each unit in the population has an equal chance of being selected.
- Definition: Population Distribution—the marginal distribution of X_i, where X_i is the value of the i^{th} unit in the sample. Note, the marginal distribution of all X_is are identical by exchangeability.

3. Random sample of size n

   - In general, a random sample of size n has many possible meanings (e.g., with replacement, without replacement, stratified, etc.).
   - We (the text and lecture) will say "random sample of size n" when we mean a sequence of independent and identically distributed (iid) random variables. This can occur if one randomly samples from a finite population with replacement, or randomly samples from an infinite population. Unless it is qualified, the phrase "random sample of size n" refers to iid random variables and does not refer to sampling without replacement from a finite population.
   - The joint pdf or pmf of a random sample of size n is denoted by

     $$f_{\mathbf{X}}(\mathbf{x}) \stackrel{\text{def}}{=} f_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n),$$

     where $\mathbf{X}$ and $\mathbf{x}$ are vectors of random variables and realizations, respectively. That is

     $$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix} \text{ and } \mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}.$$

     The transpose of a column vector $\mathbf{U}$ is denoted by $\mathbf{U}'$. For example,

     $$\mathbf{X}' = \begin{pmatrix} X_1 & X_2 & \cdots & X_n \end{pmatrix}.$$

   - Using independence,

     $$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i).$$

4. Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from an Expon(λ) distribution. Then the joint pdf of the sample is

   $$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} I_{(0,\infty)}(x_i) = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\} I_{(0,x_{(n)}]}(x_{(1)}) I_{[x_{(1)},\infty)}(x_{(n)}).$$

5. Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from Unif$[a, b]$. Then the joint pdf of the sample is

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} (b-a)^{-1} I_{[a,b]}(x_i) = (b-a)^n I_{[a,x_{(n)}]}(x_{(1)}) I_{[x_{(1)},b]}(x_{(n)}).$$

6. PMF of a random sample taken without replacement from a finite population. Consider a population of size N having $k \leq N$ distinct values. Denote the values as $v_1, v_2, \ldots, v_k$. Suppose that the population contains M_1 units with value v_1, M_2 units with value v_2, etc. Note that $N = \sum_{j=1}^{k} M_j$. Select n units at random without replacement from the population. Let X_i be the value of the i^{th} unit in the sample and denote the $n \times 1$ vector of Xs by $\mathbf{X}$. Let $\mathbf{x}$ be a realization of $\mathbf{X}$. That is, $\mathbf{x}$ is an $n \times 1$ vector whose elements are chosen from $v_1, v_2, \ldots, v_k$. Then the pmf of the sample is

$$f_{\mathbf{X}}(\mathbf{x}) = P(\mathbf{X} = \mathbf{x}) = \frac{\prod_{j=1}^{k} \binom{M_j}{y_j}}{\binom{N}{n} \binom{n}{y_1, y_2, \ldots, y_n}},$$

where y_j is the frequency of v_j in $\mathbf{x}$.

Proof: Let Y_j for $j = 1, \ldots, k$ be the frequency of v_j in $\mathbf{X}$. Note that $\sum_{j=1}^{k} Y_j = n$. Denote the vector of Ys by $\mathbf{Y}$ and the vector of ys by $\mathbf{y}$. Also, denote the number of distinct $\mathbf{x}$ sequences that yield $\mathbf{y}$ by $n_{\mathbf{y}}$. Then

$$f_{\mathbf{Y}}(\mathbf{y}) = \Pr(\mathbf{Y} = \mathbf{y}) = f_{\mathbf{X}}(\mathbf{x}) \times n_{\mathbf{y}},$$

where $f_{\mathbf{X}}(\mathbf{x})$ is the probability of any specific sequence of x's that contains y_1 units with value v_1, y_2 units with value v_2, etc. Multiplication of $f_{\mathbf{X}}(\mathbf{x})$ by $n_{\mathbf{y}}$ is correct because each permutation of $\mathbf{x}$ has the same probability (by exchangeability). Using counting rules from Stat 420, we can show that

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{\prod_{j=1}^{k} \binom{M_j}{y_j}}{\binom{N}{n}} \text{ and } n_{\mathbf{y}} = \binom{n}{y_1, y_2, \ldots, y_n}.$$

Accordingly, the pmf of the sample is

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\prod_{j=1}^{k} \binom{M_j}{y_j}}{\binom{N}{n} \binom{n}{y_1, y_2, \ldots, y_n}}.$$

7. Example: Consider the population consisting of 12 voles, M_j voles of species j for $j = 1, 2, 3$. Suppose that X_1, X_2, X_3, X_4 is a random sample taken without replacement from the population. Furthermore, suppose that

$$\mathbf{x} = \begin{pmatrix} s_3 & s_1 & s_1 & s_2 \end{pmatrix}',$$

where s_j denotes species j. The joint pdf of the sample is

$$f_{\mathbf{X}}(s_3, s_1, s_1, s_2) = \frac{\binom{M_1}{2}\binom{M_2}{1}\binom{M_3}{1}}{\binom{12}{4}\binom{4}{2,1,1}} = \frac{\frac{1}{2}M_1(M_1-1)M_2M_3}{495 \times 12}$$
$$= \frac{M_1(M_1-1)M_2(12-M_1-M_2)}{11{,}880}.$$

8.2 Likelihood

1. Family of probability distributions or models: If the joint pdf or pmf of the sample depends on the value of unknown parameters, then the joint pdf or pmf is written as

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) \text{ where } \boldsymbol{\theta} = \begin{pmatrix} \theta_1 & \theta_2 & \cdots & \theta_k \end{pmatrix}'$$

is a vector of unknown parameters. For example, if $X_1, \ldots, X_n$ is a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$, where μ and σ^2 are unknown, then the joint pdf is

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = f_{\mathbf{X}}(\mathbf{x}|\mu, \sigma^2) = \frac{\exp\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}}{(2\pi\sigma^2)^{n/2}}, \text{ where } \boldsymbol{\theta} = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}.$$

If $\boldsymbol{\theta}$ contains only one parameter, then is will be denoted as θ (i.e., no bold face).

2. Likelihood Function: The likelihood function is a measure of how likely a particular value of $\boldsymbol{\theta}$ is, given that $\mathbf{x}$ has been observed. Caution: the likelihood function is not a probability. The likelihood function is denoted by $L(\boldsymbol{\theta})$ and is obtained by

   - interchanging the roles of $\boldsymbol{\theta}$ and $\mathbf{x}$ in the joint pdf or pmf of $\mathbf{x}$, and
   - dropping all terms that do not depend on $\boldsymbol{\theta}$.

   That is,

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta}|\mathbf{x}) \propto f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}).$$

3. Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from an Expon(λ) distribution. Then the likelihood function is

$$L(\lambda) = \lambda^n \exp\left\{-\lambda \sum_{i=1}^{n} x_i\right\},$$

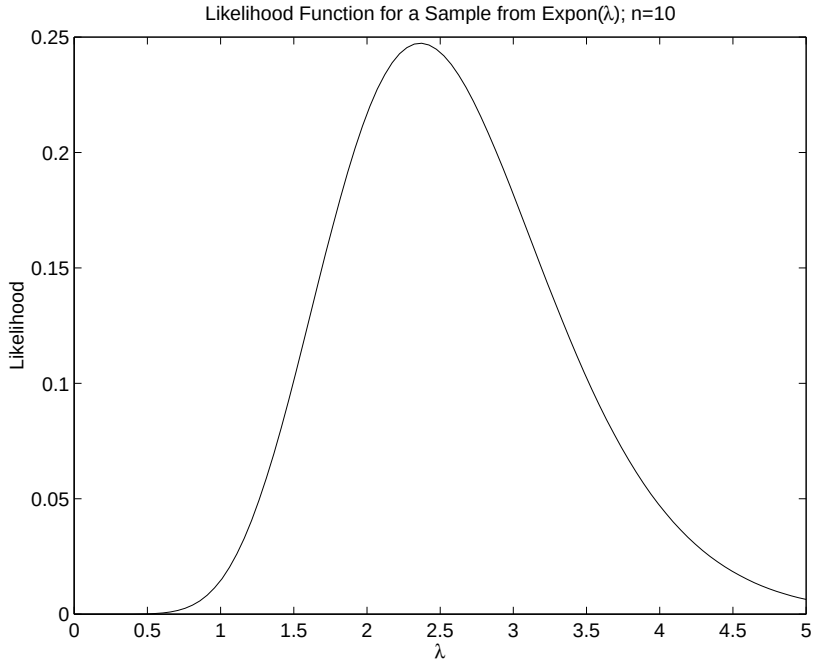
provided that all xs are in $(0, \infty)$. Note that the likelihood function and the joint pdf are identical in this example. Suppose that $n = 10$ and that

$$\begin{aligned}\mathbf{x} &= (0.4393\ 0.5937\ 0.0671\ 2.0995\ 0.1320\ 0.0148\ 0.0050 \\ &\quad 0.1186\ 0.4120\ 0.3483)'\end{aligned}$$

has been observed. The sample mean is $\bar{x} = 4.2303/10 = 0.42303$. The likelihood function is plotted below. Ratios are used to compare likelihoods. For example, the likelihood that $\lambda = 2.5$ is 1.34 times as large as the likelihood that $\lambda = 3$;

$$\frac{L(2.5)}{L(3)} = 1.3390.$$

Note: the x values actually were sampled from Expon(2).



4. Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from Unif$[\pi, b]$. Then the likelihood function is

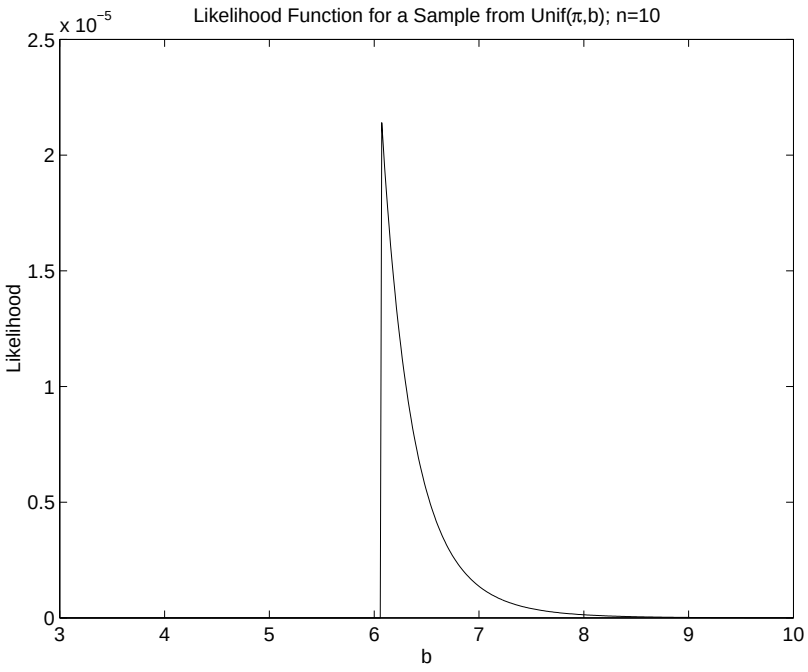
$$L(b) = (b - \pi)^{-n} I_{[x_{(n)}, \infty]}(b),$$

provided that $x_{(1)} > \pi$. Suppose that $n = 10$ and that

$$\mathbf{x} = (5.9841\ 4.9298\ 3.7507\ 5.1264\ 3.8780\ 4.8656\ 6.0682$$

4.1946 5.2010 4.3728)′

has been observed. For this sample, $x_{(1)} = 3.7507$ and $x_{(n)} = 6.0682$. The likelihood function is plotted below. Note, the x values actually were sampled from Unif$(\pi, 2\pi)$.



5. Example: Consider the population consisting consisting of 12 voles, M_j voles of species j for $j = 1, 2, 3$. Suppose that X_1, X_2, X_3, X_4 is a random sample taken without replacement from the population. Furthermore, suppose that

$$\mathbf{x} = \begin{pmatrix} s_3 & s_1 & s_1 & s_2 \end{pmatrix}',$$

where s_j denotes species j. The likelihood function is

$$L(M_1, M_2) = M_1(M_1 - 1)M_2(12 - M_1 - M_2).$$

Note, there are only two parameters, not three, because $M_1 + M_2 + M_3 = 12$. The likelihood function is displayed in the table below. Note: the x values actually were sampled from a population in which $M_1 = 5$, $M_2 = 3$, and $M_3 = 4$.

	Value of M_2										
M_1	0	1	2	3	4	5	6	7	8	9	10
0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0
2	0	18	32	42	48	50	48	42	32	18	0
3	0	48	84	108	120	120	108	84	48	0	0
4	0	84	144	180	192	180	144	84	0	0	0
5	0	120	200	240	240	200	120	0	0	0	0
6	0	150	240	270	240	150	0	0	0	0	0
7	0	168	252	252	168	0	0	0	0	0	0
8	0	168	224	168	0	0	0	0	0	0	0
9	0	144	144	0	0	0	0	0	0	0	0
10	0	90	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0

6. Likelihood Principle:

   All the information which the data provide concerning the relative merits of two hypotheses is contained in the likelihood ratio of those hypotheses on the data (Edwards, 1992).

   Another way of stating the likelihood principal is that if two experiments, each based on a model for θ, give the same likelihood, then the inference about θ should be the same in the two experiments.

7. Example

   (a) Experiment 1: Toss a 35 cent coin n independent times. Let θ be the probability of a head and let X be the number of heads observed. Then X has a binomial pmf:

   $$f_X(x|\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}I_{\{0,1,\ldots,n\}}(x),$$

   where $n = 20$. Suppose that $x = 6$ heads are observed. Then the likelihood function is

   $$L(\theta|x = 6) = \theta^6(1-\theta)^{14}.$$

   (b) Experiment 2: The 35 cent coin was tossed on independent trials until $r = 6$ heads were observed. Let Y be the number of tosses required to obtain 6 heads. Then Y has a negative binomial pmf:

   $$f_Y(y|\theta, r) = \binom{y-1}{r-1}\theta^y(1-\theta)^{y-r}I_{\{r,r+1,\ldots\}}(y),$$

   where $r = 6$. Suppose that the 6^{th} head occurred on the 20^{th} trial. Then, the likelihood function is

   $$L(\theta|y = 20) = \theta^6(1-\theta)^{14}.$$

The likelihood principal requires that any inference about θ be the same from the two experiments.

(c) Suppose that we would like to test $\mathrm{H}_0\colon \theta = 0.5$ against $\mathrm{H}_a\colon \theta < 0.5$. Based on the above two experiments, the p-values are

$$P(X \leq 6|n = 20, \theta = 0.5) = \sum_{x=0}^{6} \binom{20}{x} (1/2)^x (1 - 1/2)^{20-x} = 0.0577$$

in the binomial experiment and

$$P(Y \geq 20|r = 6, \theta = 0.5) = \sum_{y=20}^{\infty} \binom{y-1}{6-1} (1/2)^6 (1 - 1/2)^{y-6} = 0.0318$$

in the negative binomial experiment. If we fail to reject H_0 in the first experiment, but reject H_0 in the second experiment, then we have violated the likelihood principle.

8.3 Sufficient Statistics

1. Definition from the textbook: A statistic, $T = t(\mathbf{X})$, is sufficient for a family of distributions, $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, if and only if the likelihood function depends on $\mathbf{X}$ only through T:
$$L(\boldsymbol{\theta}) = h\left[t(\mathbf{X}), \boldsymbol{\theta}\right].$$

2. Usual definition: A statistic, $T = t(\mathbf{X})$, is sufficient for a family of distributions, $f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta})$, if and only if the conditional distribution of $\mathbf{X}$ given T does not depend on $\boldsymbol{\theta}$:
$$f_{\mathbf{X}|T}(\mathbf{x}|t, \boldsymbol{\theta}) = h(\mathbf{x}).$$
This definition says that after observing T, no additional functions of the data provide information about $\boldsymbol{\theta}$. It can be shown that the two definitions are equivalent.

3. Sample Space and Partitions: The sample space is the set of all possible values of $\mathbf{X}$. It is the same as the support for the joint pdf (or pmf) of $\mathbf{X}$. A statistic partitions the sample space. Each partition corresponds to a different value of of the statistic. A specific partition contains all possible values of $\mathbf{x}$ that yield the specific value of the statistic that indexes the partition. If the statistic is sufficient, then the only characteristic of the data that we need to examine is which partition the sample belongs to.

4. Non-uniqueness of the sufficient statistic: If T is a sufficient statistic, then any one-to-one transformation of T also is sufficient. Note that any transformation of T induces the same partitioning of the sample space. Accordingly, the sufficient statistic is not unique, but the partitioning that corresponds to T is unique.

5. Factorization Criterion (Neyman): A statistic, $T = t(\mathbf{X})$ is sufficient if and only if the joint pdf (pmf) factors as

$$f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) = g\left[t(\mathbf{x})|\boldsymbol{\theta}\right]\, h(\mathbf{x}).$$

   In some cases, $h(\mathbf{x})$ is a trivial function of $\mathbf{x}$. For example, $h(\mathbf{x}) = c$, where c is a constant not depending on $\mathbf{x}$.

6. Example: Bernoulli trials—Let X_i for $i = 1, \ldots, n$ be iid Bern(p) random variables. Note, $\theta = p$. The joint pmf is

$$f_{\mathbf{X}}(\mathbf{x}|p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i) = p^{y}(1-p)^{n-y} \prod_{i=1}^{n} I_{\{0,1\}}(x_i),$$

   where $y = \sum_{i=1}^{n} x_i$. Accordingly, $Y = \sum_{i=1}^{n} X_i$ is sufficient.

   For this example, it is not too hard to verify that the conditional distribution of $\mathbf{X}$ given Y does not depend on p. The conditional distribution of $\mathbf{X}$ given $Y = y$ is

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|Y = y) &= \frac{P(\mathbf{X} = \mathbf{x}) I_{\{y\}}\left(\sum x_i\right)}{P(Y = y)} \\ &= \frac{\displaystyle\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} I_{\{0,1\}}(x_i) I_{\{y\}}\left(\sum x_i\right)}{\displaystyle\binom{n}{y} p^{y}(1-p)^{n-y} I_{\{0,1,2,\ldots,n\}}(y)} \\ &= \frac{\displaystyle\prod_{i=1}^{n} I_{\{0,1\}}(x_i) I_{\{y\}}\left(\sum x_i\right)}{\displaystyle\binom{n}{y} I_{\{0,1,2,\ldots,n\}}(y)} \end{aligned}$$

   which does not depend on p. That is, the conditional distribution of X given a sufficient statistic does not depend on $\boldsymbol{\theta}$.

7. Example: Sampling from Poi(λ). Let $X_1, \ldots, X_n$ be a random sample of size n from Poi(λ). The joint pmf is

$$f_{\mathbf{X}}(\mathbf{x}|\lambda) = \frac{e^{-n\lambda}\lambda^{t}}{\displaystyle\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} I_{\{0,1,\ldots,\infty\}}(x_i), \text{ where } t = \sum_{i=1}^{n} x_i.$$

   Accordingly, the likelihood function is

$$L(\lambda) = e^{-n\lambda}\lambda^{t}$$

and $T = \sum_{i=1}^{n} X_i$ is sufficient. Recall that $T \sim \text{Poi}(n\lambda)$. Therefore, the distribution of $\mathbf{X}$ conditional on $T = t$ is

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}|T = t) &= \frac{P(\mathbf{X} = \mathbf{x}, T = t)}{P(T = t)} \\ &= \frac{e^{-n\lambda}\lambda^t I_{\{t\}}\left(\sum x_i\right) t! \prod_{i=1}^{n} I_{\{0,1,\ldots,\infty\}}(x_i)}{\left(\prod_{i=1}^{n} x_i!\right) e^{-n\lambda}(n\lambda)^t} \\ &= \binom{t}{x_1, x_2, \ldots, x_n}\left(\frac{1}{n}\right)^{x_1}\left(\frac{1}{n}\right)^{x_2}\cdots\left(\frac{1}{n}\right)^{x_n} \\ &\Longrightarrow (\mathbf{X}|T = t) \sim \text{multinom}\left(t, \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right). \end{aligned}$$

Note that the distribution of $\mathbf{X}$, conditional on the sufficient statistic does not depend on λ.

8. Example: Suppose that $X_i \sim$ iid $\text{N}(\mu, 1)$, for $i = 1, \ldots, n$. The joint pdf is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\mu) &= \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}}{(2\pi)^{\frac{n}{2}}} \\ &= \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x} + \bar{x} - \mu)^2\right\}}{(2\pi)^{\frac{n}{2}}} \\ &= \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[(x_i - \bar{x})^2 + 2(x_i - \bar{x})(\bar{x} - \mu) + (\bar{x} - \mu)^2\right]\right\}}{(2\pi)^{\frac{n}{2}}} \\ &= \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right\}}{(2\pi)^{\frac{n}{2}}} \text{ because } \sum_{i=1}^{n}(x_i - \bar{x}) = 0 \\ &= \exp\left\{-\frac{n}{2}(\bar{x} - \mu)^2\right\} \frac{\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}}{(2\pi)^{\frac{n}{2}}}. \end{aligned}$$

Accordingly, the likelihood function is

$$L(\boldsymbol{\theta}) = \exp\left\{-\frac{n}{2}(\bar{x} - \mu)^2\right\},$$

and $\overline{X}$ is sufficient for the family of distributions. This means that $\overline{X}$ contains all of the information about μ that is contained in the data. That is, if we

want to use the sample to learn about μ, we should examine $\overline{X}$ and we need not examine any other function of the data.

9. Order Statistics are sufficient: If $X_1, \ldots, X_n$ is a random sample (with or without replacement), then the order statistics are sufficient.

   Proof: By exchangeability,

   $$f_{\mathbf{X}}(\mathbf{X}|\boldsymbol{\theta}) = f_{\mathbf{X}}(X_{(1)}, X_{(2)}, \ldots, X_{(n)}|\boldsymbol{\theta}).$$

   The likelihood function is proportional to the joint pdf or pmf. Therefore, the likelihood function is a function of the order statistics and, by definition 1, the order statistics are sufficient.

   If the random sample is taken from a continuous distribution, then it can be shown that

   $$P(\mathbf{X} = \mathbf{x}|x_{(1)}, \ldots, x_{(n)}) = \frac{1}{n!}$$

   and this distribution does not depend on $\boldsymbol{\theta}$. Therefore, by definition 2 the order statistics are sufficient.

10. The One Parameter Exponential Family: The random variable X is said to have a distribution within the one parameter regular exponential family if

    $$f_X(x|\theta) = B(\theta)h(x)\exp\{Q(\theta)R(x)\},$$

    where $Q(\theta)$ is a nontrivial continuous function of θ, and $R(x)$ is a nontrivial function of x. Note that if the support of X is represented as an indicator variable, then the indicator variable is part of $h(x)$. That is, the support cannot depend on θ. Either or both of the functions $B(\theta)$ and $h(x)$ could be trivial.

    A random sample of size n from an exponential family has pdf (or pmf)

    $$f_{\mathbf{X}}(\mathbf{x}|\theta) = B(\theta)^n \exp\left\{Q(\theta)\sum_{i=1}^{n} R(x_i)\right\}\prod_{i=1}^{n} h(x_i).$$

    By the factorization criterion, $T = \sum_{i=1}^{n} R(X_i)$ is sufficient for θ.

11. Examples of one parameter exponential families and the corresponding sufficient statistic.

    - Consider a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. Then $T = \sum_{i=1}^{n} X_i$ is sufficient.
    - Consider a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$, where μ is known. Then $T = \sum_{i=1}^{n}(X_i - \mu)^2$ is sufficient.
    - Consider a random sample of size n from $\mathrm{Bern}(p)$. Then $T = \sum_{i=1}^{n} X_i$ is sufficient.

- Consider a random sample of size k from Bin(n, p). Then $T = \sum_{i=1}^{k} Y_i$ is sufficient.
- Consider a random sample of size n from Geom(p). Then $T = \sum_{i=1}^{n} X_i$ is sufficient.
- Consider a random sample of size n from NegBin(r, p), where r is known. Then $T = \sum_{i=1}^{n} X_i$ is sufficient.
- Consider a random sample of size n from Poi(λ). Then $T = \sum_{i=1}^{n} X_i$ is sufficient.
- Consider a random sample of size n from Expon(λ). Then $T = \sum_{i=1}^{n} X_i$ is sufficient.
- Consider a random sample of size n from Gam(α, λ), where α is known. Then $T = \sum_{i=1}^{n} X_i$ is sufficient.
- Consider a random sample of size n from Gam(α, λ), where λ is known. Then $T = \sum_{i=1}^{n} \ln(X_i)$ is sufficient.
- Consider a random sample of size n from Beta(α_1, α_2), where α_1 is known. Then $T = \sum_{i=1}^{n} \ln(1 - X_i)$ is sufficient.
- Consider a random sample of size n from Beta(α_1, α_2), where α_2 is known. Then $T = \sum_{i=1}^{n} \ln(X_i)$ is sufficient.

12. Examples of distributions that do not belong to the exponential family.

    - Consider a random sample of size n from Unif(a, b), where a is known. Then $T = X_{(n)}$ is sufficient by the factorization criterion.
    - Consider a random sample of size n from Unif(a, b), where b is known. Then $T = X_{(1)}$ is sufficient by the factorization criterion.
    - Consider a random sample of size n from Unif(a, b), where neither a nor b is known. Then
    $$\mathbf{T} = \begin{pmatrix} X_{(1)} \\ X_{(n)} \end{pmatrix}$$
    is sufficient by the factorization criterion.
    - Consider a random sample of size n from Unif$(\theta, \theta + 1)$. Then
    $$\mathbf{T} = \begin{pmatrix} X_{(1)} \\ X_{(n)} \end{pmatrix}$$
    is sufficient by the factorization criterion.

13. Example: consider a random sample of size n from N(μ, σ^2), where neither parameter is known. Write $(X_i - \mu)$ as
$$\begin{aligned} (X_i - \mu)^2 &= [(X_i - \overline{X}) + (\overline{X} - \mu)]^2 \\ &= (X_i - \overline{X})^2 + 2(X_i - \overline{X})(\overline{X} - \mu) + (\overline{X} - \mu)^2. \end{aligned}$$

The likelihood function can be written as

$$\begin{aligned} L(\mu,\sigma^2|\mathbf{X}) &= \frac{\exp\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2\right\}}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \\ &= \frac{\exp\left\{\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[(X_i-\overline{X})^2+2(X_i-\overline{X})(\overline{X}-\mu)+(\overline{X}-\mu)^2\right]\right\}}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}} \\ &= \frac{\exp\left\{\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2\right]\right\}}{\left(2\pi\sigma^2\right)^{\frac{n}{2}}}. \end{aligned}$$

By the factorization criterion,

$$\mathbf{T} = \begin{pmatrix} S_X^2 \\ \overline{X} \end{pmatrix}$$

is sufficient.

8.4 Sampling Distributions

Recall that a statistic is a random variable. The distribution of a statistic is called a sampling distribution. This section describes some sampling distributions that can be obtained analytically.

1. Sampling without replacement from a finite population. Consider a finite population consisting of N units, where each unit has one of just k values, $v_1,\ldots,v_k$. Of the N units, M_j have value v_j for $j=1,\ldots,k$. Note that $\sum_{j=1}^{k} M_j = N$. Take a sample of size n, one at a time at random and without replacement. Let X_i be the value of the i^{th} unit in the sample. Also, let Y_j be the number of Xs in the sample that have value v_j. If $\boldsymbol{\theta}' = \begin{pmatrix} M_1 & M_2 & \cdots & M_k \end{pmatrix}$ is the vector of unknown parameters, then the joint pmf of $X_1,\ldots,X_n$ is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) &= \frac{\prod_{j=1}^{k}\binom{M_j}{y_j}}{\binom{N}{n}\binom{n}{y_1,y_2,\ldots,y_n}} \\ &\quad \times I_{\{n\}}(\sum_{j=1}^{k} y_j)\prod_{i=1}^{n} I_{\{v_1,\ldots,v_k\}}(x_i)\prod_{j=1}^{k} I_{\{0,1,\ldots,M_j\}}(y_j). \end{aligned}$$

By the factorization theorem, $\mathbf{Y} = \begin{pmatrix} Y_1, & Y_2 & \cdots & Y_k \end{pmatrix}'$ is a sufficient statistic. The sampling distribution of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) = \frac{\prod_{j=1}^{k} \binom{M_j}{y_j}}{\binom{N}{n}} I_{\{n\}}\left(\sum_{j=1}^{k} y_j\right) \prod_{j=1}^{k} I_{\{0,1,\ldots,M_j\}}(y_j).$$

Note that $\mathbf{T} = \begin{pmatrix} Y_1, & Y_2 & \cdots & Y_{k-1} \end{pmatrix}'$ also is sufficient because $Y_k = N - \sum_{j=1}^{k-1} Y_j$ and therefore $\mathbf{Y}$ is a one-to-one function of $\mathbf{T}$ If $k = 2$, then the sampling distribution simplifies to the hypergeometric distribution.

2. Sampling with replacement from a finite population that has k distinct values or sampling without replacement from an infinite population that has k distinct values. Consider a population for which the proportion of units having value v_j is p_j, for $j = 1, \ldots, k$. Note then $\sum_{j=1}^{k} p_j = 1$. Take a sample of size n, one at a time at random and with replacement if the population is finite. Let X_i be the value of the i^{th} unit in the sample. Also, let Y_j be the number of Xs in the sample that have value v_j. Let $\boldsymbol{\theta}' = \begin{pmatrix} p_1 & p_2 & \cdots & p_k \end{pmatrix}$ be the vector of unknown parameters. The Xs are iid and the joint pmf of the sample is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\boldsymbol{\theta}) &= \prod_{i=1}^{n} \prod_{j=1}^{k} p_j^{I_{\{v_j\}}(x_i)} I_{\{v_1,\ldots,v_k\}}(x_i) \\ &= \prod_{j=1}^{k} p_j^{y_j} I_{\{n\}}\left(\sum_{j=1}^{k} y_j\right) \prod_{j=1}^{k} I_{\{0,1,\ldots,n\}}(y_j). \end{aligned}$$

Accordingly, $\mathbf{Y}' = \begin{pmatrix} Y_1, & Y_2 & \cdots & Y_k \end{pmatrix}$ is a sufficient statistic. The sampling distribution of $\mathbf{Y}$ is multinomial:

$$f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) = \binom{n}{y_1, y_2, \ldots, y_k} \prod_{j=1}^{k} p_j^{y_j} I_{\{n\}}\left(\sum_{j=1}^{k} y_j\right) \prod_{j=1}^{k} I_{\{0,1,\ldots,n\}}(y_j).$$

If $k = 2$, then the sampling distribution simplifies to the binomial distribution.

3. Sampling from a Poisson distribution. Suppose that litter size in coyotes follows a Poisson distribution with parameter λ. Let $X_1, \ldots, X_n$ be a random sample of litter sizes from n dens. The Xs are iid and the joint pmf of the sample is

$$\begin{aligned} P(\mathbf{X} &= \mathbf{x}) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!} I_{\{0,1,\ldots\}}(x_i) \\ &= \frac{e^{-n\lambda}\lambda^{y}}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} I_{\{0,1,\ldots\}}(x_i), \end{aligned}$$

where $y = \sum x_i$. Accordingly, $Y = \sum X_i$ is sufficient. The sampling distribution of Y is Poisson:

$$f_{\mathbf{Y}}(\mathbf{y}|\boldsymbol{\theta}) = \frac{e^{-n\lambda}(n\lambda)^y}{y!} I_{\{0,1,\ldots\}}(y).$$

4. Minimum of exponential random variables. Let $T_i \sim$ iid Expon(λ) for $i = 1, \ldots, n$ and let $T_{(1)}$ be the smallest order statistic. Then the sampling distribution of $T_{(1)}$ is $T_{(1)} \sim \text{Expon}(n\lambda)$. See problem 6-31.

5. Maximum of exponential random variables. As in problem 6-31 Let t_i be the failure time for the i^{th} bus. Suppose that $T_i \sim$ iid Expon(λ) for $i = 1, \ldots, n$ and let $T_{(n)}$ be the largest order statistic. The cdf of $T_{(n)}$ is

$$\begin{aligned} P(T_{(n)} \le t) &= F_{T_{(n)}}(t) = P(\text{all buses fail before time } t) \\ &= \prod_{i=1}^{n} P(T_i < t) \text{ because the failure times are } \perp\!\!\!\perp \\ &= \prod_{i=1}^{n} (1 - e^{-\lambda t}) = (1 - e^{-\lambda t})^n I_{(0,\infty)}(t). \end{aligned}$$

The pdf of $T_{(n)}$ can be found by differentiation:

$$f_{T_{(n)}}(t) = \frac{d}{dt} F_{T_{(n)}}(t) = (1 - e^{-\lambda t})^{n-1} n\lambda e^{-\lambda t} I_{(0,\infty)}(t).$$

6. Maximum of uniform random variables. Suppose that $X_i \sim$ iid Unif$(0, \theta)$. The Xs are iid and the joint pdf is

$$\begin{aligned} f_{\mathbf{X}}(\mathbf{x}|\theta) &= \prod_{i=1}^{n} \frac{1}{\theta} I_{(0,\theta)}(x_i) \\ &= \frac{1}{\theta^n} I_{(0,\theta)}(x_{(n)}) \prod_{i=1}^{n} I_{(0,x_{(n)})}(x_i). \end{aligned}$$

Accordingly, $X_{(n)}$ is sufficient. The cdf of $X_{(n)}$ is

$$\begin{aligned} P(X_{(n)} \le x) &= F_{X_{(n)}}(x) = P(\text{all } X\text{s } \le x) \\ &= \prod_{i=1}^{n} P(X_i < x) \text{ because the } X\text{s are } \perp\!\!\!\perp \\ &= \prod_{i=1}^{n} \frac{x}{\theta} = \left(\frac{x}{\theta}\right)^n I_{(0,\theta)}(x). \end{aligned}$$

The pdf of $X_{(n)}$ can be found by differentiation:

$$f_{X_{(n)}}(x) = \frac{d}{dx} F_{T_{(n)}}(x) = \frac{n x^{n-1}}{\theta^n} I_{(0,\theta)}(x).$$

8.5 Simulating Sampling Distributions

1. How to simulate a sampling distribution

   (a) Choose a population distribution of interest: example Cauchy with $\mu = 100$ and $\sigma = 10$

   (b) Choose a statistic or statistics of interest: example sample median and sample mean

   (c) Choose a sample size: example $n = 25$

   (d) Generate a random sample of size n from the specified distribution. The inverse cdf method is very useful here. For the Cauchy(μ, σ^2) distribution, the cdf is

$$F_X(x|\mu, \sigma) = \frac{\arctan\left(\frac{x-\mu}{\sigma}\right)}{\pi} + \frac{1}{2}.$$

   Accordingly, if $U \sim \text{Unif}(0, 1)$, then

$$X = \tan\left[(U - \frac{1}{2})\pi\right]\sigma + \mu \sim \text{Cauchy}(\mu, \sigma^2).$$

   (e) Compute the statistic or statistics.

   (f) Repeat the previous two steps a large number of times.

   (g) Plot, tabulate, or summarize the resulting distribution of the statistic.

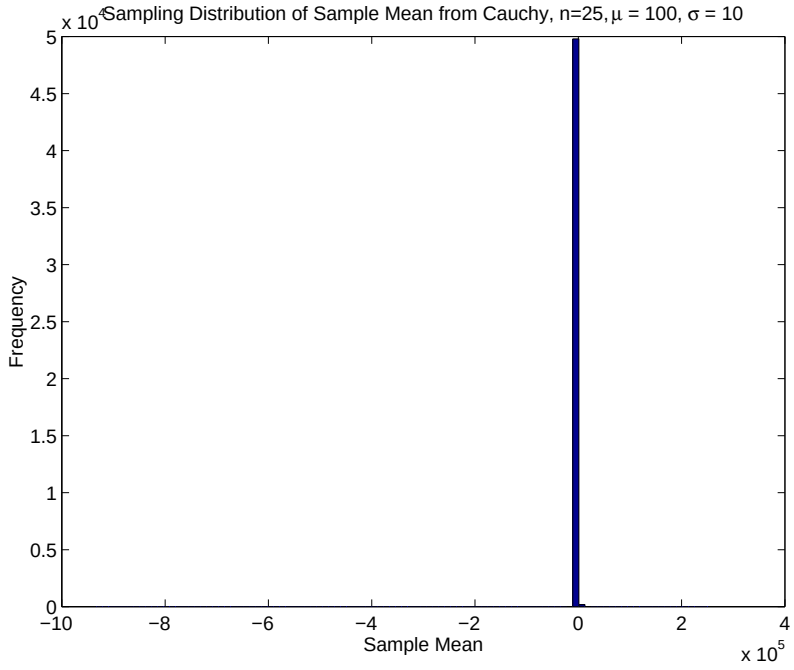
2. Example: Sampling distribution of the mean; $n = 25$, from Cauchy with $\mu = 100$ and $\sigma = 10$;

   (a) Number of samples generated: 50,000

   (b) Mean of the statistic: 85.44

   (c) Standard deviation of the statistic: 4,647.55

   (d) Plot of the statistic.



(e) Most of the distribution is centered near μ, but the tails are very fat. It can be shown that the sample mean also has a Cauchy distribution with $\mu = 100$ and $\sigma = 10$.

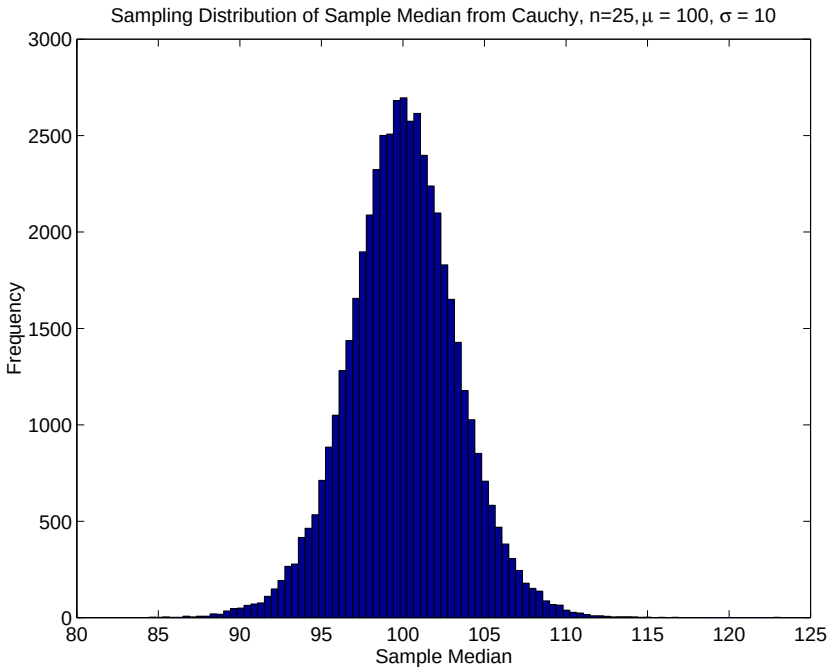
3. Example: Sampling distribution of the median; $n = 25$, from Cauchy with $\mu = 100$ and $\sigma = 10$;

(a) Number of samples generated: 50,000

(b) Mean of the statistic: 100.01

(c) Standard deviation of the statistic: 3.35

(d) Plot of the statistic.



(e) Let M_n be the sample median from a sample of size n from the Cauchy distribution with parameters μ and σ. It can be shown that as n goes to infinity, the distribution of the statistic

$$Z_n = \frac{\sqrt{n}(M_n - \mu)}{\frac{1}{2}\sigma\pi}$$

converges to $\mathrm{N}(0, 1)$. That is, for large n,

$$M_n \dot{\sim} \mathrm{N}\left[\mu, \frac{\sigma^2\pi^2}{4n}\right].$$

Note, for $n = 25$ and $\sigma = 10$, $\mathrm{Var}(M) \approx \pi^2$.

4. To generate normal random variables, the Box-Muller method can be used. see page 44 of these notes.

8.6 Order Statistics

This section examines the distribution of order statistics from continuous distributions.

1. Marginal Distributions of Order Statistics

   (a) Suppose that X_i, $i = 1, \ldots, n$ is a random sample of size n from a population with pdf $f_X(x)$ and cdf $F_X(x)$. Consider $X_{(k)}$, the k^{th} order

statistic. To find the pdf, $f_{X_{(k)}}(x)$, first partition the real line into three pieces:

$$I_1 = (\infty, x], \quad I_2 = (x, x + dx], \text{ and } I_3 = (x + dx, \infty).$$

The pdf of $f_{X_{(k)}}(x)$ is (approximately) the probability of observing $k-1$ Xs in I_1, exactly one X in I_2 and the remaining $n-k$ Xs in I_3. This probability is

$$f_{X_{(k)}}(x) \approx \binom{n}{k-1, 1, n-k} [F_X(x)]^{k-1} [f_X(x)dx]^1 [1 - F_X(x)]^{n-k}.$$

Accordingly (by the differential method), the pdf of $X_{(k)}$ is

$$f_{X_{(k)}}(x) = \binom{n}{k-1, 1, n-k} [F_X(x)]^{k-1} [1 - F_X(x)]^{n-k} f_X(x).$$

(b) Example—Smallest order statistic:

$$\begin{aligned} f_{X_{(1)}}(x) &= \binom{n}{0, 1, n-1} [F_X(x)]^0 [1 - F_X(x)]^{n-1} f_X(x) \\ &= n [1 - F_X(x)]^{n-1} f_X(x). \end{aligned}$$

(c) Example—Largest order statistic:

$$\begin{aligned} f_{X_{(n)}}(x) &= \binom{n}{n-1, 1, 0} [F_X(x)]^{n-1} [1 - F_X(x)]^0 f_X(x) \\ &= n [F_X(x)]^{n-1} f_X(x). \end{aligned}$$

(d) Example—Unif$(0, 1)$ distribution. The cdf is $F_X(x) = x$ and the pdf of the k^{th} order statistic is

$$\begin{aligned} f_{X_{(k)}}(x) &= \binom{n}{k-1, 1, n-k} x^{k-1}(1-x)^{n-k} I_{(0,1)}(x) \\ &= \frac{x^{k-1}(1-x)^{n-k}}{B(k, n-k+1)} I_{(0,1)}(x), \end{aligned}$$

where B is the beta function. That is, $X_{(k)} \sim \text{Beta}(k, n-k+1)$.

(e) Example: Find the exact pdf of the median from an odd size sample. In this case, $k = (n+1)/2$ and the pdf is

$$\begin{aligned} f_{X_{((n+1)/2)}}(x) &= \binom{n}{\frac{n-1}{2}, 1, \frac{n-1}{2}} [F_X(x)]^{(n-1)/2} [1 - F_X(x)]^{(n-1)/2} f_X(x) \\ &= \frac{[F_X(x)]^{(n-1)/2} [1 - F_X(x)]^{(n-1)/2} f_X(x)}{B\left(\frac{n-1}{2}, \frac{n-1}{2}\right)}. \end{aligned}$$

For example, if X has a a Cauchy distribution with parameters μ and σ, then the cdf is

$$F_X(x) = \frac{\arctan\left(\frac{x-\mu}{\sigma}\right)}{\pi} + \frac{1}{2}$$

and the pdf of the median, $M = X_{(\frac{n-1}{2})}$, is

$$\begin{aligned} f_M(m) &= \frac{\left[\frac{\arctan\left(\frac{m-\mu}{\sigma}\right)}{\pi} + \frac{1}{2}\right]^{(n-1)/2} \left[\frac{1}{2} - \frac{\arctan\left(\frac{m-\mu}{\sigma}\right)}{\pi}\right]^{(n-1)/2}}{B\left(\frac{n-1}{2}, \frac{n-1}{2}\right)} \\ &\times \frac{1}{\sigma\pi}\left[1 + \left(\frac{m-\mu}{\sigma}\right)^2\right]^{-1}. \end{aligned}$$

2. Joint Distributions of Order Statistics

(a) Suppose that X_i, $i = 1, \ldots, n$ is a random sample of size n from a population with pdf $f_X(x)$ and cdf $F_X(x)$. Consider $(X_{(k)}, X_{(m)})$ the k^{th} and m^{th} order statistics, where $k < m$. To find the joint pdf $f_{X_{(k)},X_{(m)}}(v, w)$, first partition the real line into five pieces:

$$\begin{aligned} I_1 &= (\infty, v], \quad I_2 = (v, v + dv], \quad I_3 = (v + dv, w], \\ I_4 &= (w, w + dw], \text{ and } I_5 = (w + dw, \infty). \end{aligned}$$

The joint pdf of $f_{X_{(k)},X_{(m)}}(v, w)$ is (approximately) the probability of observing $k-1$ Xs in I_1, exactly one X in I_2, $m-k-1$ Xs in I_3, exactly one X in I_4 and the remaining $n-m$ Xs in I_5. This probability is

$$\begin{aligned} f_{X_{(k)},X_{(m)}}(v, w) &\approx \binom{n}{k-1, 1, m-k-1, 1, n-m} [F_X(v)]^{k-1} \\ &\times [f_X(v)\, dv]^1 [F_X(w) - F_X(v)]^{m-k-1} [f_X(w) dw]^1 \\ &\times [1 - F_X(w)]^{n-m}, \end{aligned}$$

where $v < w$. Accordingly (by the differential method), the joint pdf of $X_{(k)}$ and $X_{(m)}$ is

$$\begin{aligned} f_{X_{(k)},X_{(m)}}(v, w) &= \frac{n!}{(k-1)!(m-k-1)!(n-m)!} [F_X(v)]^{k-1} \\ &\times [F_X(w) - F_X(v)]^{m-k-1} [1 - F_X(w)]^{n-m} \\ &\times f_X(v) f_X(w) I_{(v,\infty)}(w). \end{aligned}$$

(b) Example—joint distribution of smallest and largest order statistic. Let $k = 1$ and $m = n$ to obtain

$$\begin{aligned} f_{X_{(1)},X_{(n)}}(v, w) &= n(n-1) [F_X(w) - F_X(v)]^{n-2} \\ &\times f_X(v) f_X(w) I_{(v,\infty)}(w). \end{aligned}$$

(c) Example—joint distribution of smallest and largest order statistics from Unif$(0,1)$. The cdf is $F_X(x) = x$ and the joint distribution of $X_{(1)}$ and $X_{(n)}$ is

$$f_{X_{(1)},X_{(n)}}(v,w) = n(n-1)(w-v)^{n-2} I_{(v,\infty)}(w).$$

3. Distribution of Sample Range

(a) Let $R = X_{(n)} - X_{(1)}$. The distribution of this random variable is needed to construct R charts in quality control applications and to compute percentiles of Tukey's studentized range statistic (useful when making comparisons among means in ANOVA). To find the pdf of R, we will first find an expression for the cdf of R:

$$\begin{aligned} P(R \le r) &= F_R(r) = P[X_{(n)} - X_{(1)} \le r] = P[X_{(n)} \le r + X_{(1)}] \\ &= P[X_{(1)} \le X_{(n)} \le r + X_{(1)}] \\ &\quad \text{because } X_{(1)} \le X_{(n)} \text{ must be satisfied} \\ &= \int_{-\infty}^{\infty} \int_{v}^{v+r} f_{X_{(1)},X_{(n)}}(v,w)\, dw\, dv. \end{aligned}$$

To obtain $f_R(r)$, take the derivative with respect to r. Leibnitz's rule can be used.

- Leibnitz's Rule: Suppose that $a(\theta)$, $b(\theta)$, and $g(x,\theta)$ are differentiable functions of θ. Then

$$\begin{aligned} \frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} g(x,\theta) dx &= g\left[b(\theta),\theta\right] \frac{d}{d\theta} b(\theta) - g\left[a(\theta),\theta\right] \frac{d}{d\theta} a(\theta) \\ &\quad + \int_{a(\theta)}^{b(\theta)} \frac{d}{d\theta} g(x,\theta) dx. \end{aligned}$$

Accordingly,

$$\begin{aligned} f_R(r) &= \frac{d}{dr} F_R(r) = \frac{d}{dr} \int_{-\infty}^{\infty} \int_{v}^{v+r} f_{X_{(1)},X_{(n)}}(v,w)\, dw\, dv \\ &= \int_{-\infty}^{\infty} \frac{d}{dr} \int_{v}^{v+r} f_{X_{(1)},X_{(n)}}(v,w)\, dw\, dv \\ &= \int_{-\infty}^{\infty} \left[f_{X_{(1)},X_{(n)}}(v,v+r) \frac{d}{dr}(v+r) - f_{X_{(1)},X_{(n)}}(v,v) \frac{d}{dr} v \right] dv \\ &\quad + \int_{-\infty}^{\infty} \int_{v}^{v+r} \frac{d}{dr} f_{X_{(1)},X_{(n)}}(v,w)\, dw\, dv \\ &= \int_{-\infty}^{\infty} f_{X_{(1)},X_{(n)}}(v,v+r)\, dv. \end{aligned}$$

(b) Example—Distribution of sample range from Unif$(0,1)$. In this case, the support for $X_{(1)}, X_{(n)}$ is $0 < v < w < 1$. Accordingly, $f_{X_{(1)},X_{(n)}}(v,v+r)$ is

non-zero only if $0 < v < v + r < 1$. This implies that $0 < v < 1 - r$ and that $r \in (0, 1)$. The pdf of R is

$$\begin{aligned} f_R(r) &= \int_0^{1-r} n(n-1)(v+r-v)^{n-2}\,dv = n(n-1)r^{n-2}(1-r)I_{(0,1)}(r) \\ &= \frac{r^{n-2}(1-r)}{B(n-1,2)}I_{(0,1)}(r). \end{aligned}$$

That is, $R \sim \text{Beta}(n-1, 2)$.

4. Joint distribution of All Order Statistics. Employing the same procedure as for a pair if order statistics, it can be shown that the joint distribution of $X_{(1)}, \ldots, X_{(n)}$ is

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = n! \prod_{i=1}^{n} f_X(x_i) \text{ where } x_1 < x_2 < \cdots < x_n.$$

8.7 Moments of Sample Means and Proportions

Let $X_1, \ldots, X_n$ be a random sample of size n taken either with or without replacement from a population having mean μ_X and variance σ_X^2. Denote the support of the random variable X by $\mathcal{S}_X$. The following definitions are used:

$$\begin{aligned} \overline{X} &= \frac{1}{n}\sum_{i=1}^{n} X_i \\ \hat{p} &= \frac{1}{n}\sum_{i=1}^{n} X_i \text{ if } \mathcal{S}_X = \{0, 1\} \\ S_X^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \text{ and} \\ S_X^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n\hat{p}(1-\hat{p})}{n-1} \text{ if } \mathcal{S}_X = \{0, 1\}. \end{aligned}$$

This section examines the expectation and variance of $\overline{X}$ and $\hat{p}$; the expectation of S_X^2; and unbiased estimators of $\text{Var}(\overline{X})$. The following preliminary results are important and could be asked for on exams:

$$\begin{aligned} \text{E}(X_i) &= \mu_X; && (8.1) \\ \text{Var}(X_i) &= \sigma_X^2; && (8.2) \\ \text{E}(X_i^2) &= \mu_X^2 + \sigma_X^2; && (8.3) \\ \text{Var}(\overline{X}) &= n^{-2}\left[\sum_{i=1}^{n}\text{Var}(X_i) + \sum_{i\neq j}\text{Cov}(X_i, X_j)\right]; && (8.4) \end{aligned}$$

$$\begin{aligned} \mathrm{Cov}(X_i, X_j) &= \begin{cases} 0 & \text{if sampling with replacement,} \\ -\frac{\sigma_X^2}{N-1} & \text{if sampling without replacement; and} \end{cases} \qquad (8.5) \\ \mathrm{E}(\overline{X}^2) &= \mu_{\overline{X}}^2 + \mathrm{Var}(\overline{X}). \qquad (8.6) \end{aligned}$$

The result in equation 8.5 is true because $X_1, \ldots, X_n$ are iid if sampling with replacement and

$$\mathrm{Var}\left(\sum_{i=1}^{N} X_i\right) = 0 = N\,\mathrm{Var}(X_i) + N(N-1)\,\mathrm{Cov}(X_i, X_j)$$

if sampling without replacement. The remaining results in equations 8.1–8.6 follow from exchangeability and from the definition of the variance of a random variable.

Be able to use the preliminary results to prove any of the following results. See pages 63 to 64 of these notes.

1. Case I: Random Sample of size $n \Longrightarrow X_1, X_2, \ldots, X_n$ are iid.
   (a) Case Ia: Random Variable has Arbitrary Support
      - $\mathrm{E}(X_i) = \mu_X$.
      - $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$
      - $\mathrm{Var}(X) = \mathrm{E}(X_i^2) - [\mathrm{E}(X_i)]^2 = \sigma_X^2$.
      - $\mathrm{E}(\overline{X}) = \mu_X$.
      - $\mathrm{Var}(\overline{X}) = \dfrac{\sigma_X^2}{n}$.
      - $\mathrm{E}(S_X^2) = \sigma_X^2$.
      - $\mathrm{E}\left(\dfrac{S_X^2}{n}\right) = \dfrac{\sigma_X^2}{n} = \mathrm{Var}(\overline{X})$.
   (b) Case Ib: Random Variable has Support $\mathcal{S}_X = \{0, 1\}$
      - $\mathrm{E}(X_i) = p$.
      - $\mathrm{Cov}(X_i, X_j) = 0$ for $i \neq j$
      - $\mathrm{Var}(X) = \mathrm{E}(X_i^2) - [\mathrm{E}(X_i)]^2 = \sigma_X^2 = p(1-p)$.
      - $\mathrm{E}(\hat{p}) = p$.
      - $\mathrm{Var}(\hat{p}) = \dfrac{\sigma_X^2}{n} = \dfrac{p(1-p)}{n}$.
      - $\mathrm{E}(S_X^2) = \sigma_X^2 = p(1-p)$. When taking large samples from a binary population, $\sigma_X^2 = p(1-p)$ is usually estimated by $\hat{\sigma}^2 = \hat{p}(1-\hat{p})$ rather than $S_X^2 = \hat{p}(1-\hat{p})\frac{n}{n-1}$. Note that $\hat{\sigma}^2$ has bias $-p(1-p)/n$.
      - $\mathrm{E}\left(\dfrac{S_X^2}{n}\right) = \dfrac{p(1-p)}{n} = \mathrm{Var}(\hat{p})$.
2. Case II: Random Sample of size n without replacement
   (a) Case IIa: Random Variable has Arbitrary Support

- $\mathrm{E}(X_i) = \mu_X$.
- $\mathrm{Cov}(X_i, X_j) = -\dfrac{\sigma_X^2}{N}$ for $i \neq j$
- $\mathrm{Var}(X) = \mathrm{E}(X_i^2) - [\mathrm{E}(X_i)]^2 = \sigma_X^2$.
- $\mathrm{E}(\overline{X}) = \mu_X$.
- $\mathrm{Var}(\overline{X}) = \dfrac{\sigma_X^2}{n}\left(1 - \dfrac{n-1}{N-1}\right)$.
- $\mathrm{E}(S_X^2) = \sigma_X^2 \dfrac{N}{N-1}$.
- $\mathrm{E}\left[\dfrac{S_X^2}{n}\left(1 - \dfrac{n}{N}\right)\right] = \dfrac{\sigma_X^2}{n}\left(1 - \dfrac{n-1}{N-1}\right) = \mathrm{Var}(\overline{X})$.

(b) Case IIb: Random Variable has Support $\mathcal{S}_X = \{0, 1\}$

- $\mathrm{E}(X_i) = p$.
- $\mathrm{Cov}(X_i, X_j) = -\dfrac{\sigma_X^2}{N} = -\dfrac{p(1-p)}{N}$ for $i \neq j$
- $\mathrm{Var}(X) = \mathrm{E}(X_i^2) - [\mathrm{E}(X_i)]^2 = \sigma_X^2 = p(1-p)$.
- $\mathrm{E}(\hat{p}) = p$.
- $\mathrm{Var}(\hat{p}) = \dfrac{\sigma_X^2}{n}\left(1 - \dfrac{n-1}{N-1}\right) = \dfrac{p(1-p)}{n}\left(1 - \dfrac{n-1}{N-1}\right)$.
- $\mathrm{E}(S_X^2) = \mathrm{E}\left(\dfrac{n\hat{p}(1-\hat{p})}{n-1}\right) = \sigma_X^2\left(\dfrac{N}{N-1}\right) = p(1-p)\left(\dfrac{N}{N-1}\right)$.
- $\mathrm{E}\left[\dfrac{S_X^2}{n}\left(1 - \dfrac{n}{N}\right)\right] = \mathrm{E}\left[\dfrac{n\hat{p}(1-\hat{p})}{n(n-1)}\left(1 - \dfrac{n}{N}\right)\right] =$
  $\dfrac{p(1-p)}{n}\left(1 - \dfrac{n-1}{N-1}\right) = \mathrm{Var}(\hat{p})$.

8.8 The Central Limit Theorem (CLT)

Theorem Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with mean μ_X and variance σ_X^2. Then, the distribution of

$$Z_n = \frac{\overline{X} - \mu_X}{\sigma_X / \sqrt{n}}$$

converges to $\mathrm{N}(0, 1)$ as $n \to \infty$.

The importance of the CLT is that the convergence of Z_n to a normal distribution occurs regardless of the shape of the distribution of X. Transforming from Z_n to $\overline{X}$ reveals that

$$\overline{X} \dot{\sim} \mathrm{N}\left(\mu_X, \frac{\sigma_X^2}{n}\right)$$

if n is large.

1. The asymptotic distribution of $\overline{X}$ is said to be $\mathrm{N}(\mu_X, \sigma_X^2/n)$. The limiting distribution of $\overline{X}$ is degenerate $\lim_{n\to\infty} \Pr(\overline{X} = \mu_X) = 1$.

2. Another way to express the CLT is

$$\lim_{n\to\infty} \Pr\left(\frac{\sqrt{n}(\overline{X} - \mu_X)}{\sigma_X} \le c\right) = \Phi(c).$$

   Note, equation (2) on page 341 of the text is not correct. It should be

$$\lim_{n\to\infty} P(\overline{X} \le c) = \begin{cases} 0 & \text{if } c < \mu_X, \\ 1 & \text{if } c \ge \mu_X. \end{cases}$$

3. Application to Sums of iid random variables: If $X_1, X_2, \ldots, X_n$ are iid from a population with mean μ_X and variance σ_X^2, then

$$\begin{aligned} \mathrm{E}\left(\sum_{i=1}^n X_i\right) &= n\mu_X, \\ \mathrm{Var}\left(\sum_{i=1}^n X_i\right) &= n\sigma_X^2, \text{ and} \\ \lim_{n\to\infty} \Pr\left(\frac{\sum_{i=1}^n X_i - n\mu_X}{\sqrt{n}\sigma_X} \le c\right) &= \Phi(c). \end{aligned}$$

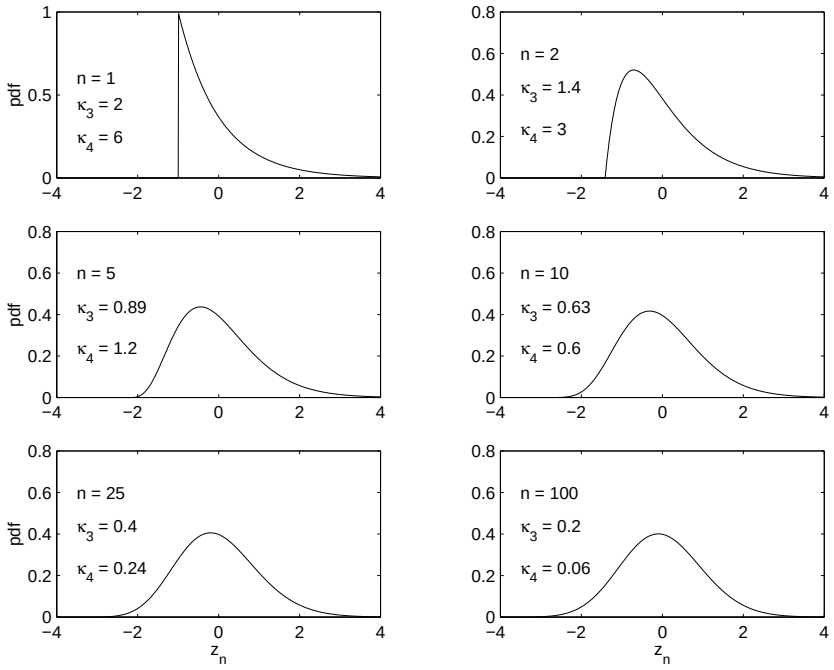
4. How large must n be before $\overline{X}$ is approximately normal? The closer the parent distribution is to a normal distribution, the smaller is the required sample size. When sampling from a normal distribution, a sample size of $n = 1$ is sufficient. Larger sample sizes are required from parent distributions with strong skewness and/or strong kurtosis. For example, suppose that $X \sim \mathrm{Expon}(\lambda)$. This distribution has skewness and kurtosis

$$\kappa_3 = \frac{\mathrm{E}(X - \mu_X)^3}{\sigma_X^{\frac{3}{2}}} = 2 \text{ and } \kappa_4 = \frac{\mathrm{E}(X - \mu_X)^4}{\sigma_X^4} - 3 = 6,$$

   where $\mu_X = 1/\lambda$ and $\sigma_X^2 = 1/\lambda^2$. The sample mean, $\overline{X}$ has distribution $\mathrm{Gam}(n, n\lambda)$. The skewness and kurtosis of $\overline{X}$ are

$$\kappa_3 = \frac{\mathrm{E}(\overline{X} - \mu_{\overline{X}})^3}{\sigma_{\overline{X}}^{\frac{3}{2}}} = \frac{2}{\sqrt{n}} \text{ and } \kappa_4 = \frac{\mathrm{E}(\overline{X} - \mu_{\overline{X}})^4}{\sigma_{\overline{X}}^4} - 3 = \frac{6}{n},$$

   where $\mu_{\overline{X}} = 1/\lambda$ and $\sigma_{\overline{X}}^2 = 1/(n\lambda^2)$. Below are plots of the pdf of Z_n for $n = 1, 2, 5, 10, 25, 100$.



5. Application to Binomial Distribution: Suppose that $X \sim \text{Bin}(n, p)$. Recall that X has the same distribution as the sum of n iid $\text{Bern}(p)$ random variables. Accordingly, for large n

$$\hat{p} \dot{\sim} \text{N}\left(p, \frac{p(1-p)}{n}\right) \text{ and}$$
$$\Pr(\hat{p} \leq c) \approx \Phi\left(\frac{\sqrt{n}(c-p)}{\sqrt{p(1-p)}}\right).$$

6. Continuity Correction. If $X \sim \text{Bin}(n, p)$, then for large n

$$\begin{aligned} X &\dot{\sim} \text{N}\left[np, np(1-p)\right] \text{ and} \\ \Pr(X = x) &= \Pr\left(x - \frac{1}{2} \leq X \leq x + \frac{1}{2}\right) \text{ for } x = 0, 1, \ldots, n \\ &\approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right). \end{aligned}$$

Adding or subtracting 0.5 is called the continuity correction. The continuity corrected normal approximations to the cdfs of X and $\hat{p}$ are

$$\begin{aligned} \Pr(X \leq x) &= \Pr\left(X \leq x + \frac{1}{2}\right) \text{ for } x = 0, 1, \ldots, n; \text{ and} \\ &\approx \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right) \end{aligned}$$

$$\begin{aligned}\Pr(\hat{p} \le c) &= \Pr\left(\hat{p} \le c + \frac{1}{2n}\right) \text{ for } c = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n} \\ &\approx \Phi\left(\frac{\sqrt{n}(c + \frac{1}{2n} - p)}{\sqrt{p(1-p)}}\right).\end{aligned}$$

8.9 Using the Moment Generating Function

1. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. We wish to find the distribution of $\overline{X}$. One approach is to find the mgf of $\overline{X}$ and (hopefully) to identify the corresponding pdf or pmf. Let $\psi_X(t)$ be the mgf of X. The mgf of $\overline{X}$ is

$$\begin{aligned}\psi_{\overline{X}}(t) &= \mathrm{E}\left(\exp\left\{\frac{t}{n}\sum_{i=1}^{n} X_i\right\}\right) \\ &= \mathrm{E}\left(\prod_{i=1}^{n} \exp\left\{\frac{t}{n}X_i\right\}\right) \\ &= \prod_{i=1}^{n} \mathrm{E}\left(\exp\left\{\frac{t}{n}X_i\right\}\right) \text{ by independence} \\ &= \prod_{i=1}^{n} \psi_{X_i}\left(\frac{t}{n}\right) \\ &= \left[\psi_X\left(\frac{t}{n}\right)\right]^n\end{aligned}$$

   because the Xs are identically distributed.

2. Example: Exponential distribution. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from Expon(λ), then

$$\psi_X(t) = \frac{\lambda}{\lambda - t} \text{ and } \psi_{\overline{X}}(t) = \left(\frac{\lambda}{\lambda - \frac{t}{n}}\right)^n = \left(\frac{n\lambda}{n\lambda - t}\right)^n$$

   which is the mgf of Gam($n, n\lambda$).

3. Example: Normal Distribution. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from N(μ_X, σ_X^2), then

$$\begin{aligned}\psi_X(t) &= \exp\left\{t\mu_X + \frac{t^2\sigma_X^2}{2}\right\} \text{ and} \\ \psi_{\overline{X}}(t) &= \left(\exp\left\{\frac{t}{n}\mu_X + \frac{t^2\sigma_X^2}{2n^2}\right\}\right)^n = \exp\left\{t\mu_X + \frac{t^2\sigma_X^2}{2n}\right\}\end{aligned}$$

   which is the mgf of N($\mu_X, \sigma_X^2/n$).

4. Example: Poisson Distribution. If $X_1, X_2, \ldots, X_n$ is a random sample from Poi(λ), then

$$\begin{aligned} \psi_X(t) &= e^{\lambda\left(e^t-1\right)} \text{ and} \\ \psi_Y(t) &= e^{n\lambda\left(e^t-1\right)}, \end{aligned}$$

where $Y = \sum_{i=1}^n X_i = n\overline{X}$. Accordingly, $n\overline{X} \sim \text{Poi}(n\lambda)$ and

$$P(\overline{X} = x) = P(n\overline{X} = nx) = \begin{cases} \dfrac{e^{-n\lambda}\lambda^{nx}}{(nx)!} & \text{for } x = \frac{0}{n}, \frac{1}{n}, \frac{2}{n}, \ldots; \\ 0 & \text{otherwise.} \end{cases}$$

5. A useful limit result. Let a be a constant and let $o\left(n^{-1}\right)$ be a term that goes to zero faster than does n^{-1}. That is,

$$\lim_{n\to\infty} \frac{o\left(n^{-1}\right)}{1/n} = \lim_{n\to\infty} no\left(n^{-1}\right) = 0.$$

Then

$$\lim_{n\to\infty} \left[1 + \frac{a}{n} + o\left(n^{-1}\right)\right]^n = e^a.$$

Proof:

$$\begin{aligned} \lim_{n\to\infty} \left[1 + \frac{a}{n} + o\left(n^{-1}\right)\right]^n &= \lim_{n\to\infty} \exp\left\{n \ln\left[1 + \frac{a}{n} + o\left(n^{-1}\right)\right]\right\} \\ &= \exp\left\{\lim_{n\to\infty} n \ln\left[1 + \frac{a}{n} + o\left(n^{-1}\right)\right]\right\}. \end{aligned}$$

The Taylor series expansion of $\ln(1+\epsilon)$ around $\epsilon = 0$ is

$$\ln(1+\epsilon) = \sum_{i=1}^{\infty} \frac{(-1)^{i+1}\epsilon^i}{i} = \epsilon - \frac{\epsilon^2}{2} + \frac{\epsilon^3}{3} - \frac{\epsilon^4}{4} + \cdots,$$

provided that $|\epsilon| < 1$. Let $\epsilon = a/n + o\left(n^{-1}\right)$. If n is large enough to satisfy $|a/n + o\left(n^{-1}\right)| < 1$, then

$$\begin{aligned} \ln\left[1 + \frac{a}{n} + o\left(n^{-1}\right)\right] &= \frac{a}{n} + o\left(n^{-1}\right) \\ &\quad -\frac{1}{2}\left[\frac{a}{n} + o\left(n^{-1}\right)\right]^2 \\ &\quad +\frac{1}{3}\left[\frac{a}{n} + o\left(n^{-1}\right)\right]^3 - \cdots \\ &= \frac{a}{n} + o\left(n^{-1}\right) \end{aligned}$$

because terms such as a^2/n^2 and $ao\left(n^{-1}\right)/n$ go to zero faster than does $1/n$. Accordingly,

$$\lim_{n\to\infty} \left[1 + \frac{a}{n} + o\left(n^{-1}\right)\right]^n = \exp\left\{\lim_{n\to\infty} n \ln\left[1 + \frac{a}{n} + o\left(n^{-1}\right)\right]\right\}$$

$$\begin{aligned} &= \exp\left\{\lim_{n\to\infty} n\left[\frac{a}{n} + o\left(n^{-1}\right)\right]\right\} \\ &= \exp\left\{\lim_{n\to\infty} a + no\left(n^{-1}\right)\right\} \\ &= \exp\{a+0\} = e^a. \end{aligned}$$

6. Heuristic Proof of CLT using MGF: Write Z_n as

$$\begin{aligned} Z_n &= \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \\ &= \frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu_X}{\sigma_X/\sqrt{n}} \\ &= \frac{\sum_{i=1}^{n} \frac{1}{n}(X_i - \mu_X)}{\sigma_X/\sqrt{n}} \\ &= \sum_{i=1}^{n} \frac{Z_i^*}{\sqrt{n}}, \text{ where } Z_i^* = \frac{X_i - \mu_X}{\sigma_X} \\ &= \sum_{i=1}^{n} U_i, \text{ where } U_i = \frac{Z_i^*}{\sqrt{n}}. \end{aligned}$$

Note that $Z_1, Z_2, \ldots, Z_n$ are iid with $\mathrm{E}(Z_i^*) = 0$ and $\mathrm{Var}(Z_i^*) = 1$. Also, $U_1, U_2, \ldots, U_n$ are iid with $\mathrm{E}(U_i) = 0$ and $\mathrm{Var}(U_i) = 1/n$. If U_i has a moment generating function, then it can be written in expanded form as

$$\begin{aligned} \psi_{U_i}(t) &= \mathrm{E}\left(e^{tU_i}\right) = \sum_{j=0}^{\infty} \frac{t^j}{j!}\mathrm{E}(U_i^j) \\ &= 1 + t\mathrm{E}(U_i) + \frac{t^2}{2}\mathrm{E}(U_i^2) + \frac{t^3}{3!}\mathrm{E}(U_i^3) + \frac{t^4}{4!}\mathrm{E}(U_i^4) + \cdots \\ &= 1 + t\frac{\mathrm{E}(Z_i^*)}{\sqrt{n}} + \frac{t^2}{2}\frac{\mathrm{E}(Z_i^{*2})}{n} + \frac{t^3}{3!}\frac{\mathrm{E}(Z_i^{*3})}{n^{\frac{3}{2}}} + \frac{t^4}{4!}\frac{\mathrm{E}(Z_i^{*4})}{n^2} + \cdots \\ &= 1 + \frac{t^2}{2n} + o\left(n^{-1}\right). \end{aligned}$$

Therefore, the mgf of Z_n is

$$\begin{aligned} \psi_{Z_n}(t) &= \mathrm{E}\left(\exp\{tZ_n\}\right) \\ &= \mathrm{E}\left(\exp\left\{t\sum_{i=1}^{n} U_i\right\}\right) \\ &= [\psi_{U_i}(t)]^n \text{ because } U_1, U_2, \ldots, U_n \text{ are iid} \\ &= \left[1 + \frac{t^2}{2n} + o\left(n^{-1}\right)\right]^n. \end{aligned}$$

Now use the limit result to take the limit of $\psi_{Z_n}(t)$ as n goes to ∞:

$$\lim_{n\to\infty} \psi_{Z_n}(t) = \lim_{n\to\infty} \left[1 + \frac{t^2}{2n} + o\left(n^{-1}\right)\right]^n = \exp\left\{\frac{t^2}{2}\right\}$$

which is the mgf of $\mathrm{N}(0,1)$. Accordingly, the distribution of Z_n converges to $\mathrm{N}(0,1)$ as $n \to \infty$.

8.10 Normal Populations

This section discusses three distributional results concerning normal distributions. Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$. Define, as usual, the sample mean and variance as

$$\overline{X} = n^{-1}\sum_{i=1}^{n} X_i \text{ and } S_X^2 = (n-1)^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Recall, that $\overline{X}$ and S_X^2 are jointly sufficient for μ and σ^2. The three distributional results are the following.

1. $\overline{X} \sim \mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$.
2. $\dfrac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}$.
3. $\overline{X} \perp\!\!\!\perp S_X^2$.

We have already verified result #1. The textbook assumes that result #3 is true and uses results #1 and #3 to prove result #2. The argument relies on another result; one that we already have verified:

$$\sum_{i=1}^{n}(X_i - \mu)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu)^2.$$

Divide both sides by σ^2 to obtain

$$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}. \tag{8.7}$$

Let

$$Z_i = \frac{X_i - \mu}{\sigma} \text{ and let } \overline{Z} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

Note that $Z_i \sim$ iid $\mathrm{N}(0,1)$ and that $\overline{Z} \sim \mathrm{N}(0,1)$. The equality in equation 8.7 can be written as

$$\sum_{i=1}^{n} Z_i^2 = \frac{(n-1)S_X^2}{\sigma^2} + \overline{Z}^2.$$

The left-hand-side above is distributed as χ^2_n and the second term of the right-hand-side is distributed as χ^2_1. If $\overline{X}$ and S^2_X are independently distributed, then the two right-hand-side terms are independently distributed. Using the second result at the top of page 248 in the text (see page 53 of these notes), it can be concluded that $(n-1)S^2_X/\sigma^2 \sim \chi^2_{n-1}$.

We will not attempt to prove result #3. It is an important result that is proven in the graduate linear models course (Stat 505).

8.11 Updating Prior Probabilities Via Likelihood

1. Overview: This section introduces the use of Bayes rule to update probabilities. Let H represent a hypothesis about a numerical parameter θ. In the frequentist tradition, the hypothesis must be either true or false because the value of θ is a fixed number. That is $P(H) = 0$ or $P(H) = 1$.

   In Bayesian analyses, prior beliefs and information are incorporated by conceptualizing θ as a realization of a random variable Θ. In this case, $P(H)$ can take on any value in $[0, 1]$. The quantity, $P(H)$ is called the prior probability. It represents the belief of the investigator prior to collecting new data. One goal of Bayesian analyses is to compute the posterior probability $P(H|\mathbf{X} = \mathbf{x})$, where $\mathbf{X}$ represents new data. By Bayes rule,

$$\begin{aligned} P(H|\mathbf{X} = \mathbf{x}) &= \frac{P(H, \mathbf{X} = \mathbf{x})}{P(\mathbf{X} = \mathbf{x})} \\ &= \frac{P(\mathbf{X} = \mathbf{x}|H)P(H)}{P(\mathbf{X} = \mathbf{x}|H)P(H) + P(\mathbf{X} = \mathbf{x}|H^c)P(H^c)}. \end{aligned}$$

   The quantity $P(\mathbf{X} = \mathbf{x}|H)$ is the likelihood function. The quantity $P(\mathbf{X} = \mathbf{x})$ does not depend on H and therefore is considered a constant (conditioning on $\mathbf{X}$ makes $\mathbf{X}$ a constant rather than a random variable). Accordingly, Bayes rule can be written as

$$P(H|\mathbf{X} = \mathbf{x}) \propto L(H|\mathbf{x})P(H).$$

   That is, the posterior is proportional to the prior times the likelihood function. Note, the functions $P(\mathbf{X} = \mathbf{x}|H)$ and $P(X = \mathbf{x})$ are either pmfs or pdfs depending on whether X is discrete or continuous.

2. Example: The pap smear is a screening test for cervical cancer. The test is not 100% accurate. Let X be the outcome of a pap smear:

$$X = \begin{cases} 0 & \text{if the test is negative, and} \\ 1 & \text{if the test is positive.} \end{cases}$$

   Studies have shown that the false negative rate of the pap smear is approximately 0.1625 and the false positive rate is approximately 0.1864.

That is, 16.25% of women without cervical cancer test positive on the pap smear and 18.64% of women with cervical cancer test negative on the pap smear. Suppose a specific woman, say Gloria, plans to have a pap smear test. Define the random variable (parameter) Θ as

$$\Theta = \begin{cases} 0 & \text{if Gloria does not have cervical cancer, and} \\ 1 & \text{if Gloria does have cervical cancer.} \end{cases}$$

The likelihood function is

$$\begin{aligned} P(X=0|\Theta=1) &= 0.1625; \quad P(X=1|\Theta=1) = 1-0.1625 = 0.8375; \\ P(X=0|\Theta=0) &= 1-0.1864 = 0.8136; \text{ and } P(X=1|\Theta=0) = 0.1864. \end{aligned}$$

Suppose that the prevalence rate of cervical cancer is 31.2 per 100,000 women. A Bayesian might use this information to specify a prior probability for Gloria, namely $P(\Theta=1) = 0.000312$. Suppose that Gloria takes the pap smear test and the test is positive. The posterior probability is

$$\begin{aligned} P(\Theta=1|X=1) &= \frac{P(X=1|\Theta=1)P(\Theta=1)}{P(X=1)} \\ &= \frac{P(X=1|\Theta=1)P(\Theta=1)}{P(X=1|\Theta=1)P(\Theta=1)+P(X=1|\Theta=0)P(\Theta=0))} \\ &= \frac{(0.8375)(0.000312)}{(0.1864)(0.999688)+(0.8375)(0.000312)} = 0.0014. \end{aligned}$$

Note that

$$\frac{P(\Theta=1|X=1)}{P(\Theta=1)} = \frac{0.0014}{0.000312} = 4.488$$

so that Gloria is approximately four and a half times more likely to have cervical cancer given the positive test than she did before the test, even though the probability that she has cervical cancer is still low. The posterior probability, like the prior probability, is interpreted as a subjective probability rather than a relative frequency probability. A relative frequency interpretation makes no sense here because the experiment can not be repeated (there is only one Gloria).

3. Bayes Factor (BF): One way of summarizing the evidence about the hypothesis H is to compute the posterior odds H divided by the prior odds H. This odds ratio is called the Bayes Factor (BF) and it is equivalent to the ratio of likelihood functions. Denote the sufficient statistic by $\mathbf{T}$. In the pap smear example, $T = X$ because there is just one observation. Denote the pdfs or pmfs of $\mathbf{T}$ given H or H^c by $f_{\mathbf{T}|H}(\mathbf{t}|H)$ and $f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)$, respectively. The marginal distribution of $\mathbf{T}$ is obtained by summing the joint distribution of $\mathbf{T}$ and the hypothesis over H and H^c:

$$m_{\mathbf{T}}(\mathbf{t}) = f_{\mathbf{T}|H}(\mathbf{t}|H)P(H) + f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)P(H^c).$$

The Posterior odds of H are

$$\begin{aligned} \text{posterior Odds of } H &= \frac{P(H|\mathbf{T}=\mathbf{t})}{1-P(H|\mathbf{T}=\mathbf{t})} = \frac{P(H|\mathbf{T}=\mathbf{t})}{P(H^c|\mathbf{T}=\mathbf{t})} \\ &= \frac{f_{\mathbf{T}|H}(\mathbf{t}|H)P(H)}{m_{\mathbf{T}}(\mathbf{t})} \div \frac{f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)P(H^c)}{m_{\mathbf{T}}(\mathbf{t})} \\ &= \frac{f_{\mathbf{T}|H}(\mathbf{t}|H)P(H)}{f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)P(H^c)}. \end{aligned}$$

The prior odds of H are

$$\text{Prior Odds of } H = \frac{P(H)}{1-P(H)} = \frac{P(H)}{P(H^c)}.$$

4. Result: The Bayes Factor is equivalent to the ratio of likelihood functions,

$$\text{BF} = \frac{f_{\mathbf{T}|H}(\mathbf{t}|H)}{f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)}.$$

Proof:

$$\begin{aligned} \text{BF} &= \frac{\text{Posterior odds of } H}{\text{Prior odds of } H} = \frac{P(H|\mathbf{T}=\mathbf{t})/P(H^c|\mathbf{T}=\mathbf{t})}{P(H)/P(H^c)} \\ &= \frac{f_{\mathbf{T}|H}(\mathbf{t}|H)P(H)}{f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)P(H^c)} \div \frac{P(H)}{P(H^c)} = \frac{f_{\mathbf{T}|H}(\mathbf{t}|H)}{f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)} \end{aligned}$$

which is the ratio of likelihood functions.

Frequentist statisticians refer to this ratio as the likelihood ratio. A Bayes factor greater than 1 means that the data provide evidence for H relative to H^c and a Bayes factor less than 1 means that the data provide evidence for H^c relative to H, For the cervical cancer example, the hypothesis is $\Theta = 1$ and the Bayes factor is

$$\text{BF} = \frac{P(X=1|\Theta=1)}{P(X=1|\Theta=0)} = \frac{0.8375}{0.1864} = 4.493.$$

The above Bayes factor is nearly the same as the ratio of the posterior probability to the prior probability of H because the prior probability is nearly zero. In general, these ratios will not be equal.

8.12 Some Conjugate Families

1. Overview: Let $X_1, X_2, \ldots, X_n$ be a random sample (with or without replacement) from a population having pdf or pmf $f_X(x|\boldsymbol{\theta})$. A first step in making inferences about $\boldsymbol{\theta}$ is to reduce the data by finding a sufficient

statistic. Let $\mathbf{T}$ be the sufficient statistic and denote the pdf or pmf of $\mathbf{T}$ by $f_{\mathbf{T}|\boldsymbol{\Theta}}(\mathbf{t}|\boldsymbol{\theta})$. Suppose that prior beliefs about $\boldsymbol{\theta}$ can be represented as the prior distribution $g_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. By the definition of conditional probability, the posterior distribution of $\boldsymbol{\Theta}$ is

$$g_{\boldsymbol{\Theta}|\mathbf{T}}(\boldsymbol{\theta}|\mathbf{t}) = \frac{f_{\boldsymbol{\Theta},\mathbf{T}}(\boldsymbol{\theta},\mathbf{t})}{m_{\mathbf{T}}(\mathbf{t})} = \frac{f_{\mathbf{T}|\boldsymbol{\Theta}}(\mathbf{t}|\boldsymbol{\theta})\, g_{\boldsymbol{\Theta}}(\boldsymbol{\theta})}{m_{\mathbf{T}}(\mathbf{t})},$$

where $m_{\mathbf{T}}(\mathbf{t})$ is the marginal distribution of $\mathbf{T}$ which can be obtained as

$$\begin{aligned} m_{\mathbf{T}}(\mathbf{t}) &= \int f_{\mathbf{T},\boldsymbol{\Theta}}(\mathbf{t},\boldsymbol{\theta})\, d\boldsymbol{\theta} \\ &= \int f_{\mathbf{T}|\boldsymbol{\Theta}}(\mathbf{t}|\boldsymbol{\theta}) g_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}. \end{aligned}$$

Integration should be replaced by summation if the prior distribution is discrete.

In practice, obtaining an expression for the marginal distribution of $\mathbf{T}$ may be unnecessary. Note that $m_{\mathbf{T}}(\mathbf{t})$ is just a constant in the posterior distribution. Accordingly, the posterior distribution is

$$g_{\boldsymbol{\Theta}|\mathbf{T}}(\boldsymbol{\theta}|\mathbf{t}) \propto L(\boldsymbol{\theta}|\mathbf{t})\, g_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \text{ because } L(\boldsymbol{\theta}|\mathbf{t}) \propto f_{\mathbf{T}|\boldsymbol{\Theta}}(\mathbf{t}|\boldsymbol{\theta}). \tag{8.8}$$

2. The kernel of a pdf or pmf is proportional to the pmf or pdf and is the part of the function that depends on the random variable. That is, the kernel is obtained by deleting any multiplicative terms that do not depend on the random variable. The right-hand-side of equation (8.8) contains the kernel of the posterior. If the kernel can be recognized, then the posterior distribution can be obtained without first finding the marginal distribution of $\mathbf{T}$.

   The kernels of some well-known distributions are given below.

   (a) If $\Theta \sim \text{Unif}(a,b)$, then the kernel is $I_{(a,b)}(\theta)$.

   (b) If $\Theta \sim \text{Expon}(\lambda)$, then the kernel is $e^{-\lambda\theta} I_{(0,\infty)}(\theta)$.

   (c) If $\Theta \sim \text{Gamma}(\alpha,\lambda)$, then the kernel is $\theta^{\alpha-1} e^{-\lambda\theta} I_{(0,\infty)}(\theta)$.

   (d) If $\Theta \sim \text{Poi}(\lambda)$, then the kernel is $\dfrac{\lambda^\theta}{\theta!} I_{\{0,1,2,\ldots\}}(\theta)$.

   (e) If $\Theta \sim \text{Beta}(\alpha,\beta)$, then the kernel is $\theta^{\alpha-1}(1-\theta)^{\beta-1} I_{(0,1)}(\theta)$.

   (f) If $\Theta \sim \text{N}(\mu,\sigma^2)$, then the kernel is $e^{-\frac{1}{2\sigma^2}(\theta^2 - 2\theta\mu)}$.

3. Conjugate Families: A family of distributions is conjugate for a likelihood function if the prior and posterior distributions both are in the family.

4. Example 1. Consider the problem of making inferences about a population proportion, θ. A random sample $X_1, X_2, \ldots, X_n$ will be obtained from a Bernoulli(θ) population. By sufficiency, the data can be reduced to $Y = \sum X_i$,

and conditional on $\Theta = \theta$, the distribution of Y is $Y \sim \text{Bin}(n, \theta)$ One natural prior distribution for Θ is Beta(α, β). Your textbook gives plots of several beta pdfs on page 356. The lower left plot is not correct. The beta parameters must be greater than zero. The limiting distribution as α and β go to zero is

$$\lim_{\alpha \to 0, \beta \to 0} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} = \begin{cases} \frac{1}{2} & \theta \in \{0,1\} \\ 0 & \text{otherwise.} \end{cases}$$

The parameter $\alpha - 1$ can be conceptualized as the number of prior successes and $\beta - 1$ can be conceptualized as the number of prior failures.

(a) Prior:

$$g_\Theta(\theta|\alpha,\beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} I_{(0,1)}(\theta).$$

(b) Likelihood Function:

$$f_{Y|\Theta}(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} I_{\{0,1,\ldots,n\}}(y) \text{ and}$$
$$L(\theta|y) = \theta^y(1-\theta)^{n-y}.$$

(c) Posterior:

$$g_{\Theta|Y}(\theta|\alpha,\beta,y) \propto \theta^y(1-\theta)^{n-y}\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} I_{(0,1)}(\theta).$$

The kernel of the posterior is $\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}$. Accordingly, the posterior distribution is Beta$(y+\alpha, n-y+\beta)$. Note that the posterior mean (a point estimator) is

$$\text{E}(\Theta|Y=y) = \frac{y+\alpha}{n+\alpha+\beta}.$$

(d) Note that both the prior and posterior are beta distributions. Accordingly, the beta family is conjugate for the binomial likelihood.

5. Example 2. Consider the problem of making inferences about a population mean, θ, when sampling from a normal distribution having known variance, σ^2. By sufficiency, the data can be reduced to $\overline{X}$. One prior for Θ is $\text{N}(\nu, \tau^2)$.

(a) Prior:

$$g_\Theta(\theta|\nu,\tau) = \frac{e^{-\frac{1}{2\tau^2}(\theta-\nu)^2}}{\sqrt{2\pi\tau^2}}.$$

(b) Likelihood Function:

$$f_{\overline{X}|\Theta}(\bar{x}|\theta,\sigma^2) = \frac{\exp\{-\frac{n}{2\sigma^2}(\bar{x}-\theta)^2\}}{\sqrt{2\pi\frac{\sigma^2}{n}}} \text{ and}$$
$$L(\theta|x) = e^{-\frac{n}{2\sigma^2}(\theta^2 - 2\bar{x}\theta)}.$$

(c) Posterior:

$$g_{\Theta|\overline{X}}(\theta|\sigma,\nu,\tau,\bar{x}) \propto \frac{e^{-\frac{1}{2\tau^2}(\theta-\nu)^2}}{\sqrt{2\pi\tau^2}} e^{-\frac{n}{2\sigma^2}(\theta^2-2\bar{x}\theta)}.$$

The combined exponent, after dropping multiplicative terms that do not depend on θ is

$$\begin{aligned}
&-\frac{1}{2}\left\{\frac{n}{\sigma^2}(\theta^2-2\theta\bar{x})+\frac{1}{\tau^2}(\theta^2-2\theta\nu)\right\} \\
&= -\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left\{\theta^2-2\theta\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\bar{x}n}{\sigma^2}+\frac{\nu}{\tau^2}\right)+C\right\} \\
&= -\frac{1}{2}\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)\left\{\theta-\left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\bar{x}n}{\sigma^2}+\frac{\nu}{\tau^2}\right)\right\}^2+C^*,
\end{aligned}$$

where C and C^* are terms that do not depend on θ. Note, we have "completed the square." This is the kernel of a normal distribution with mean and variance

$$\mathrm{E}(\Theta|\bar{x}) = \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}\left(\frac{\bar{x}n}{\sigma^2}+\frac{\nu}{\tau^2}\right) \text{ and } \mathrm{Var}(\Theta|\bar{x}) = \left(\frac{n}{\sigma^2}+\frac{1}{\tau^2}\right)^{-1}$$

(d) Note that both the prior and the posterior are normal distributions. Accordingly, the normal family is conjugate for the normal likelihood when σ^2 is known.

(e) Precision. An alternative expression for the posterior mean and variance uses what is called the precision of a random variable. Precision is defined as the reciprocal of the variance. Thus, as the variance increases, the precision decreases. Your textbook uses the symbol π to stand for precision. For this problem

$$\pi_{\overline{X}} = \frac{n}{\sigma^2}, \quad \pi_\Theta = \frac{1}{\tau^2}, \text{ and } \pi_{\Theta|\overline{X}} = \frac{n}{\sigma^2}+\frac{1}{\tau^2} = \pi_{\overline{X}}+\pi_\Theta.$$

That is, the precision of the posterior is the sum of the precision of the prior plus the precision of the data. Using this notation, the posterior mean and variance are

$$\mathrm{E}(\Theta|\bar{x}) = \frac{\pi_{\overline{X}}}{\pi_{\overline{X}}+\pi_\Theta}\bar{x}+\frac{\pi_\Theta}{\pi_{\overline{X}}+\pi_\Theta}\nu \text{ and } \mathrm{Var}(\Theta|\bar{x}) = (\pi_{\overline{X}}+\pi_\Theta)^{-1}.$$

Note that the posterior mean is a weighted average of the prior mean and that data mean.

8.13 Predictive Distributions

1. The goal in this section is to make predictions about future observations, $Y_1, Y_1, \ldots, Y_k$. We may have current observations $X_1, X_1, \ldots, X_n$ to aid us.

2. Case I: No data available. If the value of $\boldsymbol{\theta}$ is known, then the predictive distribution is simply the pdf (or pmf), $f_{\mathbf{Y}|\boldsymbol{\theta}}(\mathbf{y}|\boldsymbol{\theta})$. In most applications $\boldsymbol{\theta}$ is not known. The Bayesian solution is to integrate $\boldsymbol{\theta}$ out of the joint distribution of $(\boldsymbol{\Theta}, \mathbf{Y})$. That is, the Bayesian predictive distribution is

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}) &= \mathrm{E}_{\boldsymbol{\Theta}}\left[f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})\right] \\ &= \int f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}) g_{\boldsymbol{\Theta}}(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \end{aligned}$$

where $g_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is the prior distribution of $\boldsymbol{\Theta}$. Replace integration by summation if the distribution of $\boldsymbol{\Theta}$ is discrete.

3. Case II: Data available. Suppose that $X_1, X_2, \ldots, X_n$ has been observed from $f_{X|\boldsymbol{\Theta}}(x|\boldsymbol{\theta})$. Denote the sufficient statistic by $\mathbf{T}$ and denote the pdf (pmf) of $\mathbf{T}$ given $\boldsymbol{\theta}$ by $f_{\mathbf{T}|\boldsymbol{\Theta}}(\mathbf{t}|\boldsymbol{\theta})$. The Bayesian posterior predictive distribution is given by

$$\begin{aligned} f_{\mathbf{Y}|\mathbf{T}}(\mathbf{y}|\mathbf{t}) &= \mathrm{E}_{\boldsymbol{\Theta}|\mathbf{T}}\left[f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta})\right] \\ &= \int f_{\mathbf{Y}|\boldsymbol{\Theta}}(\mathbf{y}|\boldsymbol{\theta}) g_{\boldsymbol{\Theta}|\mathbf{T}}(\boldsymbol{\theta}|\mathbf{t})\, d\boldsymbol{\theta}, \end{aligned}$$

where $g_{\boldsymbol{\Theta}|\mathbf{T}}(\boldsymbol{\theta}|\mathbf{t})$ is the posterior distribution of $\boldsymbol{\Theta}$. Replace integration by summation if the distribution of $\boldsymbol{\Theta}$ is discrete. The posterior distribution of $\boldsymbol{\Theta}$ is found by Bayes rule

$$g_{\boldsymbol{\Theta}|\mathbf{T}}(\boldsymbol{\theta}|\mathbf{t}) \propto \mathrm{L}(\boldsymbol{\theta}|\mathbf{t}) g_{\boldsymbol{\Theta}}(\boldsymbol{\theta}).$$

4. Example of case I. Consider the problem of predicting the number of successes in k Bernoulli trials. Thus, conditional on $\Theta = \theta$, the distribution of the sum of the Bernoulli random variables is $Y \sim \mathrm{Bin}(k, \theta)$. The probability of success, θ is not known, but suppose that the prior belief function can be represented by a beta distribution. Then the Bayesian predictive distribution is

$$\begin{aligned} f_Y(y) &= \int_0^1 \binom{k}{y} \theta^y (1-\theta)^{k-y} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} d\theta \\ &= \binom{k}{y} \frac{B(\alpha+y, \beta+k-y)}{B(\alpha,\beta)} I_{\{0,1,\ldots,k\}}(y). \end{aligned}$$

This predictive pmf is known as the beta-binomial pmf. It has expectation

$$\mathrm{E}(Y) = \mathrm{E}_{\Theta}\left[\mathrm{E}(Y|\Theta)\right] = \mathrm{E}_{\Theta}(k\Theta) = k\frac{\alpha}{\alpha+\beta}.$$

For example, suppose that the investigator has no prior knowledge and believes that Θ is equally likely to be anywhere in the $(0, 1)$ interval. Then an

appropriate prior is Beta$(1,1)$, the uniform distribution. The Bayesian predictive distribution is

$$f_Y(y) = \binom{k}{y} \frac{B(1+y, 1+k-y)}{B(1,1)} I_{\{0,1,\ldots,k\}}(y) = \frac{1}{k+1} I_{\{0,1,\ldots,k\}}(y)$$

which is a discrete uniform distribution with support $\{0, 1, \ldots, k\}$. The expectation of Y is $\mathrm{E}(Y) = k/2$.

5. Example of case II. Consider the problem of predicting the number of successes in k Bernoulli trials. Thus, conditional on $\Theta = \theta$, the distribution of the sum of the Bernoulli random variables is $Y \sim \mathrm{Bin}(k, \theta)$. A random sample of size n from Bern(θ) has been obtained. The sufficient statistic is $T = \sum X_i$ and $T \sim \mathrm{Bin}(n, \theta)$. The probability of success, θ is not known, but suppose that the prior belief function can be represented by a beta distribution. Then the posterior distribution of Θ is

$$g_{\Theta|T}(\theta|t) \propto \binom{n}{t} \theta^t (1-\theta)^{n-t} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)}.$$

By recognizing the kernel, it is clear that the posterior distribution of Θ is beta$(t+\alpha, n-t+\beta)$. The Bayesian posterior predictive distribution is

$$\begin{aligned} f_{Y|T}(y|t) &= \int_0^1 \binom{k}{y} \theta^y (1-\theta)^{k-y} \frac{\theta^{t+\alpha-1}(1-\theta)^{n-t+\beta-1}}{B(t+\alpha, n-t+\beta)} d\theta \\ &= \binom{k}{y} \frac{B(\alpha+t+y, \beta+n-t+k-y)}{B(t+\alpha, n-t+\beta)} I_{\{0,1,\ldots,k\}}(y). \end{aligned}$$

This is another beta-binomial pmf. It has expectation

$$\mathrm{E}(Y) = \mathrm{E}_\Theta\left[\mathrm{E}(Y|\Theta)\right] = \mathrm{E}_\Theta(k\Theta) = k\frac{\alpha+t}{n+\alpha+\beta}.$$

For example, suppose that the investigator has no prior knowledge and believes that Θ is equally likely to be anywhere in the $(0,1)$ interval. Then an appropriate prior is Beta$(1,1)$, the uniform distribution. One Bernoulli random variable has been observed and its value is $x = 0$. That is, the data consist of just one failure; $n = 1, t = 0$. The posterior distribution of Θ is Beta$(1,2)$ and the Bayesian posterior predictive distribution is

$$f_{Y|T}(y|t) = \binom{k}{y} \frac{B(1+y, 2+k-y)}{B(1,1)} I_{\{0,1,\ldots,k\}}(y) = \frac{2(k+1+y)}{(k+1)(k+2)} I_{\{0,1,\ldots,k\}}(y).$$

The expectation of Y is $\mathrm{E}(Y) = k/3$.

Chapter 9

ESTIMATION

9.1 Errors in Estimation

1. Estimator versus Estimate: An estimator of a population parameter, say θ, is a function of the data and is a random variable. An estimate is a realization of the random variable.

2. Variance and Bias: Let $T = T(\mathbf{X})$ be an estimator of θ. The bias of T is $b_T = \mathrm{E}(T - \theta) = \mathrm{E}(T) - \theta$. If $b_T = 0$, then T is unbiased for θ. The variance of T is $\sigma_T^2 = \mathrm{E}(T - \mu_T)^2$, where $\mu_T = \mathrm{E}(T)$.

3. Mean Square Error: The mean square error of T is $MSE_T(\theta) = \mathrm{E}(T - \theta)^2 = \mathrm{E}\left[(T - \theta)^2\right]$.

4. Result: $MSE_T(\theta) = \mathrm{Var}(T) + b_T^2$.

   Proof:

$$\begin{aligned} MSE_T(\theta) &= \mathrm{E}[(T - \mu_T) + (\mu_T - \theta)]^2 \\ &= \mathrm{E}[(T - \mu_T)^2 + 2(T - \mu_T)(\mu_T - \theta) + (\mu_T - \theta)^2] \\ &= \mathrm{E}(T - \mu_T)^2 + 2(\mu_T - \theta)\mathrm{E}(T - \mu_T) + (\mu_T - \theta)^2 \\ &= \mathrm{Var}(T) + b_T^2. \end{aligned}$$

5. Root Mean Square Error: $RMSE_T(\theta) = \sqrt{MSE_T(\theta)}$.

6. Example: Sample Variance. Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$. Compare two estimators of σ^2:

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

   We know that S^2 is unbiased for σ^2 (this result does not depend on normality). Therefore

$$b_{S^2} = 0 \text{ and } b_V = \mathrm{E}\left(\frac{n-1}{n}S^2\right) - \sigma^2 = -\frac{\sigma^2}{n}.$$

Recall that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Therefore

$$\text{Var}\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \text{Var}\left(\sigma^2\chi^2_{n-1}\right) = \sigma^4 2(n-1)$$

because $\text{Var}(\chi^2_{n-1}) = 2(n-1)$. The MSEs of S^2 and V are

$$\begin{aligned} MSE_{S^2}(\sigma^2) &= \text{Var}(S^2) = \frac{2(n-1)\sigma^4}{(n-1)^2} = \frac{2\sigma^4}{n-1} \text{ and} \\ MSE_V(\sigma^2) &= \text{Var}(V) + b_V^2 = \frac{2(n-1)\sigma^4}{n^2} + \frac{\sigma^4}{n^2} \\ &= \frac{(2n-1)\sigma^4}{n^2} = \frac{2\sigma^4}{n-1}\left(1 - \frac{3n-1}{2n^2}\right). \end{aligned}$$

Note that $MSE_{S^2}(\sigma^2) > MSE_V(\sigma^2)$ even though S^2 is unbiased and V is biased.

7. Standard Error: An estimator (or estimate) of the standard deviation of an estimator is called the standard error of the estimator.

8. Example of standard errors: In the following table, $f = \dfrac{n-1}{N-1}$.

Parent Distribution	Parameter	Estimator (T)	Var(T)	$SE(T)$
Any, Sampling w replacement	μ	$\overline{X}$	$\dfrac{\sigma^2}{n}$	$\dfrac{S}{\sqrt{n}}$
Bernoulli, Sampling w replacement	p	$\hat{p}$	$\dfrac{p(1-p)}{n}$	$\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n-1}}$
Any finite pop., Sampling w/o replacement	μ	$\overline{X}$	$\dfrac{\sigma^2}{n}(1-f)$	$\sqrt{\dfrac{S^2}{n}(1-f)}$
Finite Bern., Sampling w/o replacement	p	$\hat{p}$	$\dfrac{p(1-p)}{n}(1-f)$	$\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n-1}\left(1-\dfrac{n}{N}\right)}$
Normal	σ^2	S^2	$\dfrac{2\sigma^4}{n-1}$	$S^2\sqrt{\frac{2}{n-1}}$

9.2 Consistency

1. Chebyshev's Inequality: Suppose that X is a random variable with pdf or pmf $f_X(x)$. Let $h(X)$ be a non-negative function of X whose expectation exists and let k be any positive constant. Then

$$P\left[h(X) \geq k\right] \leq \frac{\text{E}\left[h(X)\right]}{k}.$$

Proof: Suppose that X is a continuous rv. Let R be the set $R = \{x; x \in \mathcal{S}_X;\ h(x) \geq k\}$. Then

$$\begin{aligned} \mathrm{E}\left[h(X)\right] &= \int_{\mathcal{S}_X} h(x) f_X(x)\, dx \geq \int_R h(x) f_X(x)\, dx \\ &\geq k \int_R f_X(x)\, dx = kP\left[h(X) \geq k\right] \\ \Longrightarrow \frac{\mathrm{E}\left[h(X)\right]}{k} &\geq P\left[h(X) \geq k\right]. \end{aligned}$$

If X is discrete, then replace integration by summation. A perusal of books in my office reveals that Chebyshev's inequality also is known as

(a) Tchebichev's inequality (Roussas, *Introduction to Probability and Statistical Inference*, 2003, Academic Press),

(b) Tchebysheff's theorem (Mendenhall et al., *A Brief Introduction to Probability and Statistics*, 2002, Duxbury; Freund & Wilson, *Statistical Methods*, 2003, Academic Press),

(c) Tchebychev's inequality (Schervish, *Theory of Statistics*, 1995, Springer-Verlag),

(d) Chebychev's inequality (Casella & Berger, *Statistical Inference*, 2002, Duxbury), and

(e) possibly other variants.

2. Application 1: Suppose that $\mathrm{E}(X) = \mu_X$ and $\mathrm{Var}(X) = \sigma_X^2 < \infty$. Then

$$P\left[\frac{|X - \mu_X|^2}{\sigma_X^2} \geq k^2\right] \leq \frac{1}{k^2}.$$

Proof: Choose $h(X)$ to be

$$h(X) = \frac{(X - \mu_X)^2}{\sigma_X^2}.$$

By the definition of $\mathrm{Var}(X)$, it follows that $\mathrm{E}\left[h(X)\right] = 1$. Also,

$$\begin{aligned} P\left[\frac{|X - \mu_X|}{\sigma_X} \geq k\right] &= P\left[|X - \mu_X| \geq k\sigma_X\right] \\ = P\left[\frac{|X - \mu_X|^2}{\sigma_X^2} \geq k^2\right] &\leq \frac{1}{k^2} \text{ by Chebyshev} \\ \Longrightarrow P\left[|X - \mu_X| < k\sigma_X\right] &\geq 1 - \frac{1}{k^2}. \end{aligned}$$

3. Application 2: Suppose that T is a random variable (estimator of the unknown parameter θ) with $\mathrm{E}(T) = \mu_T$ and $\mathrm{Var}(T) = \sigma_T^2 < \infty$. Then

$$P\left[|X - \theta| < \epsilon\right] \geq 1 - \frac{MSE_X(\theta)}{\epsilon^2},$$

Proof: Choose $h(X)$ to be

$$h(X) = (X - \theta)^2.$$

Then $\mathrm{E}\left[h(T)\right] = MSE_T(\theta)$ and

$$\begin{aligned} P\left[|T - \theta| \geq \epsilon\right] &= P\left[|T - \theta|^2 \geq \epsilon^2\right] \\ &\leq \frac{MSE_T(\theta)}{\epsilon^2} \text{ by Chebyshev} \\ &= \frac{\sigma_T^2 + \left[\mathrm{E}(T) - \theta\right]^2}{\epsilon^2} \\ \Longrightarrow P\left[|T - \theta| < \epsilon\right] &\geq 1 - \frac{MSE_T(\theta)}{\epsilon^2}. \end{aligned}$$

4. Consistency Definition: A sequence of estimators, $\{T_n\}$, is consistent for θ if

$$\lim_{n \to \infty} P\left[|T_n - \theta| < \epsilon\right] = 1$$

for every $\epsilon > 0$.

5. Converge in Probability Definition: A sequence of estimators, $\{T_n\}$, converges in probability to θ if the sequence is consistent for θ. Convergence in probability is usually written as

$$T_n \xrightarrow{\text{prob}} \theta.$$

6. Law of Large Numbers If $\overline{X}$ is the sample mean based on a random sample of size n from a population having mean μ_X, then

$$\overline{X} \xrightarrow{\text{prob}} \mu_X.$$

We will prove the law of large numbers for the special case in which the population variance is finite (see # 9 below). The more general result when the population variance is infinite is sometimes called Khintchine's Theorem.

7. Mean Square Consistent Definition: An estimator of θ is mean square consistent if

$$\lim_{n \to \infty} MSE_{T_n}(\theta) = 0.$$

8. Result: If an estimator is mean square consistent, then it is consistent.

   Proof: Let T_n be an estimator of θ. Assume that T_n has finite mean and variance. Then it follows from Chebyshev's Theorem that

   $$P\left[|T_n - \theta| < \epsilon\right] \geq 1 - \frac{MSE_{T_n}(\theta)}{\epsilon^2},$$

   where ϵ is any positive constant. In T_n, is mean square consistent for θ, then

   $$\lim_{n\to\infty} P\left[|T_n - \theta| < \epsilon\right] \geq \lim_{n\to\infty} 1 - \frac{MSE_{T_n}(\theta)}{\epsilon^2} = 1$$
   $$\text{because } \lim_{n\to\infty} \frac{MSE_{T_n}(\theta)}{\epsilon^2} = 0$$

   for any $\epsilon > 0$.

9. Application: The sample mean based on a random sample of size n from a population with finite mean and variance has mean μ_X and variance σ_X^2/n. Accordingly,

   $$MSE_{\overline{X}}(\mu_X) = \frac{\sigma_X^2}{n} \text{ and } \lim_{n\to\infty} MSE_{\overline{X}}(\mu_X) = 0$$

   which reveals that $\overline{X}$ is mean square consistent. It follows from the result in (8), that $\overline{X} \xrightarrow{\text{prob}} \mu_X$.

9.3 Large Sample Confidence Intervals

1. General setting: Suppose that T_n is an estimator of θ and that

   $$\lim_{n\to\infty} P\left[\frac{T_n - \theta}{\sigma_{T_n}} \leq c\right] = \Phi(c).$$

   That is, $T_n \dot{\sim} \text{N}(\theta, \sigma_{T_n}^2)$ provided that sample size is sufficiently large. Suppose that $\sigma_{T_n}^2 = \omega^2/n$ and that W_n^2 is a consistent estimator of ω^2. That is, $S_{T_n} = SE(T_n) = W_n/\sqrt{n}$ and $W_n^2 \xrightarrow{\text{prob}} \omega^2$. Then, it can be shown that

   $$\lim_{n\to\infty} P\left[\frac{T_n - \theta}{S_{T_n}} \leq c\right] = \Phi(c).$$

   We will not prove the above result. It is an application of Slutsky's theorem which is not covered in Stat 424..

2. Constructing a confidence interval: Denote the $100(1-\alpha/2)$ percentile of the standard normal distribution by $z_{\alpha/2}$. That is

   $$\Phi^{-1}(1 - \alpha/2) = z_{\alpha/2}.$$

Then, using the large sample distribution of T_n, it follows that

$$P\left[-z_{\alpha/2} \leq \frac{T_n - \theta}{S_{T_n}} \leq z_{\alpha/2}\right] \approx 1 - \alpha.$$

Using simple algebra to manipulate the three sides of the above equation yields

$$P\left[T_n - z_{\alpha/2} S_{T_n} \leq \theta \leq T_n + z_{\alpha/2} S_{T_n}\right] \approx 1 - \alpha.$$

The above random interval is a large sample $100(1-\alpha)\%$ confidence interval for θ. The interval is random because T_n and S_{T_n} are random variables.

3. Interpretation of the interval: Let t_n and s_{T_n} be realizations of T_n and S_{T_n}. Then
$$(t_n - z_{\alpha/2} s_{T_n}, t_n + z_{\alpha/2} s_{T_n})$$
is a realization of the random interval. We say that we are $100(1-\alpha)\%$ confident that the realization captures the parameter θ. The $1-\alpha$ probability statement applies to the interval estimator, but not to the interval estimate (i.e., a realization of the interval).

4. Example 1: Confidence interval for μ_X. Let $X_1, \ldots, X_n$ be a random sample of size n from a population having mean μ_X and variance σ_X^2. If sample size is large, then $\overline{X} \dot{\sim} \mathrm{N}\left(\mu_X, \sigma_X^2/n\right)$ by the CLT. The estimated variance of $\overline{X}$ is $S_{\overline{X}}^2 = S_X^2/n$. It can be shown that
$$\mathrm{Var}(S_X^2) = \frac{2\sigma_X^4}{n-1}\left[1 + \frac{(n-1)\kappa_4}{2n}\right],$$
where κ_4 is the standardized kurtosis of the parent distribution. Recall, that if X is normal, then $\kappa_4 = 0$. If κ_4 is finite, then Chebyshev's inequality reveals that $S_X^2 \xrightarrow{\text{prob}} \sigma^2$. It follows that $S_X \xrightarrow{\text{prob}} \sigma$. Accordingly
$$\lim_{n\to\infty} P\left[\frac{\overline{X} - \mu_X}{S_X/\sqrt{n}} \leq c\right] = \Phi(c)$$
and
$$P\left[\overline{X} - z_{\alpha/2}\frac{S_X}{\sqrt{n}} \leq \mu_X \leq \overline{X} + z_{\alpha/2}\frac{S_X}{\sqrt{n}}\right] \approx 1 - \alpha.$$

5. Example 2: Confidence interval for a population proportion, p. Let $X_1, \ldots, X_n$ be a random sample of size n from Bern(p). If sample size is large, then $\hat{p} \dot{\sim} \mathrm{N}\left(p, p(1-p)/n\right)$ by the CLT. The usual estimator of σ_X^2 is $V_X = \hat{p}(1-\hat{p})$. We know that $\hat{p} \xrightarrow{\text{prob}} p$ (by the law of large numbers). It follows that $\hat{p}(1-\hat{p}) \xrightarrow{\text{prob}} p(1-p)$ and, therefore, $V_X = \hat{p}(1-\hat{p})$ is consistent for σ_X^2. Accordingly
$$\lim_{n\to\infty} P\left[\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \leq c\right] = \Phi(c)$$

and

$$P\left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] \approx 1-\alpha.$$

9.4 Determining Sample Size

1. Margin of Error Suppose that, for large n, $T_n \dot{\sim} \mathrm{N}(\theta, \omega^2/n)$ and that ω^2 is known. The large sample $100(1-\alpha)\%$ confidence interval for θ is

$$T_n \pm M_\alpha, \text{ where } M_\alpha = z_{\alpha/2}\frac{\omega}{\sqrt{n}}.$$

   The quantity M_α is $\frac{1}{2}$ of the confidence interval width and is called the margin of error.

2. Choosing n: Suppose that the investigator would like to estimate θ to within $\pm m$ with confidence $100(1-\alpha)\%$. The required sample size is obtained by equating m to M_α and solving for n. The solution is

$$n = \left(\frac{z_{\alpha/2}\omega}{m}\right)^2.$$

   If the solution is not an integer, then round up.

3. Application 1: Let $X_1, \ldots, X_n$ be a random sample from a distribution with unknown mean μ_X and known variance σ_X^2. If n is large, then $\overline{X} \dot{\sim} \mathrm{N}(\mu_X, \sigma_X^2/n)$ by the CLT. To estimate μ_X to within $\pm m$ with $100(1-\alpha)\%$ confidence, use sample size

$$n = \left(\frac{z_{\alpha/2}\sigma_X}{m}\right)^2.$$

4. Application 2: Let $X_1, \ldots, X_n$ be a random sample from a distribution with unknown mean μ_X and unknown variance σ_X^2. If n is large, then $\overline{X} \dot{\sim} \mathrm{N}(\mu_X, \sigma_X^2/n)$. The investigator desires to estimate μ_X to within $\pm m$ with $100(1-\alpha)\%$ confidence. To make any progress, something must be known about σ_X. If the likely range of the data is known, then a rough estimate of σ_X is the range divided by four. Another approach is to begin data collection and then use S_X to estimate σ_X after obtaining several observations. The sample size formula can be used to estimate the number of additional observations that must be taken. The sample size estimate can be updated after collecting more data are re-estimating σ_X.

5. Application 3: Let $X_1, \ldots, X_n$ be a random sample from a Bern(p) distribution. If n is large, then $\hat{p} \dot{\sim} \mathrm{N}(p, p(1-p)/n)$ by the CLT. To estimate p to within $\pm m$ with $100(1-\alpha)\%$ confidence, it would appear that we should use sample size

$$n = \left(\frac{z_{\alpha/2}\sqrt{p(1-p)}}{m}\right)^2.$$

The right-hand-side above, however, cannot be computed because p is unknown and, therefore, $p(1-p)$ also is unknown. Note that $p(1-p)$ is a quadratic function that varies between 0 (when $p = 0$ or $p = 1$) and 0.25 (when $p = 0.5$). A conservative approach is to use $p(1-p) = 0.25$ in the sample size formula. This approach ensures that the computed sample size is sufficiently large, but in most cases it will be larger than necessary. If it is known that $p > p_0$ or that $p < p_0$, then p_0 may be substituted in the sample size formula.

9.5 Small Sample Confidence Intervals for μ_X

1. Setting: Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu_X, \sigma_X^2)$, where neither the mean nor the variance is known. It is desired to construct a $100(1-\alpha)\%$ confidence interval for μ_X. If n is small, then the large sample procedure will not work well because S_X may differ substantially from σ_X.

2. Solution: Consider the random variable

$$\begin{aligned} T &= \frac{\overline{X} - \mu_X}{S_X/\sqrt{n}} = \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \times \frac{\sigma}{S_X} \\ &= \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \div \left(\frac{(n-1)S_X^2}{\sigma^2(n-1)} \right)^{\frac{1}{2}}. \end{aligned}$$

   Recall that

$$\frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \sim \mathrm{N}(0,1), \quad \frac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}, \text{ and}$$
$$\frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \perp\!\!\!\perp \frac{(n-1)S_X^2}{\sigma^2}.$$

   The independence result follows from $\overline{X} \perp\!\!\!\perp S_X^2$. Accordingly, the random variable T has the same distribution as the ratio $Z/\sqrt{W/(n-1)}$ where $Z \sim \mathrm{N}(0,1)$ and $W \sim \chi^2_{n-1}$. This quantity has a t distribution with $n-1$ degrees of freedom.

3. Solution to the problem. Let $t_{\alpha/2,n-1}$ be the $100(1-\alpha/2)$ percentile of the t_{n-1} distribution. That is $F_T^{-1}(1-\alpha/2) = t_{\alpha/2,n-1}$, where $F_T(\,\cdot\,)$ is the cdf of T. Then, using the symmetry of the t distribution around 0, it follows that

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}} \leq t_{\alpha/2,n-1} \right) = 1 - \alpha.$$

   Algebraic manipulation reveals that

$$P\left[\overline{X} - t_{\alpha/2,n-1}\frac{S_X}{\sqrt{n}} \leq \mu_X \leq \overline{X} + t_{\alpha/2,n-1}\frac{S_X}{\sqrt{n}} \right] = 1 - \alpha.$$

Accordingly, an exact $100(1-\alpha)\%$ confidence interval for μ_X is

$$\overline{X} \pm t_{\alpha/2,n-1}\frac{S_X}{\sqrt{n}}.$$

4. Caution: The above confidence interval is correct if one is sampling from a normal distribution. If sample size is small and skewness or kurtosis is large, then the true confidence can differ substantially from $100(1-\alpha)\%$.

9.6 The Distribution of T

Recall that if $Z \sim \mathrm{N}(0,1)$, $Y \sim \chi^2_k$ and $Z \perp\!\!\!\perp Y$, then

$$T = \frac{Z}{\sqrt{\frac{Y}{k}}} \sim t_k.$$

In this section, we will derive the pdf of T.

The strategy that we will use is (a) first find an expression for the cdf of T and then (b) differentiate the cdf to obtain the pdf. The cdf of T is

$$\begin{aligned} P(T \le t) &= F_T(t) = P\left[\frac{Z}{\sqrt{\frac{Y}{k}}} \le t\right] \\ &= P\left[Z \le t\sqrt{\frac{Y}{k}}\right] = \int_0^\infty \int_{-\infty}^{t\sqrt{y/k}} f_Z(z) f_Y(y) dz\, dy \end{aligned}$$

using $f_{Z,Y}(z,y) = f_Z(z)f_Y(y)$ which follows from $Y \perp\!\!\!\perp Z$. Using Leibnitz's rule, the pdf of T is

$$\begin{aligned} f_T(t) &= \frac{d}{dt}F_T(t) = \int_0^\infty \frac{d}{dt}\int_{-\infty}^{t\sqrt{y/k}} f_Z(z) f_Y(y) dz\, dy \\ &= \int_0^\infty \sqrt{y/k} f_Z\left(t\sqrt{y/k}\right) f_Y(y) dy. \end{aligned}$$

Substituting the pdfs for Z and Y yields

$$\begin{aligned} f_T(t) &= \int_0^\infty \sqrt{y/k}\frac{e^{-\frac{1}{2}\frac{t^2}{k}y} y^{\frac{k}{2}-1} e^{-\frac{1}{2}y}}{\sqrt{2\pi}\,\Gamma\left(\frac{k}{2}\right)2^{\frac{k}{2}}} dy = \int_0^\infty \frac{y^{\frac{k+1}{2}-1} e^{-\frac{1}{2}\left(\frac{t^2}{k}+1\right)y}}{\sqrt{k\pi}\,\Gamma\left(\frac{k}{2}\right)2^{\frac{k+1}{2}}} dy \\ &= \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)\sqrt{k\pi}\left(\frac{t^2}{k}+1\right)^{\frac{k+1}{2}}} I_{(-\infty,\infty)}(t). \end{aligned}$$

The last integral is evaluated by recognizing the kernel of a gamma distribution. That is,

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

9.7 Pivotal Quantities

1. Definition: A pivotal quantity is a function of a statistic and a parameter. The distribution of the pivotal quantity does not depend on any unknown parameters.

2. How to construct confidence intervals. Suppose that $Q(\mathbf{T};\theta)$ is a pivotal quantity. The distribution of Q is known, so percentiles of Q can be computed. Let q_1 and q_2 be percentiles that satisfy

   $$P\left[q_1 \le Q(\mathbf{T};\theta) \le q_2\right] = 1-\alpha.$$

   If $Q(\mathbf{T};\theta)$ is a monotonic increasing or decreasing function of θ for each realization of $\mathbf{T}$, then the inverse function $Q^{-1}\left[Q(\mathbf{T};\theta)\right] = \theta$ exists, and

   $$P\left[Q^{-1}(q_1) \le \theta \le Q^{-1}(q_2)\right] = 1-\alpha$$
   if $Q(\mathbf{T};\theta)$ is an increasing function of θ and
   $$P\left[Q^{-1}(q_2) \le \theta \le Q^{-1}(q_1)\right] = 1-\alpha$$
   if $Q(\mathbf{T};\theta)$ is a decreasing function of θ.

3. Example 1: Suppose that $X_1,\ldots,X_n$ is a random sample from $\mathrm{N}(\mu,\sigma^2)$.

   (a) $Q(\overline{X},S_X;\mu) = \dfrac{\overline{X}-\mu}{S_X/\sqrt{n}} \sim t_{n-1}$ which reveals that Q is a pivotal quantity. Note that

   $$\mathbf{T} = \begin{pmatrix}\overline{X}\\ S_X\end{pmatrix}$$

   is two-dimensional. Also, Q is a decreasing function of μ,

   $$Q^{-1}(Q) = \overline{X} - \frac{S_X}{\sqrt{n}}Q = \mu \text{ and}$$
   $$P\left[\overline{X} - \frac{S_X}{\sqrt{n}}\,q_2 \le \overline{X} - \frac{S_X}{\sqrt{n}}\,q_1\right] = 1-\alpha,$$

   where q_1 and q_2 are appropriate percentiles of the t_{n-1} distribution.

   (b) $Q(S_X^2;\sigma^2) = \dfrac{(n-1)S_X^2}{\sigma^2} \sim \chi^2_{n-1}$ which reveals that Q is a pivotal quantity. Furthermore, Q is a decreasing function of σ,

   $$Q^{-1}(Q) = \sqrt{\frac{(n-1)S_X^2}{Q}} = \sigma \text{ and}$$

$$P\left[\sqrt{\frac{(n-1)S_X^2}{q_2}} \leq \sigma \leq \sqrt{\frac{(n-1)S_X^2}{q_1}}\right] = 1-\alpha,$$

where q_1 and q_2 are appropriate percentiles of the χ^2_{n-1} distribution.

4. Example 2: Suppose that $X_1, \ldots, X_n$ is a random sample from Unif$(0, \theta)$. It is easy to show that $X_{(n)}$ is sufficient statistic. Note that $X_i/\theta \sim$ Unif$(0, 1)$. Accordingly, $X_{(n)}/\theta$ is distributed as the largest order statistic from a Unif$(0, 1)$ distribution. That is, $Q(X_{(n)}; \theta) = X_{(n)}/\theta \sim$ Beta$(n, 1)$ which reveals that Q is a pivotal quantity. Furthermore, Q is a decreasing function of θ,

$$Q^{-1}(Q) = \frac{X_{(n)}}{Q} = \theta \text{ and}$$
$$P\left[\frac{X_{(n)}}{q_2} \leq \theta \leq \frac{X_{(n)}}{q_1}\right] = 1-\alpha,$$

where q_1 and q_2 are appropriate percentiles of the Beta$(n, 1)$ distribution.

(a) Note, $q_2 = 1$ is the 100^{th} percentile of Beta$(n, 1)$ and $q_1 = \alpha^{1/n}$ is the 100α percentile of Beta$(n, 1)$. Accordingly, a $100(1-\alpha)$ confidence interval for θ can be based on

$$P\left[X_{(n)} \leq \theta \leq \frac{X_{(n)}}{\alpha^{1/n}}\right] = 1-\alpha.$$

(b) Note, $q_1 = 0$ is the 0^{th} percentile of Beta$(n, 1)$ and $q_2 = (1-\alpha)^{1/n}$ is the $100(1-\alpha)$ percentile of Beta$(n, 1)$. Accordingly, a $100(1-\alpha)$ one-sided confidence interval for θ can be based on

$$P\left[\frac{X_{(n)}}{(1-\alpha)^{1/n}} \leq \theta \leq \infty\right] = 1-\alpha.$$

9.8 Estimating a Mean Difference

1. Setting: Suppose that $T_{1,n_1} \dot{\sim} \text{N}(\theta_1, \omega_1^2/n_1)$; $T_{2,n_2} \dot{\sim} \text{N}(\theta_2, \omega_2^2/n_2)$; and $T_{1,n_1} \perp\!\!\!\perp T_{2,n_2}$. The goal is to construct a confidence interval for $\theta_1 - \theta_2$. Note that

$$T_{1,n_1} - T_{2,n_2} \dot{\sim} \text{N}\left(\theta_1 - \theta_2, \frac{\omega_1^2}{n_1} + \frac{\omega_2^2}{n_2}\right).$$

If W_1^2 and W_2^2 are consistent estimators of ω_1^2 and ω_2^2 (i.e., $W_1^2 \xrightarrow{\text{prob}} \omega_1^2$ and $W_2^2 \xrightarrow{\text{prob}} \omega_2^2$), then

$$\frac{(T_{1,n_1} - T_{2,n_2}) - (\theta_1 - \theta_2)}{\sqrt{\frac{W_1^2}{n_1} + \frac{W_2^2}{n_2}}} \dot{\sim} \text{N}(0, 1).$$

A large sample $100(1-\alpha)\%$ confidence interval for $\theta_1 - \theta_2$ can be based on

$$P\left[T_{1,n_1} - T_{2,n_2} - z_{\alpha/2}\text{SE} \le \theta_1 - \theta_2 \le T_{1,n_1} - T_{2,n_2} + z_{\alpha/2}\text{SE}\right] \approx 1-\alpha,$$

$$\text{where SE} = \text{SE}(T_{1,n_1} - T_{2,n_2}) = \sqrt{\frac{W_1^2}{n_1} + \frac{W_2^2}{n_2}}.$$

2. Application 1: Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from a population having mean μ_1 and variance σ_1^2 and that $X_{21}, X_{22}, \ldots, X_{2n_1}$ is an independent random sample from a population having mean μ_2 and variance σ_2^2. Then

$$\frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} \text{N}(0,1).$$

A large sample $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ can be based on

$$P\left[\overline{X}_1 - \overline{X}_2 - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \le \mu_1 - \mu_2 \le \overline{X}_1 - \overline{X}_2 + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right] \approx 1-\alpha.$$

3. Application 2: Suppose that $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from Bern(p_1) and that $X_{21}, X_{22}, \ldots, X_{2n_1}$ is an independent random sample from Bern(p_2). Then

$$\frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \dot{\sim} \text{N}(0,1).$$

A large sample $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ can be based on

$$P\left[\hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\text{SE} \le p_1 - p_2 \le \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\text{SE}\right] \approx 1-\alpha, \text{ where}$$

$$\text{SE} = \text{SE}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

9.9 Estimating Variability

Most of this section is a review of earlier material. The only new material is concerned with the distribution of the sample range when sampling from N$(0,1)$.

1. Let $X_1, X_2, \ldots, X_n$ be a random sample from N(μ, σ^2). Define Z_i as $Z_i = (X_i - \mu)/\sigma$. Then $Z_i \sim$ iid N$(0,1)$. Note that the joint distribution of $Z_1, \ldots, Z_n$ does not depend on μ or σ. Accordingly, the distribution of

$$R_Z = Z_{(n)} - Z_{(1)} = \frac{X_{(n)} - \mu}{\sigma} - \frac{X_{(1)} - \mu}{\sigma} = \frac{X_{(n)} - X_{(1)}}{\sigma} = \frac{R_X}{\sigma}$$

does not depend on μ or σ. That is, R_X/σ is a pivotal quantity.

2. Percentiles of $W = R_X/\sigma$ for various sample sizes are listed in Table XIII. They can be used to make probability statements such as

$$P(w_1 \leq W \leq w_2) = 1 = \alpha,$$

where w_1 and w_2 are appropriate percentiles of the distribution of W. Note that W is a decreasing function of σ and that $W^{-1}(W) = (X_{(n)} - X_{(1)})/W = \sigma$. Therefore, confidence intervals can be based on

$$P\left[\frac{X_{(n)} - X_{(1)}}{w_2} \leq \sigma \leq \frac{X_{(n)} - X_{(1)}}{w_1}\right] = 1 - \alpha.$$

3. Table XIII also gives $\mathrm{E}(W)$ and σ_W. These values can be used to obtain a point estimator of σ and to compute the standard error of the estimator. The point estimator is

$$\hat{\sigma} = \frac{X_{(n)} - X_{(1)}}{\mathrm{E}(W)}.$$

The estimator $\hat{\sigma}$ is unbiased for σ because

$$\mathrm{E}(\hat{\sigma}) = \frac{\mathrm{E}(R_X)}{\mathrm{E}(W)} = \frac{\sigma\mathrm{E}(R_X)}{\mathrm{E}(R_X)} = \sigma.$$

The variance of $\hat{\sigma}$ is

$$\begin{aligned}\mathrm{Var}(\hat{\sigma}) &= \mathrm{Var}\left(\frac{R_X}{\mathrm{E}(W)}\right) = \mathrm{Var}\left(\frac{\sigma W}{\mathrm{E}(W)}\right) \\ &= \sigma^2\frac{\mathrm{Var}(W)}{[\mathrm{E}(W)]^2}.\end{aligned}$$

Accordingly,

$$\mathrm{SE}(\hat{\sigma}) = \hat{\sigma}\frac{\sigma_W}{\mathrm{E}(W)}$$

is an estimator of $\sqrt{\mathrm{Var}(\hat{\sigma})}$.

9.10 Deriving Estimators

1. Method of Moments

   (a) Setting: Suppose that $X_i, X_2, \ldots, X_n$ is a random sample from $f_X(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters. The goal is to derive an estimator of $\boldsymbol{\theta}$.

   (b) Let M_j' be the j^{th} sample moment about the origin and let μ_j' be the corresponding population moment. That is

$$M_j' = \frac{1}{n}\sum_{i=1}^{n} X_i^j \text{ and } \mu_j' = \mu_j'(\boldsymbol{\theta}) = \mathrm{E}(X^j)$$

for $j = 1, 2, \ldots$. The population moments are denoted by $\mu_j'(\boldsymbol{\theta})$ because they are functions of the components of $\boldsymbol{\theta}$.

(c) The method of moments estimator consists of equating sample moments to population moments and solving for $\boldsymbol{\theta}$. That is, solve

$$M_j' = \mu_j'(\widehat{\boldsymbol{\theta}}),\ j = 1, \ldots, k$$

for $\widehat{\boldsymbol{\theta}}$.

(d) Central Moments: It sometimes is more convenient to use central moments. For the special case of $k = 2$, solve

$$\begin{aligned} M_j &= \mu_j(\widehat{\boldsymbol{\theta}}),\ j = 1, 2 \text{ where} \\ M_1 &= M_1' = \frac{1}{n}\sum_{i=1}^{n} X_i; \quad M_2 = S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2; \\ \mu_1 &= \mu_1' = \mathrm{E}(X);\ \text{and } \mu_2 = \sigma_X^2 = \mathrm{E}(X - \mu_1)^2. \end{aligned}$$

(e) Example: Gamma Distribution. Suppose $X_1, \ldots, X_n$ is a random sample from Gamma(α, λ). The moments about the origin are

$$\mu_j'(\boldsymbol{\theta}) = \mathrm{E}(X^j) = \frac{\Gamma(\alpha + j)}{\Gamma(\alpha)\lambda^j}.$$

The first two central moments are

$$\mu_1(\boldsymbol{\theta}) = \mathrm{E}(X) = \frac{\alpha}{\lambda} \text{ and } \mu_2(\boldsymbol{\theta}) = \mathrm{Var}(X) = \frac{\alpha}{\lambda^2}.$$

The method of moments estimators of α and λ are obtained by solving

$$\overline{X} = \frac{\widehat{\alpha}}{\widehat{\lambda}} \text{ and } S_X^2 = \frac{\widehat{\alpha}}{\widehat{\lambda}^2}$$

for $\widehat{\alpha}$ and $\widehat{\lambda}$. The solutions are

$$\widehat{\alpha} = \frac{\overline{X}^2}{S_X^2} \text{ and } \widehat{\lambda} = \frac{\overline{X}}{S_X^2}.$$

(f) Example: Beta Distribution. Suppose $X_1, \ldots, X_n$ is a random sample from Beta(α_1, α_2). The moments about the origin are

$$\mu_j'(\boldsymbol{\theta}) = \mathrm{E}(X^j) = \frac{B(\alpha_1 + j, \alpha_2)}{B(\alpha_1, \alpha_2)} = \frac{\Gamma(\alpha_1 + j)\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_1 + \alpha_2 + j)}.$$

The first two central moments are

$$\begin{aligned} \mu_1(\boldsymbol{\theta}) &= \mathrm{E}(X) = \frac{\alpha_1}{\alpha_1 + \alpha_2} \text{ and} \\ \mu_2(\boldsymbol{\theta}) &= \mathrm{Var}(X) = \frac{\alpha_1\alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}. \end{aligned}$$

The method of moments estimators of α_1 and α_2 are obtained by solving

$$\overline{X} = \frac{\widehat{\alpha}_1}{\widehat{\alpha}_1 + \widehat{\alpha}_2} \text{ and } S_X^2 = \frac{\widehat{\alpha}_1\widehat{\alpha}_2}{(\widehat{\alpha}_1 + \widehat{\alpha}_2)^2(\widehat{\alpha}_1 + \widehat{\alpha}_2 + 1)}$$

for $\widehat{\alpha}_1$ and $\widehat{\alpha}_2$. The solutions are

$$\widehat{\alpha}_1 = \overline{X}\left[\frac{\overline{X}(1-\overline{X})}{S_X^2} - 1\right] \text{ and } \widehat{\alpha}_2 = (1-\overline{X})\left[\frac{\overline{X}(1-\overline{X})}{S_X^2} - 1\right].$$

2. Maximum Likelihood Estimators (MLEs)

   (a) Setting: Suppose that $X_i, X_2, \ldots, X_n$ is a random sample from $f_X(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters. The goal is to derive an estimator of $\boldsymbol{\theta}$.

   (b) Definition: A maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$ is any value $\widehat{\boldsymbol{\theta}}$ that maximizes the likelihood function and is a point in the parameter space or on the boundary of the parameter space.

   (c) If the likelihood function is a differentiable function of $\boldsymbol{\theta}$, then the maximum likelihood estimator is a solution to

$$\left.\frac{\partial L(\boldsymbol{\theta}|\mathbf{X})}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = \mathbf{0}.$$

   (d) Note, any maximizer of $L(\boldsymbol{\theta}|\mathbf{X})$ also is a maximizer of $\ln[L(\boldsymbol{\theta}|\mathbf{X})]$. Accordingly, one can maximize the log likelihood function rather than the likelihood function.

   (e) Example: Suppose that $X_1, \ldots, X_n$ is a random sample from Expon(λ). The log likelihood function is

$$\ln[L(\lambda|\mathbf{X})] = n\ln(\lambda) - \lambda\sum_{i=1}^{n} X_i.$$

   Taking the derivative with respect to λ; setting it to zero; and solving for $\widehat{\lambda}$ yields

$$\widehat{\lambda} = \frac{1}{\overline{X}}.$$

   (f) Example: Suppose that $X_1, \ldots, X_n$ is a random sample from Unif(θ). The likelihood function is

$$L(\theta|X_{(n)}) = \frac{1}{\theta^n} I_{(X_{(n)},\infty)}(\theta).$$

   Plotting the likelihood function reveals that $\widehat{\theta} = X_{(n)}$ is the MLE. Note, taking derivatives does not work in this case.

(g) Example: Suppose that $X_1, \ldots, X_n$ is a random sample from $\text{N}(\mu, \sigma^2)$. The log likelihood function is

$$\ln\left[L\left(\mu, \sigma^2 | \overline{X}, S_X^2\right)\right] = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 - \frac{n}{2\sigma^2}(\overline{X} - \mu)^2.$$

Taking the derivatives with respect to μ and σ^2 and setting them to zero yields two equations to solve:

$$\frac{n}{\widehat{\sigma}^2}(\overline{X} - \widehat{\mu}) = 0 \text{ and}$$
$$-\frac{n}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{n}(X_i - \overline{X})^2 + \frac{n}{2\widehat{\sigma}^4}(\overline{X} - \widehat{\mu})^2 = 0.$$

Solving the first equation for $\widehat{\mu}$ yields $\widehat{\mu} = \overline{X}$. Substituting $\widehat{\mu} = \overline{X}$ into the second equation and solving for $\widehat{\sigma}^2$ yields

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(h) Invariance property of MLEs. Let $g(\boldsymbol{\theta})$ be a function of $\boldsymbol{\theta}$. Then, the MLE of $g(\boldsymbol{\theta})$ is $g(\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$.

Proof: We will prove the invariance property only for the special case in which $g(\boldsymbol{\theta})$ is a one-to-one function. Note, if the dimension of $\boldsymbol{\theta}$ is k, then the dimension of $g(\boldsymbol{\theta})$ also must be k. Let $\boldsymbol{\eta} = g(\boldsymbol{\theta})$. Then $\boldsymbol{\theta} = g^{-1}(\boldsymbol{\eta})$ because g is one-to-one. Define $L^*(\boldsymbol{\eta}|\mathbf{X})$ as the likelihood function when $g(\boldsymbol{\theta}) = \boldsymbol{\eta}$. That is,

$$L^*(\boldsymbol{\eta}|\mathbf{X}) = f_{\mathbf{X}}[\mathbf{X}|g^{-1}(\boldsymbol{\eta})] = L[g^{-1}(\boldsymbol{\eta})|\mathbf{X}], \text{ where } g^{-1}(\boldsymbol{\eta}) = \boldsymbol{\theta}.$$

Note that,

$$\max_{\boldsymbol{\eta}} L^*(\boldsymbol{\eta}|\mathbf{X}) = \max_{\boldsymbol{\eta}} L[g^{-1}(\boldsymbol{\eta})|\mathbf{X}] = \max_{\boldsymbol{\theta}} L[\boldsymbol{\theta}|\mathbf{X}].$$

That is, the maximized likelihood is the same whether one maximizes L^* with respect to $\boldsymbol{\eta}$ or maximizes L with respect to $\boldsymbol{\theta}$. Accordingly, if $\widehat{\boldsymbol{\theta}}$ maximizes the likelihood function $L(\boldsymbol{\theta}|\mathbf{X})$, then $\widehat{\boldsymbol{\eta}} = g(\widehat{\boldsymbol{\theta}})$ maximizes the likelihood function $L^*(\boldsymbol{\eta}|\mathbf{X})$.

Example Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from $\text{N}(\mu, \sigma^2)$. Find the MLE of the coefficient of variation $g(\mu, \sigma^2) = 100\sigma/\mu$. Solution:

$$\widehat{g} = 100\frac{\widehat{\sigma}}{\overline{X}}, \text{ where } \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(i) Properties of MLEs: Under certain regularity conditions, it can be shown (we will not do so) that

i. MLEs are consistent,

ii. MLEs have asymptotic normal distributions, and

iii. MLEs are functions of sufficient statistics. This property is easy to prove because the likelihood function depends on the data solely through a sufficient statistic.

3. Rao-Blackwell Theorem If $U(\mathbf{X})$ is unbiased for θ (a scalar) and $T(\mathbf{X})$ is sufficient, then $V = V(\mathbf{T}) = \mathrm{E}(U|T)$ is

   (a) a statistic,

   (b) unbiased for θ, and

   (c) $\mathrm{Var}(V) \leq \mathrm{Var}(U)$, with strict inequality iff and only if U is a function of $\mathbf{T}$.

   Proof.

   (a) V is a statistic because the distribution of $\mathbf{X}$ conditional on $\mathbf{T}$ does not depend on θ. Accordingly, the expectation of V with respect to the distribution of $\mathbf{X}$ conditional on $\mathbf{T}$ does not depend on θ.

   (b) Note, $\mathrm{E}(U) = \theta$ because U is unbiased. Now use iterated expectation:

$$\theta = \mathrm{E}(U) = \mathrm{E}_T\left[\mathrm{E}(U|\mathbf{T})\right] = \mathrm{E}_T(V) = \mathrm{E}(V).$$

   (c) Use iterated variance:

$$\begin{aligned} \mathrm{Var}(U) &= \mathrm{E}\left[\mathrm{Var}(U|\mathbf{T})\right] + \mathrm{Var}\left[\mathrm{E}(U|\mathbf{T})\right] \\ &= \mathrm{E}\left[\mathrm{Var}(U|\mathbf{T})\right] + \mathrm{Var}(V) \geq \mathrm{Var}(V). \end{aligned}$$

   Example. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from $\mathrm{Poi}(\lambda)$. The goal is to find a good unbiased estimator of of $P(X = 0) = e^{-\lambda}$. Recall that $T = \sum_{i=1}^{n} X_i$ is sufficient and that $T \sim \mathrm{Poi}(n\lambda)$. Consider

$$U = I_{\{0\}}(X_1) = \begin{cases} 1 & \text{if } X_1 = 0 \\ 0 & \text{if } X_1 = 1. \end{cases}$$

   The support of U is $\{0, 1\}$ and the expectation of U is

$$\mathrm{E}(U) = 1 \times P(U = 1) + 0 \times P(U = 0) = 1 \times P(X_1 = 0) = e^{-\lambda}.$$

   Thus, U is unbiased for $e^{-\lambda}$. To find a better estimator, use the Rao-Blackwell theorem. It was shown on page 75 of these notes that the conditional distribution of $X_1, X_2, \ldots, X_k$ given $T = t$ is

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n | T = t)$$

$$= \binom{t}{x_1, x_2, \ldots, x_n} \left(\frac{1}{n}\right)^{x_1} \left(\frac{1}{n}\right)^{x_2} \cdots \left(\frac{1}{n}\right)^{x_n}.$$

That is, given $T = t$, the X's have a multinomial distribution with t trials, n categories, and probability $1/n$ for each category. Note that the conditional distribution of the data given T does not depend on λ. This is because T is sufficient. The conditional distribution of X_1 given $T = t$ is binomial:

$$X_1|(T = t) \sim \text{Bin}\left[T, \frac{1}{n}\right].$$

The expectation of U given $T = t$ is

$$\begin{aligned} \text{E}(U|T = t) &= 1 \times P(U = 1|T = 1) + 0 \times P(U = 0|T) \\ &= P(U = 1|T = t) = P(X_1 = 0|T = t) \\ &= \binom{t}{0} \left(\frac{1}{n}\right)^0 \left(1 - \frac{1}{n}\right)^{t-0} = \left(1 - \frac{1}{n}\right)^t. \end{aligned}$$

Accordingly, an unbiased estimator of $e^{-\lambda}$ that has smaller variance than U is

$$\text{E}(U|T) = \left(1 - \frac{1}{n}\right)^T.$$

9.11 Bayes Estimators

1. Setting: Suppose that we have (a) data, $X_1, \ldots, X_n$, (b) the corresponding likelihood function, $L(\theta|\mathbf{T})$, where $\mathbf{T}$ is sufficient, and (c) a prior distribution for θ, $g_\Theta(\theta)$. Then, if we are skillful, we can find the posterior $g_{\Theta|T}(\theta|\mathbf{T})$. We would like to find point and interval estimators of θ.

2. Point Estimators: In general, we can use some characteristic of the posterior distribution as our point estimator. Suitable candidates are the mean, median, or mode. How do we choose which one to use?

   (a) Loss Function: Suppose that we can specify a loss function that describes the penalty for missing the mark when estimating θ. Denote our estimator as a or $a(t)$ because it will depend on t. Two possible loss functions are

   $$\ell_1(\theta, a) = |\theta - a| \text{ and } \ell_2(\theta, a) = (\theta - a)^2.$$

   (b) Bayes Estimator: Recall that Θ is a random variable. A posterior estimator, a, is a Bayes estimator with loss function ℓ if a minimizes the posterior expected loss (Bayes loss):

   $$\text{Bayes Loss } = B(a) = \text{E}_{\Theta|T}\left[\ell(\Theta - a)\right].$$

(c) From prior results (see page 24 of these notes), the Bayes estimator for loss ℓ_1 is known to be the median of the posterior distribution.
Recall that $\mathrm{E}(X)$ is the minimizer of $\mathrm{E}(X-a)^2$ with respect to a. Accordingly, the Bayes estimator for loss ℓ_2 is the mean of the posterior distribution

(d) Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from Bern(θ). Recall that $T = \sum_{i=1}^n X_i$ is sufficient. Furthermore, suppose that the prior on θ is $\Theta \sim \text{beta}(\alpha_1, \alpha_2)$. Then the posterior, conditional on $T = t$ is $\Theta \sim \text{beta}(\alpha_1 + t, \alpha_2 + n - t)$. The mean of the posterior is

$$\mathrm{E}(\Theta|T=t) = \frac{\alpha_1 + t}{\alpha_1 + \alpha_2 + n}.$$

If $n = 10$, $\alpha_1 = \alpha_2 = 1$, and $t = 6$, then

$$\mathrm{E}(\Theta|T=t) = \frac{7}{12} = 0.5833$$

and the median of the beta$(7, 5)$ distribution is 0.5881, obtained by using Matlab's betainv function.

3. Interval Estimator: Use the posterior distribution to find lower and upper limits, h_1 and h_2 such that

$$P(h_1 \leq \Theta \leq h_2 | T = t) = 1 - \alpha.$$

The above interval is a Bayesian $100(1-\alpha)\%$ confidence interval. In the statistical literature, this interval usually is called a credibility interval. Unlike the frequentist confidence interval, the credibility interval is interpreted as a probability. That is, we say that Θ is contained in the interval with probability $1 - \alpha$. This is the interpretation that 216 students often give to frequentist intervals and we give them no credit when they do so.

Example 1 In the binomial example, the posterior distribution of Θ is beta$(7, 5)$. Using Matlab's betainv function, a 95% Bayesian confidence interval is

$$P(0.3079 \leq \Theta \leq 0.8325) = 0.95.$$

Example 2 Suppose that $X_1 \ldots, X_n$ is a random sample from $\mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. If the prior on μ is $\mu \sim \mathrm{N}(\nu, \tau^2)$, then the posterior distribution is

$$\mu|\bar{x} \sim \mathrm{N}\left[\frac{\pi_{\bar{x}}}{\pi_{\bar{x}} + \pi_\mu}\bar{x} + \frac{\pi_\mu}{\pi_{\bar{x}} + \pi_\mu}\nu, \left(\pi_{\bar{x}} + \pi_\mu\right)^{-1}\right],$$

where the precisions are

$$\pi_{\bar{x}} = \frac{n}{\sigma^2}; \text{ and } \pi_\mu = \frac{1}{\tau^2}.$$

A 95% Bayesian confidence interval for μ is

$$P\left[\widehat{\mu} - 1.96\sqrt{(\pi_{\bar{x}} + \pi_\mu)^{-1}} \leq \mu \leq \widehat{\mu} + 1.96\sqrt{(\pi_{\bar{x}} + \pi_\mu)^{-1}} \,\middle|\, \overline{X} = \bar{x}\right] = 0.95 \text{ where}$$

$$\widehat{\mu} = \frac{\pi_{\bar{x}}}{\pi_{\bar{x}} + \pi_\mu}\bar{x} + \frac{\pi_\mu}{\pi_{\bar{x}} + \pi_\mu}\nu.$$

As τ increases (indicating less and less a priori knowledge about μ), the Bayesian confidence interval approaches

$$P\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}} \,\middle|\, \overline{X} = \bar{x}\right] = 0.95.$$

This is interpreted as a fixed interval that has probability 0.95 of capturing the random variable μ. Note that the above Bayesian credibility interval is identical to the usual 95% frequentist confidence interval for μ when σ^2 is known.

9.12 Efficiency

1. This section is concerned with optimal estimators. For example, suppose that we are interested in estimating $g(\theta)$, for some function g. The question to be addressed is—how do we know if we have the best possible estimator? A partial answer is given by the Cramér-Rao inequality. It gives a lower bound on the variance of an unbiased estimator. If our estimator attains the Cramér-Rao lower bound, then we know that we have the best unbiased estimator.

2. Cramér-Rao Inequality (Information Inequality). Suppose that the joint pdf (pmf) of $X_1, X_2, \ldots, X_n$ is $f_{\mathbf{X}}(\mathbf{x}|\theta)$, where θ is a scalar and the support of $\mathbf{X}$ does not depend on θ. Furthermore, suppose that the statistic $T(\mathbf{X})$ is an unbiased estimator of a differentiable function of θ. That is, $\mathrm{E}(T) = g(\theta)$. Then, under mild regularity conditions,

$$\mathrm{Var}(T) \geq \frac{\left[\frac{\partial g(\theta)}{\partial \theta}\right]^2}{I_\theta}, \text{ where } I_\theta = \mathrm{E}\left[\left(\frac{\partial \ln f_{\mathbf{X}}(\mathbf{X}|\theta)}{\partial \theta}\right)^2\right].$$

The quantity I_θ is called Fisher's information and it is an index of the amount of information that $\mathbf{X}$ has about θ.

Proof. Define the random variable S as

$$S = S(\mathbf{X}, \theta) = \frac{\partial \ln f_{\mathbf{X}}(\mathbf{X}|\theta)}{\partial \theta} = \frac{1}{f_{\mathbf{X}}(\mathbf{X}|\theta)}\frac{\partial f_{\mathbf{X}}(\mathbf{X}|\theta)}{\partial \theta}.$$

This quantity is called the score function (not in your text). Your text denotes the score function by W.

(a) Result: The expected value of the score function is zero.

Proof: This result can be shown by interchanging integration and differentiation (justified if the regularity conditions are satisfied):

$$\begin{aligned}
\mathrm{E}(S) &= \int S(\mathbf{x},\theta) f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} \\
&= \int \frac{1}{f_{\mathbf{X}}(\mathbf{x}|\theta)} \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} \\
&= \int \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta}\, d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} \int f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} 1 = 0.
\end{aligned}$$

because the integral of the joint pdf over the entire sample space is 1. Substitute summation for integration if the random variables are discrete.

(b) Result: The variance of S is I_θ.

Proof: This result follows from the first result and from the definition of I_θ:

$$\mathrm{Var}(S) = \mathrm{E}(S^2) - [\mathrm{E}(S)]^2 = \mathrm{E}(S^2) = I_\theta.$$

(c) Result: The covariance between S and T is

$$\mathrm{Cov}(S,T) = \frac{\partial g(\theta)}{\partial \theta}.$$

Proof: To verify this result, again we will interchange integration and differentiation. First, note that $\mathrm{Cov}(S,T) = \mathrm{E}(ST) - \mathrm{E}(S)\mathrm{E}(T) = \mathrm{E}(ST)$ because $\mathrm{E}(S) = 0$. Accordingly,

$$\begin{aligned}
\mathrm{Cov}(S,T) &= \mathrm{E}(ST) = \int S(\mathbf{x},\theta) T(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} \\
&= \int \frac{1}{f_{\mathbf{X}}(\mathbf{x}|\theta)} \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} T(\mathbf{x}) f_{\mathbf{X}}(\mathbf{x}|\theta)\, d\mathbf{x} \\
&= \int \frac{\partial f_{\mathbf{X}}(\mathbf{x}|\theta)}{\partial \theta} T(\mathbf{x})\, d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} \int f_{\mathbf{X}}(\mathbf{x}|\theta) T(\mathbf{x})\, d\mathbf{x} \\
&= \frac{\partial}{\partial \theta} \mathrm{E}(T) = \frac{\partial g(\theta)}{\partial \theta}.
\end{aligned}$$

(d) Result: Cramér-Rao Inequality:

$$\mathrm{Var}(T) \geq \frac{\left[\frac{\partial g(\theta)}{\partial \theta}\right]^2}{I_\theta}.$$

The right-hand-side of the above equation is called the Cramér-Rao Lower Bound (CRLB). That is,

$$\text{CRLB } = \frac{\left[\frac{\partial g(\theta)}{\partial \theta}\right]^2}{I_\theta}.$$

Proof: If ρ is a correlation coefficient, then from the Cauchy-Schwartz inequality it is known that $0 \leq \rho^2 \leq 1$. Accordingly,

$$\rho^2_{S,T} = \frac{[\text{Cov}(S,T)]^2}{\text{Var}(S)\,\text{Var}(T)} \leq 1$$

$$\implies \text{Var}(T) \geq \frac{[\text{Cov}(S,T)]^2}{\text{Var}(S)} = \frac{\left[\frac{\partial g(\theta)}{\partial \theta}\right]^2}{I_\theta}.$$

(e) If $g(\theta) = \theta$, then the inequality simplifies to

$$\text{Var}(T) \geq \frac{1}{I_\theta} \text{ because } \frac{\partial}{\partial \theta}\theta = 1.$$

(f) Note: if $X_1, X_2, \ldots, X_n$ are iid, then the score function can be written as

$$\begin{aligned} S(\mathbf{X}|\theta) &= \frac{\partial}{\partial \theta}\sum_{i=1}^{n} \ln f_X(X_i|\theta) \\ &= \sum_{i=1}^{n} \frac{\partial \ln f_X(X_i|\theta)}{\partial \theta} \\ &= \sum_{i=1}^{n} S_i(X_i, \theta), \text{ where } S_i = \frac{\partial \ln f_X(X_i|\theta)}{\partial \theta} \end{aligned}$$

is the score function for X_i. The score functions S_i for $i = 1, \ldots, n$ are iid, each with mean zero. Accordingly,

$$\begin{aligned} \text{Var}(S) = \text{Var}\left(\sum_{i=1}^{n} S_i\right) &= \sum_{i=1}^{n} \text{Var}(S_i) \text{ by independence} \\ &= \sum_{i=1}^{n} \text{E}(S_i^2) = n\text{E}(S_1^2), \end{aligned}$$

where S_1 is the score function for X_1.

3. Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from Poi(λ). The score function for a single X is

$$S(X_i, \lambda) = \frac{\partial\,[-\lambda + X_i \ln(\lambda) - \ln(X_i!)]}{\partial \lambda} = -1 + \frac{X_i}{\lambda}.$$

Accordingly, the information is

$$I_\theta = n\,\text{Var}(-1 + X_i/\lambda) = n\frac{\text{Var}(X_i)}{\lambda^2} = n\frac{\lambda}{\lambda^2} = \frac{n}{\lambda}.$$

Suppose that the investigator would like to estimate $g(\lambda) = \lambda$. The MLE of λ is $\overline{X}$ (homework) and $\text{E}(\overline{X}) = \lambda$, so the MLE is unbiased. The variance of a Poisson random variable is λ and therefore $\text{Var}(\overline{X}) = \lambda/n$. The CRLB for estimating λ is

$$\text{CRLB} = \frac{\left[\frac{\partial}{\partial\lambda}\lambda\right]^2}{n/\lambda} = \frac{\lambda}{n}.$$

Therefore, the MLE attains the CRLB.

4. Efficiency: The efficiency of an unbiased estimator of $g(\theta)$ is the ratio of the CRLB to the variance of the estimator. That is, suppose that T is an unbiased estimator of $g(\theta)$. Then the efficiency of T is

$$\begin{aligned}\text{Efficiency} &= \frac{\text{CRLB}}{\text{Var}(T)} \\ &= \frac{\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2}{I_\theta} \div \text{Var}(T) = \frac{\left(\frac{\partial g(\theta)}{\partial\theta}\right)^2}{I_\theta\,\text{Var}(T)}.\end{aligned}$$

If this ratio is one, then the estimator is said to be efficient. Efficiency always is less than or equal to one.

5. Exponential Family Results: Recall, if the distribution of X belongs to the one parameter exponential family and $X_1, X_2, \ldots, X_n$ is a random sample, then the joint pdf (pmf) is

$$f_{\mathbf{X}}(\mathbf{X}|\theta) = [B(\theta)]^n \left[\prod_{i=1}^n h(X_i)\right] \exp\left\{Q(\theta)\sum_{i=1}^n R(X_i)\right\}.$$

   (a) The score function is

$$S(T,\theta) = n\frac{\partial \ln B(\theta)}{\partial\theta} + T\frac{\partial Q(\theta)}{\partial\theta}, \text{ where } T = \sum_{i=1}^n R(X_i).$$

   (b) Note that the score function is a linear function of T:

$$S = a + bT, \text{ where } a = n\frac{\partial \ln B(\theta)}{\partial\theta} \text{ and } b = \frac{\partial Q(\theta)}{\partial\theta}.$$

   (c) Recall that $\text{E}(S) = 0$. It follows that

$$\begin{aligned} n\frac{\partial \ln B(\theta)}{\partial\theta} + \text{E}(T)\frac{\partial Q(\theta)}{\partial\theta} &= 0 \text{ and} \\ \text{E}(T) &= -n\frac{\partial \ln B(\theta)}{\partial\theta}\left[\frac{\partial Q(\theta)}{\partial\theta}\right]^{-1}.\end{aligned}$$

(d) Result: Suppose that $g(\theta)$ is chosen to be

$$g(\theta) = \mathrm{E}(T) = -n\frac{\partial \ln B(\theta)}{\partial \theta}\left[\frac{\partial Q(\theta)}{\partial \theta}\right]^{-1}.$$

Then,

$$\mathrm{Var}(T) = \frac{\left(\frac{\partial g(\theta)}{\partial \theta}\right)^2}{I_\theta} = \text{ CRLB}$$

and T is the minimum variance unbiased estimator of $g(\theta)$.

Proof: First, note that T is unbiased for $\mathrm{E}(T)$. Second, note that

$$S = a + bT \Longrightarrow \rho^2_{S,T} = 1 \Longrightarrow \frac{\left[\frac{\partial g(\theta)}{\partial \theta}\right]^2}{\mathrm{Var}(T)\, I_\theta} = 1$$

$$\Longrightarrow \mathrm{Var}(T) = \frac{\left(\frac{\partial g(\theta)}{\partial \theta}\right)^2}{I_\theta} = \text{ CRLB.}$$

6. Example. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample form Geom(θ). The pmf of X_i is

$$f_X(x_i|\theta) = (1-\theta)^{x_i - 1}\theta I_{\{1,2,\ldots\}}(x_i) = \frac{\theta}{1-\theta} I_{\{1,2,\ldots\}}(x_i)\exp\{\ln(1-\theta)x_i\}.$$

Accordingly, the distribution of X belongs to the exponential family with

$$B(\theta) = \frac{\theta}{1-\theta}; \quad h(x_i) = I_{\{1,2,\ldots\}}(x_i); \quad Q(\theta) = \ln(1-\theta); \text{ and } R(x_i) = x_i.$$

The score function for the entire sample is

$$\begin{aligned} S(\mathbf{X}, \theta) &= n\frac{\partial \ln \frac{\theta}{1-\theta}}{\partial \theta} + T\frac{\partial \ln(1-\theta)}{\partial \theta} \\ &= n\left(\frac{1}{\theta} + \frac{1}{1-\theta}\right) - \frac{T}{1-\theta}, \end{aligned}$$

where $T = \sum_{i=1}^n X_i$. It follows that

$$\mathrm{E}(T) = g(\theta) = \frac{n}{\theta}; \quad \mathrm{E}\left(\frac{1}{n}T\right) = \mathrm{E}(\overline{X}) = \frac{1}{\theta};$$

and that T is the minimum variance unbiased estimator of n/θ. Equivalently, $T/n = \overline{X}$ is the minimum variance unbiased estimator of $1/\theta$. The variance of T can be obtained from the moment generating function. the result is

$$\mathrm{Var}(X_i) = \frac{1-\theta}{\theta^2} \Longrightarrow \mathrm{Var}(T/n) = \frac{1-\theta}{n\theta^2}.$$

7. Another Exponential Family Result: Suppose that the joint pdf (pmf) of $X_1, X_2, \ldots, X_n$ is $f_{\mathbf{X}}(\mathbf{x}|\theta)$; $T(\mathbf{X})$ is a statistic that is unbiased for $g(\theta)$; and $\text{Var}(T)$ attains the Cramér-Rao lower bound. Then T is sufficient and the joint pdf (pmf) belongs to the one parameter exponential family.

   Proof: If $\text{Var}(T)$ attains the Cramér-Rao lower bound, then it must be true that $\rho^2_{S,T} = 1$ and that

   $$S(\mathbf{X}, \theta) = a(\theta) + b(\theta)T(\mathbf{X})$$

   for some functions a and b. Integrating S with respect to θ gives

   $$\begin{aligned} \int S(\mathbf{X}, \theta)\, d\theta &= \ln f_{\mathbf{X}}(\mathbf{X}|\theta) + K_1(\mathbf{X}) \text{ for some function } K_1(\mathbf{X}) \\ &= \int a(\theta) + b(\theta)T(\mathbf{X})\, d\theta = A(\theta) + B(\theta)T(\mathbf{X}) + K_2(\mathbf{X}) \\ \implies f_{\mathbf{X}}(\mathbf{X}|\theta) &= \exp\{A(\theta)\} \exp\{[K_2(\mathbf{X}) - K_1(\mathbf{X})]\} \exp\{B(\theta)T(\mathbf{X})\}. \end{aligned}$$

   which shows that the distribution belongs to the exponential family and that T is sufficient.

Chapter 10

SIGNIFICANCE TESTING

This chapter describes hypothesis testing from a Fisherian viewpoint. The main ingredients are hypotheses, test statistics, and p-values.

10.1 Hypotheses

1. H_0 and H_a are statements about probability models or, equivalently, about population characteristics.

2. The null hypothesis, H_0, usually says no effect, no difference, etc. In terms of parameters, it usually is written as $H_0 : \theta = \theta_0$; $H_0 : \theta \leq \theta_0$; or $H_0 : \theta \geq \theta_0$, where θ_0 is a value specified by the investigator. It is important that the null contains a point of equality.

3. The alternative states that the null is false and usually is written as $H_a : \theta \neq \theta_0$; $H_a : \theta > \theta_0$; or $H_a : \theta < \theta_0$.

10.2 Assessing the Evidence

1. A significance test is a test of H_0. The strategy is as follows:

   (a) Translate the scientific hypothesis into H_0 and H_a.

   (b) Begin with the assumption that H_0 is true.

   (c) Collect data.

   (d) Determine whether or not the data contradict H_0. The p-value is a measure of how strongly the data contradict the null. A small p-value is strong evidence against H_0.

2. Test statistic: Definition: A test statistic is a function of the data and θ_0. The test statistic is chosen to discriminate between H_0 and H_a. Usually, it incorporates an estimator of θ. Familiar test statistics are

(a) $Z = \dfrac{\widehat{\theta} - \theta_0}{SE(\widehat{\theta}|\mathrm{H}_0)}$

(b) $t = \dfrac{\overline{X} - \mu_0}{S_X/\sqrt{n}}$

(c) $X^2 = \dfrac{(n-1)S_X^2}{\sigma_0^2}$

3. p-value Denote the test statistic as T and denote the realized value of T as t_{obs}. The p-value is a measure of consistency between the data and H_0. It is defined as

$$p\text{-value } = P\left(T \text{ is as or more extreme than } t_{\text{obs}} \text{ in the direction of } \mathrm{H}_a \middle| \mathrm{H}_0\right).$$

A small p-value means that the data are not consistent with H_0. That is, small p-values are taken to be evidence against H_0.

4. Common Error I: Many investigators interpret a large p-value to be evidence for H_0. This is not correct. A large p-value means that there is little or no evidence against H_0. For example, consider a test of $\mathrm{H}_0\colon \mu = 100$ versus $\mathrm{H}_a\colon \mu \neq 100$ based on a sample of size $n = 1$ from $\mathrm{N}(0, 20^2)$. Suppose that the true mean is $\mu = 105$ and that $X = 101$ is observed ($4/20 = 1/5\sigma$ below the true mean). The Z statistic is

$$z_{\text{obs}} = \frac{101 - 100}{20} = 0.05$$

and the p-value is

$$p\text{-value } = 2[1 - \Phi(0.05)] = 2(1 - 0.5199) = 0.9601.$$

The p-value is large, but the data do not provide strong evidence that H_0 is true.

5. Common Error II: Many investigators interpret a very small p-value to mean that a large (important) effect has been found. This is not correct. A very small p-value is strong evidence against H_0. It is not evidence that a large effect has been found. Example: Suppose that the standard treatment for the common cold reduces symptoms in 62% of the population. An investigator develops a new treatment and wishes to test $\mathrm{H}_0\colon p = 0.62$ against $\mathrm{H}_0\colon p > 0.62$. The usual test statistic is

$$Z = \frac{\widehat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

where $p_0 = 0.62$. If the new treatment reduces symptoms in 62.1% of a sample of size $n = 3{,}000{,}000$, then the observed value of the test statistic is

$$z_{\text{obs}} = \frac{.621 - .62}{\sqrt{\frac{.62(1-.62)}{3{,}000{,}000}}} = 3.5684.$$

The p-value is $P(Z > 3.5684|\mathrm{H}_0) = 0.00018$. This is strong evidence against H_0, but the effect size is very small.

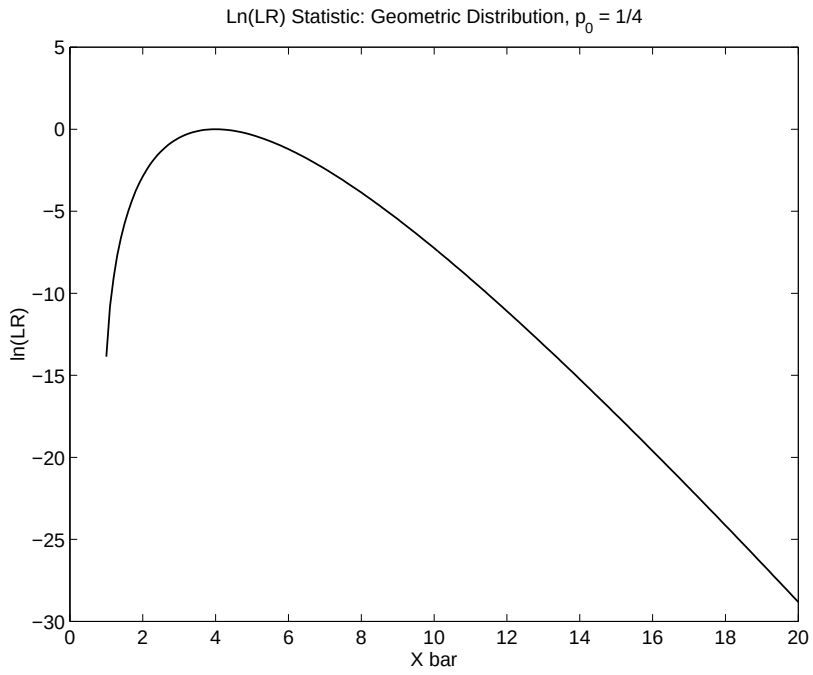
6. Using likelihood to find a test statistic. One way to find a test statistic for testing $\mathrm{H}_0 : \theta = \theta_0$ against a one or two-sided alternative is to examine the likelihood ratio

$$\mathrm{LR} = \frac{L(\theta_0|\mathbf{X})}{\max_{\theta} L(\theta|\mathbf{X})},$$

where the maximization in the denominator is over all θ that satisfy H_a. The LR is ratio of the probability of the data under H_0 to the largest possible probability of the data under H_a. The LR satisfies $\mathrm{LR} \in (0, 1)$. Small values of LR are interpreted to be evidence against H_0. Example: suppose that $\mathrm{H}_0 : p = p_0$ is to be tested against $\mathrm{H}_a : p \neq p_0$ using a random sample from Geom(p). Recall that the MLE of p is $\widehat{p} = 1/\overline{X}$. The likelihood ratio is

$$\mathrm{LR} = \frac{(1-p_0)^{n(\overline{X}-1)} p_0^n}{(1-1/\overline{X})^n \overline{X}^{-n}} = \left(\frac{1-p_0}{\overline{X}-1}\right)^{n(\overline{X}-1)} \overline{X}^{n\overline{X}} p_0^n.$$

In the following display, the log of the LR statistic is plotted against $\overline{X}$ for the special case $n = 10$ and $p_0 = 0.25$.



The above plot reveals that the log likelihood ratio statistic is small if $\overline{X}$ is substantially larger or substantially smaller that 4. Note that $\overline{X} = 4 \Longrightarrow 1/\overline{X} = 0.25 = p_0$ and that the LR statistic is 1 if $\overline{X} = 4$; i.e., the log of the LR statistic is zero.

To use the likelihood ratio as a test statistic, it is necessary to determining its sampling distribution under a true null. This can be quite difficult, but fortunately there is an easy to use large sample result. If θ is a scalar, then under certain regularity conditions, the asymptotic null distribution of $-2\ln(LR)$ is χ^2_1.

10.3 One Sample Z Tests

1. Form of the test Statistic. Suppose that it is of interest to test $\mathrm{H}_0\colon \theta = \theta_0$ against either $\mathrm{H}_a\colon \theta \neq \theta_0$, $\mathrm{H}_a\colon \theta > \theta_0$, or $\mathrm{H}_a\colon \theta < \theta_0$. Further, suppose that we have an estimator of θ that satisfies

$$T|\mathrm{H}_0 \dot{\sim} \mathrm{N}\left(\theta, \frac{\omega_0^2}{n}\right)$$

and an estimator of ω_0^2, say W_0^2 that satisfies $W_0^2 \xrightarrow{\text{prob}} \omega_0^2$ whenever H_0 is true. The subscript on ω is a reminder that the variance of T under H_0 could be a function of θ_0.

2. A reasonable test statistic is

$$Z = \frac{T - \theta_0}{W_0/\sqrt{n}}.$$

If sample size is large, then the distribution of Z under H_0 is approximately $\mathrm{N}(0,1)$.

3. p-values Let z_{obs} be the observed test statistic. Then the p-value for testing H_0 against H_a is

   (a) $1 - P(-|z_{\text{obs}}| \leq Z \leq |z_{\text{obs}}|) = 2\left[1 - \Phi(|z_{\text{obs}}|)\right]$ if the alternative hypothesis is $\mathrm{H}_a\colon \theta \neq \theta_0$;

   (b) $P(z_{\text{obs}} \leq Z) = 1 - \Phi(z_{\text{obs}})$ if the alternative hypothesis is $\mathrm{H}_a\colon \theta > \theta_0$; and

   (c) $P(Z \leq z_{\text{obs}}) = \Phi(z_{\text{obs}})$ if the alternative hypothesis is $\mathrm{H}_a\colon \theta < \theta_0$.

4. Example: $X_1, X_2, \ldots, X_n$ is a random sample from a population with mean μ and variance σ^2. To test $\mathrm{H}_0\colon \mu = \mu_0$ against either $\mathrm{H}_a\colon \mu \neq \mu_0$, $\mathrm{H}_a\colon \mu > \mu_0$, or $\mathrm{H}_a\colon \mu < \mu_0$, use the test statistic

$$Z = \frac{\overline{X} - \mu_0}{S_X/\sqrt{n}}.$$

5. Example: $X_1, X_2, \ldots, X_n$ is a random sample from Bern(p). To test $\mathrm{H}_0\colon p = p_0$ against either $\mathrm{H}_a\colon p \neq p_0$, $\mathrm{H}_a\colon p > p_0$, or $\mathrm{H}_a\colon p < p_0$, use the test statistic

$$Z = \frac{\widehat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Note that $\omega_0^2 = p_0(1-p_0)$.

10.4 One Sample t Tests

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from $\mathrm{N}(\mu, \sigma^2)$. To test $\mathrm{H}_0\colon \mu = \mu_0$ against either $\mathrm{H}_a\colon \mu \neq \mu_0$, $\mathrm{H}_a\colon \mu > \mu_0$, or $\mathrm{H}_a\colon \mu < \mu_0$, use the test statistic

$$T = \frac{\overline{X} - \mu_0}{S_X/\sqrt{n}}.$$

Under H_0, the test statistic has a t distribution with $n-1$ degrees of freedom.

10.5 Some Nonparametric Tests

We will examine only the sign test. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from a continuous distribution having median $\widetilde{\mu}$. A test of $\mathrm{H}_0\colon \widetilde{\mu} = \widetilde{\mu}_0$ against either $\mathrm{H}_a\colon \widetilde{\mu} \neq \widetilde{\mu}_0$, $\mathrm{H}_a\colon \widetilde{\mu} > \widetilde{\mu}_0$, or $\mathrm{H}_a\colon \widetilde{\mu} < \widetilde{\mu}_0$ is desired. Let

$$U_i = I_{(-\infty, \widetilde{\mu}_0]}(X_i) = \begin{cases} 1 & \text{if } X_i \leq \widetilde{\mu}_0; \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Under H_0, $U_i \sim$ iid Bern(0.5) and $Y = \sum_{i=1}^n U_i \sim \mathrm{Bin}(n, 0.5)$. Accordingly, the test statistic

$$Z = \frac{\widehat{p} - 0.5}{\sqrt{.25/n}}$$

is distributed approximately $\mathrm{N}(0,1)$ under H_0, where $\widehat{p} = Y/n$.

10.6 Probability of the Null Hypothesis

When using the frequentist approach, it is not correct to interpret the p-value as the probability that H_0 is true. When using the Bayesian approach, then the posterior probability that H_0 is true can be computed.

Suppose, for example, that $X_i \sim$ iid Bern(θ) for $i = 1, \ldots, n$. Furthermore, suppose that the prior on θ is $\Theta \sim \mathrm{Beta}(\alpha_1, \alpha_2)$. The sufficient statistic is $Y = \sum_{i=1}^n X_i$ and the posterior is $\Theta|(Y = y) \sim \mathrm{Beta}(\alpha_1 + y, \alpha_2 + n - y)$. Suppose that one wants to test $\mathrm{H}_0\colon \theta \leq \theta_0$ against $\mathrm{H}_a\colon \theta > \theta_0$. Then

$$P(\mathrm{H}_0|Y = y) = P(\Theta \leq \theta_0|Y = y) = \int_0^{\theta_0} g_{\Theta|Y}(\theta|y)\, d\theta.$$

For example, if $\alpha_1 = \alpha_2 = 1$, n=40, y=30, and we want to test $\mathrm{H}_0\colon \theta \leq 0.6$ against $\mathrm{H}_a\colon \theta > 0.6$, then $\Theta|(Y = 30) \sim \mathrm{Beta}(31, 11)$ and

$$P(\mathrm{H}_0|Y = 30) = \int_0^{0.6} \frac{\theta^{30}(1-\theta)^{10}}{B(31,11)} d\theta = 0.0274.$$

The test statistic for computing the frequentist p-value is

$$z_{\text{obs}} = \frac{0.75 - 0.6}{\sqrt{\frac{0.6(1-0.6)}{40}}} = 1.9365.$$

The p-value is

$$p\text{-value } = P(Z > 1.9365) = 0.0264.$$

If the correction for continuity is employed, then

$$z_{obs} = \frac{0.75 - \frac{1}{80} - 0.6}{\sqrt{\frac{0.6(0.4)}{40}}} = 1.7751 \text{ and 4-value } = 0.0379.$$

There are some complications if one wants a Bayesian test of $\mathrm{H}_0\colon \theta = \theta_0$ against $\mathrm{H}_a\colon \theta \neq \theta_0$. Your textbook gives one example. We will not have time to discuss this issue.

Chapter 11

TESTS AS DECISION RULES

This chapter introduces the Neyman-Pearson theory of tests as decision rules. In this chapter it is assumed that a decision about H_0 verses H_a must be made. Based on the data, the investigator either will reject H_0 and act as though H_a is true or fail to reject H_0 and act as though H_0 is true. The latter decision is called "accept H_0."

11.1 Rejection Regions and Errors

1. Suppose that the data consist of $X_1, X_2, \ldots, X_n$. The joint sample space of $\mathbf{X}$ is partitioned into two pieces, the rejection region and the acceptance region. In practice, the partitioning is accomplished by using a test statistic.

   (a) Rejection region: the set of values of the test statistic that call for rejecting H_0.

   (b) Acceptance region: the set of values of the test statistic that call for accepting H_0. The acceptance region is the complement of the rejection region.

2. Errors

   (a) Type I: rejecting a true H_0.

   (b) Type II: accepting a false H_0.

3. Size of the Test:

$$\text{size} = \alpha = P(\text{reject } H_0 | H_0 \text{ true}).$$

4. Type II error probability:

$$P(\text{type II error}) = \beta = P(\text{accept } H_0 | H_0 \text{ false}).$$

5. Power:

$$\text{power} = 1 - \beta = P(\text{reject } H_0 | H_0 \text{ false}).$$

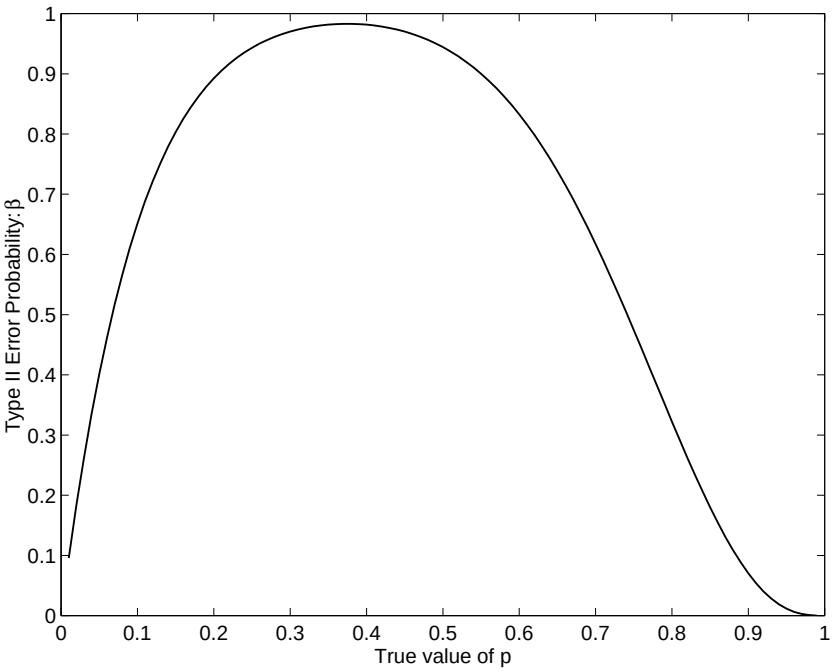
6. Example; Consider a test of $H_0: p = 0.4$ versus $H_a: p \neq 0.4$ based on a random sample of size 10 from Bern(p). Let $Y = \sum X_i$. If the rejection rule is to reject H_0 if $Y \leq 0$ or $Y \geq 8$, then the test size is

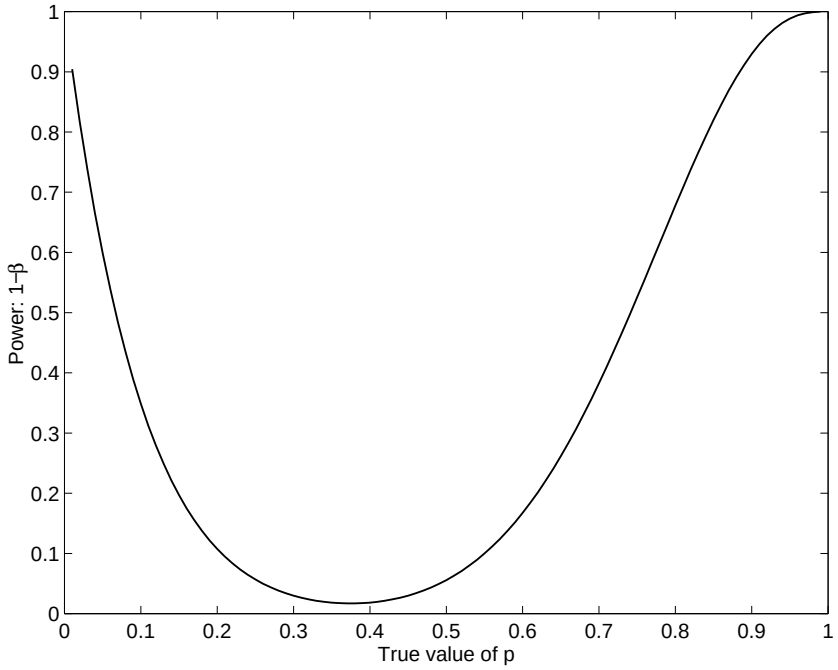
$$\alpha = 1 - P(1 \leq Y \leq 7 | p = 0.4) = 0.0183.$$

The type II error probability and the power depend on the value of p when H_0 is not true. Their values are

$$P(\text{type II error}) = P(1 \leq Y \leq 7|p) \text{ and power} = 1 - P(1 \leq Y \leq 7|p).$$

The two plots below display the type II error probabilities and power for various values of p.





11.2 The Power function

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from $f_X(x|\theta)$. Denote the parameter space of θ by Θ and let Θ_0 and Θ_a be two disjoint subspaces of Θ. Consider a test of $\mathrm{H}_0\colon \theta \in \Theta_0$ against $\mathrm{H}_a\colon \theta \in \Theta_a$. The power function is a function of θ and is defined by

$$\pi(\theta) = P(\text{reject } \mathrm{H}_0|\theta).$$

The function usually is used when $\theta \in \Theta_a$, but the function is defined for all $\theta \in \Theta$.

For example, consider a test of $\mathrm{H}_0\colon \mu = \mu_0$ against $\mathrm{H}_a\colon \mu > \mu_0$ based on a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. Let $\Phi^{-1}(1-\alpha) = z_{1-\alpha}$ be the $100(1-\alpha)$ percentile of the standard normal distribution. Then a one-sample Z test of H_0 will reject H_0 if $Z > z_{1-\alpha}$, where

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

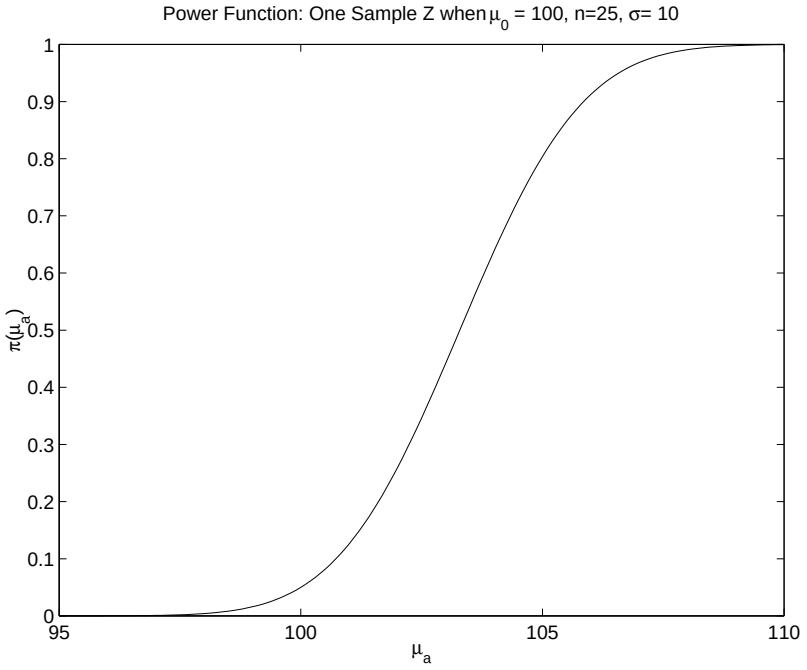
The power function is

$$\begin{aligned}
\pi(\mu_a) &= P(Z > z_{1-\alpha}|\mu = \mu_a) = 1 - P\left[\frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \le z_{1-\alpha}\middle|\, \mu = \mu_a\right] \\
&= 1 - P\left[\frac{\overline{X} - \mu_a + \mu_a - \mu_0}{\sigma/\sqrt{n}} \le z_{1-\alpha}\middle|\, \mu = \mu_a\right] \\
&= 1 - P\left[\frac{\overline{X} - \mu_a}{\sigma/\sqrt{n}} \le z_{1-\alpha} - \frac{\mu_a - \mu_0}{\sigma/\sqrt{n}}\middle|\, \mu = \mu_a\right] \\
&= 1 - \Phi\left[z_{1-\alpha} - \frac{\mu_a - \mu_0}{\sigma/\sqrt{n}}\right].
\end{aligned}$$

As an illustration, if $\sigma = 10$, $\mu_0 = 100$, $n = 25$, and $\alpha = 0.05$, then the power function is

$$\pi(\mu_a) = 1 - \Phi\left[1.645 - \frac{\mu_a - 100}{2}\right].$$

This function is displayed below for various values of μ_a.

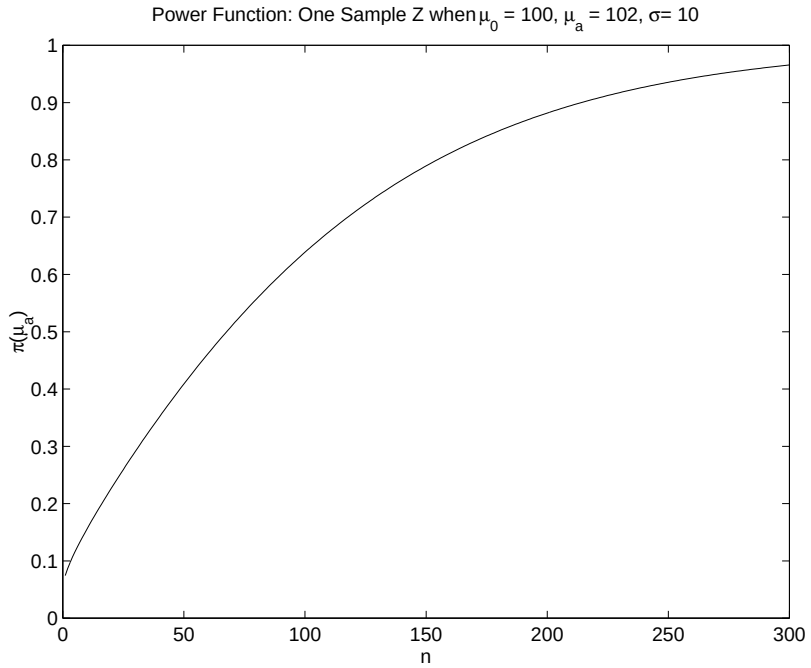


11.3 Choosing a Sample Size

An investigator may want to plan a study so that power will be adequate to detect a meaningful difference. Consider the power function from the last section. Suppose that the investigator decides that the minimal difference of importance is two points. That is, if $\mu_a \geq 102$, then the investigator would like to reject H_0. If μ_a is fixed at $\mu_a = 102$, then the power of the test as a function of n is

$$\pi(\mu_a) = 1 - \Phi\left[1.645 - \frac{102 - 100}{10/\sqrt{n}}\right] = 1 - \Phi\left[1.645 - \frac{\sqrt{n}}{5}\right].$$

This function is plotted below for various values of n.



If the investigator has decided that a specific power is necessary, then the required sample size can be read from the above display.

In general, the required sample size for a one sample Z test of $\mathrm{H}_0\colon \mu = \mu_0$ against $\mathrm{H}_0\colon \mu > \mu_0$ can be obtained by equating the power function to the desired value of $1-\beta$ and solving for n. Denote the 100β percentile of the standard normal distribution by $\Phi^{-1}(\beta) = z_\beta$. That is

$$\begin{aligned}
\pi(\mu_a) &= 1 - \Phi\left[z_{1-\alpha} - \frac{\mu_a - \mu_0}{\sigma/\sqrt{n}}\right] = 1-\beta \\
&\iff \Phi\left[z_{1-\alpha} - \frac{\mu_a - \mu_0}{\sigma/\sqrt{n}}\right] = \beta \\
&\iff z_{1-\alpha} - \frac{\mu_a - \mu_0}{\sigma/\sqrt{n}} = \Phi^{-1}(\beta) = z_\beta \\
&\iff n = \frac{\sigma^2(z_{1-\alpha} - z_\beta)^2}{(\mu_a - \mu_0)^2}.
\end{aligned}$$

For example, if $\mu_0 = 100$, $\mu_a = 102$, $\alpha = 0.05$, $\beta = 0.10$, and $\sigma = 10$, then

$$n = \frac{100(1.645 + 1.282)^2}{(102-100)^2} = 214.18.$$

A sample size of $n = 215$ is required.

11.4 Quality Control

Skip this section.

11.5 Most Powerful tests

1. Definition: Simple Hypothesis. A simple hypothesis is one that completely specifies the joint distribution of the data. That is, there are no unknown parameters under a simple hypothesis. For example $\mathrm{H}_0\colon Y \sim \mathrm{Bin}(25, \frac{1}{3})$ is a simple hypothesis. In this section, we will examine the test of a simple H_0 against a simple H_a.

2. Definition: Most Powerful Test. A test of a simple H_0 versus a simple H_a is a most powerful test of size α if no other test which has size $\leq \alpha$ has greater power.

3. Neyman-Pearson Lemma: Consider the hypotheses $\mathrm{H}_0\colon \mathbf{X} \sim f_0(\mathbf{x})$ against $\mathrm{H}_a\colon \mathbf{X} \sim f_1(\mathbf{x})$, where f_0 and f_1 are the joint pdfs (pmfs) under H_0 and H_a, respectively. Then the most powerful test is to reject H_0 if

$$\Lambda(\mathbf{x}) < K, \text{ where } \Lambda(\mathbf{x}) = \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}$$

is the likelihood ratio. Furthermore, the size of the test is

$$\alpha = \int_R f_0(\mathbf{x})\, d\mathbf{x}, \text{ where } R = \{\mathbf{x}; \Lambda(\mathbf{x}) < K\}.$$

Proof: Consider any other test that has size $\leq \alpha$. Denote the rejection region of the competing test by R^* and denote the power of the competing test by $1 - \beta^*$. Then

$$\int_{R^*} f_0(\mathbf{x})\, d\mathbf{x} = \alpha^* \leq \alpha \text{ and}$$
$$(1-\beta) - (1-\beta^*) = \beta^* - \beta = \int_R f_1(\mathbf{x})\, d\mathbf{x} - \int_{R^*} f_1(\mathbf{x})\, d\mathbf{x}.$$

We will show that the above difference is greater than or equal to zero. Note that $R = (R \cap R^*) \cup (R \cap R^{*c})$ and that $(R \cap R^*)$ is disjoint from $(R \cap R^{*c})$. Similarly, $R^* = (R^* \cap R) \cup (R^* \cap R^c)$ and $(R^* \cap R)$ is disjoint from $(R^* \cap R^c)$. Accordingly,

$$\int_R f_1(\mathbf{x})\, d\mathbf{x} = \int_{R \cap R^*} f_1(\mathbf{x})\, d\mathbf{x} + \int_{R \cap R^{*c}} f_1(\mathbf{x})\, d\mathbf{x},$$
$$\int_{R^*} f_1(\mathbf{x})\, d\mathbf{x} = \int_{R^* \cap R^c} f_1(\mathbf{x})\, d\mathbf{x} + \int_{R^* \cap R} f_1(\mathbf{x})\, d\mathbf{x}, \text{ and}$$
$$\beta^* - \beta = \int_{R \cap R^*} f_1(\mathbf{x})\, d\mathbf{x} + \int_{R \cap R^{*c}} f_1(\mathbf{x})\, d\mathbf{x} - \int_{R^* \cap R} f_1(\mathbf{x})\, d\mathbf{x} - \int_{R^* \cap R^c} f_1(\mathbf{x})\, d\mathbf{x}$$
$$= \int_{R \cap R^{*c}} f_1(\mathbf{x})\, d\mathbf{x} - \int_{R^* \cap R^c} f_1(\mathbf{x})\, d\mathbf{x}.$$

Note that $(R \cap R^*) \in R$ so that $f_1(\mathbf{x}) > K^{-1} f_0(\mathbf{x})$ in the first integral. Also, $(R^* \cap R^c) \in R^c$ so that $f_1(\mathbf{x}) < K^{-1} f_0(\mathbf{x})$ in the second integral. Therefore,

$$\begin{aligned}
\beta^* - \beta &\geq \frac{1}{K} \int_{R \cap R^{*c}} f_0(\mathbf{x})\, d\mathbf{x} - \frac{1}{K} \int_{R^* \cap R^c} f_0(\mathbf{x})\, d\mathbf{x} \\
&= \left[\frac{1}{K} \int_{R \cap R^{*c}} f_0(\mathbf{x})\, d\mathbf{x} + \frac{1}{K} \int_{R \cap R^*} f_0(\mathbf{x})\, d\mathbf{x} \right] \\
&- \left[\frac{1}{K} \int_{R^* \cap R^c} f_0(\mathbf{x})\, d\mathbf{x} + \frac{1}{K} \int_{R^* \cap R} f_0(\mathbf{x})\, d\mathbf{x} \right] \text{ by adding zero} \\
&= \frac{1}{K} \int_R f_0(\mathbf{x})\, d\mathbf{x} - \frac{1}{K} \int_{R^*} f_0(\mathbf{x})\, d\mathbf{x} = \frac{1}{K} (\alpha - \alpha^*) \geq 0
\end{aligned}$$

because the size of the competing test is $\alpha^* \leq \alpha$.

4. Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from $\text{NegBin}(k, \theta)$, where k is known. Find the most powerful test of $\text{H}_0\colon \theta = \theta_0$ against $\text{H}_a\colon \theta = \theta_a$, where $\theta_a > \theta_0$. Solution: The likelihood ratio test statistic is

$$\begin{aligned}
\Lambda(\mathbf{x}) &= \frac{\displaystyle\prod_{i=1}^{n} \binom{x_i - 1}{k-1} \theta_0^k (1-\theta_0)^{x_i - k}}{\displaystyle\prod_{i=1}^{n} \binom{x_i - 1}{k-1} \theta_a^k (1-\theta_a)^{x_i - k}} \\
&= \frac{\theta_0^{nk} (1-\theta_0)^{n(\bar{x}-k)}}{\theta_a^{nk} (1-\theta_a)^{n(\bar{x}-k)}} \\
&= \left(\frac{\theta_0 (1-\theta_a)}{\theta_a (1-\theta_0)} \right)^{nk} \left(\frac{1-\theta_0}{1-\theta_a} \right)^{n(\bar{x}-k)}.
\end{aligned}$$

Note that the likelihood ratio test statistic depends on the data solely through $\bar{x}$ and that

$$\theta_a > \theta_0 \Longrightarrow \frac{1-\theta_0}{1-\theta_a} > 1.$$

Accordingly $\Lambda(\mathbf{x})$ is an increasing function of $\bar{x}$. Rejecting H_0 for small values of $\Lambda(\mathbf{x})$ is equivalent to rejecting H_0 for small values of $\bar{x}$ and the most powerful test is to reject H_0 if $\bar{x} < K^*$ where K^* is determined by the relation

$$P(\overline{X} < K^* | \theta = \theta_0) \leq \alpha.$$

The above probability can be evaluated without too much difficulty because

$$n\overline{X} = \sum_{i=1}^{n} X_i \sim \text{ NegBin}(nk, \theta_0)$$

under H_0.

11.6 Randomized Tests

Skip this section.

11.7 Uniformly Most Powerful tests

1. Consider the problem of testing $H_0\colon \theta = \theta_0$ against $H_0\colon \theta > \theta_0$ based on a random sample of size n from $f_X(x|\theta)$.

2. Uniformly Most Powerful (UMP) Test. Definition: If a test of $H_0\colon \theta = \theta_0$ against $H_a\colon \theta = \theta_a$, is a most powerful test for every $\theta_a > \theta_0$ among all tests with size $\leq \alpha$, then the test is uniformly most powerful for testing $H_0\colon \theta = \theta_0$ against $H_a\colon \theta > \theta_0$.

3. Approach to finding a UMP test. First use the Neyman-Pearson Lemma to find a most powerful test of $H_0\colon \theta = \theta_0$ against $H_a\colon \theta = \theta_a$ for some $\theta_a > \theta_0$. If the form of the test is the same for all θ_a, then the test is UMP.

4. Example: Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from NegBin(k, θ), where k is known. Find the UMP test of $H_0\colon \theta = \theta_0$ against $H_a\colon \theta > \theta_0$. Solution: The most powerful test of $H_0\colon \theta = \theta_0$ against $H_a\colon \theta = \theta_a$, where $\theta_a > \theta_0$ is to reject H_0 if $\bar{x} < K^*$ where K^* is determined by the relation

$$P(\overline{X} < K^* | \theta = \theta_0) \leq \alpha.$$

   Note that the form of the test does not depend on the particular value of θ_a. Accordingly, the test that rejects H_0 for small values of $\bar{x}$ is the UMP test of $H_0\colon \theta = \theta_0$ against $H_a\colon \theta > \theta_0$.

5. A similar argument shows that in the negative binomial example, the UMP test of $H_0\colon \theta = \theta_0$ against $H_a\colon \theta < \theta_0$ is to reject H_0 for large values of $\bar{x}$.

6. The UMP idea can be extended to tests of $H_0\colon \theta \leq \theta_0$ against $H_a\colon \theta > \theta_0$. If the power function is monotonic in θ, then the UMP test of $H_0\colon \theta = \theta_0$ against $H_a\colon \theta > \theta_0$ also is UMP for testing $H_0\colon \theta \leq \theta_0$ against $H_a\colon \theta > \theta_0$. The size of the test is

$$\alpha = \sup_{\theta \leq \theta_0} P(\text{reject } H_0 | \theta) = \sup_{\theta \leq \theta_0} \pi(\theta) = P(\text{reject } H_0 | \theta_0)$$

   because the power function is monotonic It can be shown that the power function is monotone in θ if the distribution of X belongs to the one parameter exponential family.

11.8 Likelihood Ratio Tests

1. Consider the problem of testing $\mathrm{H}_0\colon \boldsymbol{\theta} \in \Theta_0$ against $\mathrm{H}_a\colon \boldsymbol{\theta} \in \Theta_a$, where Θ_0 and Θ_a are disjoint subspaces of the parameter space. The parameter $\boldsymbol{\theta}$ may be a vector.

2. The generalized likelihood ratio test of H_0 against H_a is to reject H_0 for small values of the likelihood ratio test statistic

$$\begin{aligned} \Lambda(\mathbf{X}) &= \frac{L(\widehat{\boldsymbol{\theta}}_0|\mathbf{X})}{L(\widehat{\boldsymbol{\theta}}_a|\mathbf{X})}, \text{ where} \\ L(\widehat{\boldsymbol{\theta}}_0|\mathbf{X}) &= \sup_{\boldsymbol{\theta}\in\Theta_0} L(\boldsymbol{\theta}|\mathbf{X}) \text{ and } L(\widehat{\boldsymbol{\theta}}_a|\mathbf{X}) = \sup_{\boldsymbol{\theta}\in\Theta_0\cup\Theta_a} L(\boldsymbol{\theta}|\mathbf{X}). \end{aligned}$$

   That is, the likelihood function is maximized twice; first under the null, and second under the union of the null and alternative.

3. Properties of $\Lambda(\mathbf{X})$.

   (a) $\Lambda(\mathbf{X}) \in [0, 1]$.

   (b) Small values of Λ are evidence against H_0.

   (c) Under mild regularity conditions, the asymptotic null distribution of $-2\ln[\Lambda(\mathbf{X})]$ is χ^2 with degrees of freedom equal to the number of restrictions under H_0 minus the number of restrictions under H_a.

   (d) The decision rule is to reject H_0 for large values of $-2\ln(\Lambda)$.

4. Example 1: Suppose that $X_{ij} \overset{\text{ind}}{\sim} \text{Bern}(p_i)$ for $j = 1, \ldots, n_i$. That is, we have independent samples from each of two Bernoulli populations. Consider the problem of testing $\mathrm{H}_0\colon p_1 = p_2$ against $\mathrm{H}_a\colon p_1 \neq p_2$. The sufficient statistics are $Y_1 = \sum_{j=1}^{n_1} X_{1j}$ and $Y_2 = \sum_{j=1}^{n_2} X_{2j}$. These statistics are independently distributed as $Y_i \sim \text{Bin}(n_i, p_i)$. The likelihood function is

$$L(p_1, p_2|y_1, y_2) = p_1^{y_1}(1-p_1)^{n_1-y_1} p_2^{y_2}(1-p_2)^{n_2-y_2}.$$

   Under H_0, the MLE of the common value $p = p_1 = p_2$ is $\widehat{p} = (y_1 + y_2)/(n_1 + n_2)$. Under the union of H_0 and H_a, there are no restrictions on p_1 and p_2 and the MLEs are $\widehat{p}_1 = y_1/n_1$ and $\widehat{p}_2 = y_2/n_2$. The likelihood ratio test statistic is

$$\Lambda(y_1, y_2) = \frac{\widehat{p}^{\,y_1+y_2}(1-\widehat{p})^{n_1+n_2-y_1-y_2}}{\widehat{p}_1^{\,y_1}(1-\widehat{p}_1)^{n_1-y_1}\widehat{p}_2^{\,y_2}(1-\widehat{p}_2)^{n_2-y_2}}.$$

   If H_0 is true and sample sizes are large, then $-2\ln[\Lambda(Y_1, Y_2)]$ is approximately distributed as a χ^2 random variable. There are no restrictions under H_a and one restriction under H_0, so the χ^2 random variable has 1 degree of freedom. For example, if $n_1 = 30$, $n_2 = 40$, $y_1 = 20$, and $y_2 = 35$, then

$\Lambda(y_1, y_2) = 0.1103$, $-2\ln(\Lambda) = 4.4087$, and the p-value is 0.0358. For comparison, the familiar large sample test statistic is

$$z = \frac{\widehat{p}_1 - \widehat{p}_2}{\sqrt{\widehat{p}(1-\widehat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 2.1022$$

and the p-value is 0.0355. Note, $Z^2 \sim \chi^2_1$ and $z^2 = 4.4192$ which is very close to the LR test statistic.

5. Example 2: Your textbook (page 476) shows that the usual one-sample t test of $\mathrm{H}_0\colon \mu = \mu_0$ against $\mathrm{H}_a\colon \mu \neq \mu_0$ when sampling from a normal distribution with unknown variance is the likelihood ratio test. Below is another version of the proof.

   (a) Lemma: Let a be a constant or a variable that does not depend on σ^2 and let n be a positive constant. Then, the maximizer of

   $$h(\sigma^2; a) = \frac{e^{-\frac{1}{2\sigma^2}a}}{\left(\sigma^2\right)^{\frac{n}{2}}}$$

   with respect to σ^2 is

   $$\hat{\sigma}^2 = \frac{a}{n}.$$

   Also,

   $$\max_{\sigma^2} h(\sigma^2; a) = h\left(\frac{a}{n}, a\right) = \frac{e^{-\frac{n}{2}}}{\left(\frac{a}{n}\right)^{\frac{n}{2}}}.$$

   Proof: The natural log of h is

   $$\ln(h) = -\frac{1}{2\sigma^2}a - \frac{n}{2}\ln(\sigma^2).$$

   Equating the first derivative of $\ln(h)$ to zero and solving for σ^2 yields

   $$\begin{aligned} \frac{\partial}{\partial\,\sigma^2}\ln(h) &= \frac{1}{2\sigma^4}a - \frac{n}{2\sigma^2}, \\ \frac{\partial}{\partial\,\sigma^2}\ln(h) = 0 &\Longrightarrow \frac{a}{\sigma^2} - n = 0 \\ &\Longrightarrow \sigma^2 = \frac{a}{n}. \end{aligned}$$

   The solution is a maximizer because the second derivative of $\ln(h)$ evaluated at $\sigma^2 = a/n$ is negative:

   $$\begin{aligned} \frac{\partial^2}{(\partial\,\sigma^2)^2}\ln(h) &= \frac{\partial}{\partial\,\sigma^2}\left(\frac{a}{2\sigma^4} - \frac{n}{2\sigma^2}\right) = -\frac{a}{\sigma^6} + \frac{n}{2\sigma^4}, \\ \frac{\partial^2}{(\partial\,\sigma^2)^2}\ln(h)\bigg|_{\sigma^2=a/n} &= -\frac{a}{(a/n)^3} + \frac{n}{2(a/n)^2} = -\frac{n^3}{a^2} < 0. \end{aligned}$$

   To obtain the maximum, substitute a/n for σ^2 in $h(\sigma^2, a)$.

(b) Theorem: The generalized likelihood ratio test of $\mathrm{H}_0\colon \mu = \mu_0$ against $\mathrm{H}_a\colon \mu \neq \mu_0$ based on a random sample of size n from a normal distribution with unknown variance is to reject H_0 for large $|T|$, where

$$T = \frac{\sqrt{n}(\overline{X} - \mu_0)}{S_X},$$
$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \text{ and } S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Proof: The likelihood function of μ and σ^2 given $X_i \overset{\text{iid}}{\sim} \mathrm{N}(\mu, \sigma^2)$ for $i = 1, \ldots, n$ is

$$L(\mu, \sigma^2|\mathbf{X}) = \frac{\exp\left\{-\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^{n}(X_i - \mu)^2\right\}}{\left(\sigma^2\right)^{\frac{n}{2}}(2\pi)^{\frac{n}{2}}}.$$

Under $\mathrm{H}_0\colon \mu = \mu_0$, the likelihood function is

$$L(\mu_0, \sigma^2|\mathbf{X}) = \frac{\exp\left\{-\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^{n}(X_i - \mu_0)^2\right\}}{\left(\sigma^2\right)^{\frac{n}{2}}(2\pi)^{\frac{n}{2}}}.$$

Using the Lemma with

$$a = \sum_{i=1}^{n}(X_i - \mu_0)^2$$

yields

$$\hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_0)^2 \text{ and}$$
$$\max_{\sigma^2} L(\mu_0, \sigma^2|\mathbf{X}) = L(\mu_0, \hat{\sigma}_0^2|\mathbf{X}) = \frac{e^{-\frac{n}{2}}}{\left(\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(X_i - \mu_0)^2\right)^{\frac{n}{2}}(2\pi)^{\frac{n}{2}}}.$$

Under $\mathrm{H}_0 \cup \mathrm{H}_a$, the likelihood function must be maximized with respect to μ and σ^2. Note that the sign of the exponent in the numerator of L is negative. Accordingly, to maximize L with respect to μ, we must minimize

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

with respect to μ. By the parallel axis theorem, it is known that the minimizer is $\mu = \overline{X}$. Substitute $\overline{X}$ for μ in the likelihood function and use the Lemma with

$$a = \sum_{i=1}^{n}(X_i - \overline{X})^2$$

to obtain

$$\hat{\sigma}_a^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \text{ and}$$

$$\max_{\sigma^2,\mu} L(\mu, \sigma^2|\mathbf{X}) = L(\overline{X}, \hat{\sigma}_a^2|\mathbf{X}) = \frac{e^{-\frac{n}{2}}}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^{\frac{n}{2}}(2\pi)^{\frac{n}{2}}}.$$

The likelihood ratio test statistic is

$$\Lambda = \frac{L(\mu_0, \hat{\sigma}_0^2)}{L(\overline{X}, \hat{\sigma}_a^2)} = \left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \mu_0)^2}\right)^{\frac{n}{n}}.$$

Recall, that from the parallel axis theorem,

$$\sum_{i=1}^{n}(X_i - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2.$$

Accordingly, the Likelihood Ratio Test (LRT) is to reject H_0 for small Λ, where

$$\Lambda = \left(\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2 + n(\overline{X} - \mu_0)^2}\right)^{\frac{n}{2}}.$$

Any monotonic transformation of Λ also can be used as the LRT statistic. In particular,

$$\left(\Lambda^{-\frac{2}{n}} - 1\right)(n-1) = \frac{n(\overline{X} - \mu_0)^2}{S_X^2} = T^2$$

is a decreasing function of Λ. Therefore, the LRT rejects H_0 for large T^2 or, equivalently, for large $|T|$.

11.9 Bayesian Testing

1. To develop a Bayes Test, first make the following definitions:

   (a) Loss Function: $\ell(\theta, act)$ = loss incurred if action *act* is performed when the state of nature is θ. The action *act* will be either "reject H_0" or "accept H_0." For example, $\ell(H_0, \text{reject } H_0)$ is the loss incurred when a true H_0 is rejected. It is assumed that making the correct action incurs no loss.

   (b) Parameter Space: Denote the support of θ under H_0 by Θ_0 and denote the support of θ under H_a by Θ_a. For example, if the hypotheses are $H_0\colon \mu \leq 100$ and $H_a\colon \mu > 100$, then, $\Theta_0 = (\infty, 100]$ and $\Theta_a = (100, \infty)$.

   (c) Prior: Before new data are collected, the prior pdf or pmf for θ is denoted by $g(\theta)$.

   (d) Posterior: After new data have been collected, the posterior pdf or pmf for θ is denoted by $g(\theta|\mathbf{X})$.

   (e) Bayes Loss: The posterior Bayes Loss for action *act* is

$$B(act\,|\mathbf{X}) = \mathrm{E}_\theta\left[\ell(\theta, act)\right]$$
$$= \begin{cases} \displaystyle\int_{\Theta_0} \ell(\theta, \text{reject } H_0) g(\theta|\mathbf{X})\, d\theta & \text{if } act = \text{reject } H_0, \\ \displaystyle\int_{\Theta_a} \ell(\theta, \text{accept } H_0) g(\theta|\mathbf{X})\, d\theta & \text{if } act = \text{accept } H_0. \end{cases}$$

   (f) Bayes Test: A Bayes test is the rule that minimizes Bayes Loss.

2. Theorem: When testing a simple null against a simple alternative, the Bayes test is a Neyman-Pearson test and a Neyman-Pearson rejection region, $f_0/f_a < K$, corresponds to a Bayes test for some prior.

   Proof: The null and alternative can be written as $H_0\colon f(\mathbf{x}) = f_0(\mathbf{x})$ versus $H_1\colon f(\mathbf{x}) = f_1(\mathbf{x})$. Also, the support of θ has only two points:

$$\Theta = \{f_0, f_1\}, \quad \Theta_0 = \{f_0\}, \text{ and } \Theta_a = \{f_1\} \text{ or, equivalently,}$$
$$\Theta = \{H_0, H_1\}, \quad \Theta_0 = \{H_0\}, \text{ and } \Theta_a = \{H_1\}.$$

   Denote the prior probabilities of θ as

$$g_0 = P(H_0) \text{ and } g_1 = P(H_1).$$

   The posterior probabilities are

$$f(H_0|\mathbf{x}) = \frac{f(H_0, \mathbf{x})}{f(\mathbf{x})} = \frac{f(\mathbf{x}|H_0) f(H_0)}{f(\mathbf{x}|H_0) f(H_0) + f(\mathbf{x}|H_1) f(H_1)}$$
$$= \frac{f_0(\mathbf{x}) g_0}{f_0(\mathbf{x}) g_0 + f_1(\mathbf{x}) g_1} \text{ and}$$

$$f(\mathrm{H}_1|\mathbf{x}) = \frac{f(\mathrm{H}_1, \mathbf{x})}{f(\mathbf{x})} = \frac{f(\mathbf{x}|\mathrm{H}_1)f(\mathrm{H}_1)}{f(\mathbf{x}|\mathrm{H}_0)f(\mathrm{H}_0) + f(\mathbf{x}|\mathrm{H}_1)f(\mathrm{H}_1)}$$
$$= \frac{f_1(\mathbf{x})g_1}{f_0(\mathbf{x})g_0 + f_1(\mathbf{x})g_1}.$$

Denote the losses for incorrect decisions by

$$\ell(\mathrm{H}_0, \text{reject } \mathrm{H}_0) = \ell_0 \text{ and } \ell(\mathrm{H}_1, \text{accept } \mathrm{H}_0) = \ell_1.$$

Note that ℓ_0 and ℓ_1 are merely scalar constants. Then, the posterior Bayes losses are

$$B(\text{reject } \mathrm{H}_0|\mathbf{x}) = \ell_0 \times \left(\frac{f_0(\mathbf{x})g_0}{f_0(\mathbf{x})g_0 + f_1(\mathbf{x})g_1}\right) \text{ and}$$
$$B(\text{accept } \mathrm{H}_0|\mathbf{x}) = \ell_1 \times \left(\frac{f_1(\mathbf{x})g_1}{f_0(\mathbf{x})g_0 + f_1(\mathbf{x})g_1}\right).$$

The Bayes test consists of choosing the action that has the smallest Bayes loss. Alternatively, the ratio of Bayes losses can be examined:

$$\frac{B(\text{reject } \mathrm{H}_0|\mathbf{x})}{B(\text{accept } \mathrm{H}_0|\mathbf{x})} = \frac{\ell_0 f_0(\mathbf{x})g_0}{\ell_1 f_1(\mathbf{x})g_1}.$$

If the ratio is smaller than 1, then the Bayes test is to reject H_0, otherwise accept H_0. That is, H_0 is rejected if

$$\frac{\ell_0 f_0(\mathbf{x})g_0}{\ell_1 f_1(\mathbf{x})g_1} < 1 \text{ or, equivalently,}$$
$$\frac{f_0(\mathbf{x})}{f_1(\mathbf{x})} < K, \text{ where } K = \frac{\ell_1 g_1}{\ell_0 g_0}.$$

Accordingly, the Bayes test is a Neyman-Pearson test. Also, a Neyman-Pearson rejection region, $f_0/f_1 < K$, corresponds to a Bayes test, where the priors and losses satisfy

$$K = \frac{\ell_1 g_1}{\ell_0 g_0}.$$

3. Example 11.9a (with details) A machine that fills bags with flour is adjusted so that the mean weight in a bag is 16 ounces. To determine whether the machine is at the correct setting, a sample of bags can be weighed. There is a constant cost for readjusting the machine. The cost is due to shutting down the production line, etc. If the machine is not adjusted, then the company may be over-filling the bags with cost $2(\mu - 16)$ or under-filling the bags with cost $16 - \mu$. The under-filling cost is due to customer dissatisfaction.

   Consider testing $\mathrm{H}_0\colon \mu \leq 16$ against $\mathrm{H}_a\colon \mu > 16$ based on a random sample of size n from $\mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. Furthermore, suppose that the prior

on μ is $\mathrm{N}(\nu, \tau^2)$. Using the result in Example 2 on page 101 of these notes, the posterior distribution of μ is normal with

$$\mathrm{E}(\mu|\bar{x}) = \left(\frac{n\tau^2}{n\tau^2+\sigma^2}\right)\bar{x} + \left(1 - \frac{n\tau^2}{n\tau^2+\sigma^2}\right)\nu \text{ and}$$
$$\mathrm{Var}(\mu|\bar{x}) = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)^{-1}.$$

If $n = 5$, $\bar{x} = 16.4$, $\sigma^2 = 0.05^2$, $\nu = 16$, and $\tau^2 = 0.002$, then the posterior distribution of μ is $\mathrm{N}(16.32, 0.02^2)$. Suppose that the loss functions are

$$\ell(\mu, \text{reject } \mathrm{H}_0) = \begin{cases} 2(16-\mu) & \text{if } \mu \le 16 \\ \mu - 16 & \text{if } \mu > 16, \end{cases}$$
$$\text{and } \ell(\mu, \text{accept } \mathrm{H}_0) = \ell_1.$$

The Bayes losses are

$$B(\text{reject } \mathrm{H}_0|\mathbf{x}) = \mathrm{E}_{\mu|\mathbf{x}}\left[\ell(\mu, \text{reject } \mathrm{H}_0)\right]$$
$$= \ell_1 \times P(\text{reject } \mathrm{H}_0|\mu \neq 16, \mathbf{x}) + \ell_1 \times P(\text{reject } \mathrm{H}_0|\mu = 16, \mathbf{x}) = \ell_1, \text{ and}$$
$$B(\text{accept } \mathrm{H}_0|\mathbf{x}) = \mathrm{E}_{\mu|\mathbf{x}}\left[\ell(\mu, \text{accept } \mathrm{H}_0)\right]$$
$$= \int_{-\infty}^{16} 2(16-\mu) f_{\mu|\mathbf{x}}(\mu|\mathbf{x})\, d\mu + \int_{16}^{\infty} (\mu - 16) f_{\mu|\mathbf{x}}(\mu|\mathbf{x})\, d\mu.$$

The latter integral can be computed as follows. Denote the conditional mean and variance of μ as $\mu_{\mu|\mathbf{x}}$ and $\sigma^2_{\mu|\mathbf{x}}$. That is,

$$\mu_{\mu|\mathbf{x}} = \mathrm{E}(\mu|\mathbf{x}) \text{ and } \sigma^2_{\mu|\mathbf{x}} = \mathrm{Var}(\mu|\mathbf{x}).$$

Transform from μ to

$$z = \frac{\mu - \mu_{\mu|\mathbf{x}}}{\sigma_{\mu|\mathbf{x}}}.$$

Denote the pdf of the standard normal distribution as $\varphi(z)$. Then,

$$\mu = z\sigma_{\mu|\mathbf{x}} + \mu_{\mu|\mathbf{x}},$$
$$d\mu = \sigma_{\mu|\mathbf{x}}\, dz, \text{ and}$$
$$B(\text{accept } \mathrm{H}_0|\mathbf{x}) = \int_{-\infty}^{(16-\mu_{\mu|\mathbf{x}})/\sigma_{\mu|\mathbf{x}}} 2(16 - \sigma_{\mu|\mathbf{x}} z - \mu_{\mu|\mathbf{x}})\varphi(z)\, dz$$
$$+ \int_{(16-\mu_{\mu|\mathbf{x}})/\sigma_{\mu|\mathbf{x}}}^{\infty} (\sigma_{\mu|\mathbf{x}} z + \mu_{\mu|\mathbf{x}} - 16)\varphi(z)\, dz$$
$$= -\int_{-\infty}^{-16} 2(0.32 + 0.02z)\varphi(z)\, dz + \int_{-16}^{\infty} (0.02z + 0.32))\varphi(z)\, dz$$
$$\approx \int_{-16}^{\infty} (0.02z + 0.32))\varphi(z)\, dz$$

$$\approx \mathrm{E}(0.02Z + 0.32) = 0.032,$$

because essentially the entire standard normal distribution lies in the interval $(-16, \infty)$ and essentially none of the distribution lies in the interval $(-\infty, -16)$. The Bayes test rejects H_0 if $\ell_1 \leq 0.32$ and accepts H_0 if $\ell_1 > 0.32$.

A Matlab program to compute the integrals together with the program output are listed below.

```
n=5;
xbar=16.4;
sigma2=.05^2;
nu=16;
tau2=0.002;
w=n*tau2/(n*tau2+sigma2);
m=w*xbar+(1-w)*nu;
v2=(n/sigma2 + 1/tau2)^(-1);
v=sqrt(v2);
disp(['Conditional Mean and SD of mu are'])
disp([m v])
g1 = inline('2*(16-z*s-m).*normpdf(z)', 'z','s','m');
g2 = inline('(z*s+m-16).*normpdf(z)','z','s','m');

z0=(16-m)/v;
tol=1.e-10;
Integral_1=quadl(g1,-30,z0,tol,[],v,m)
Integral_2=quadl(g2,z0,30,tol,[],v,m)
Bayes_Loss = Integral_1+Integral_2

Conditional Mean and SD of mu are  16.3200    0.0200
Integral_1 =  5.6390e-67
Integral_2 =  0.3200
Bayes_Loss =  0.3200
```

4. Example, Problem 11-33 (with details): The goal is to conduct a Bayes Test of $\mathrm{H}_0\colon p \leq \frac{1}{2}$ against $\mathrm{H}_a\colon p > \frac{1}{2}$ based on a random sample of size n from Bern(p). The losses are

$$\ell(p, act) = 0 \text{ if the correct decision is made}$$
$$\ell(\mathrm{H}_0, \text{reject } \mathrm{H}_0) = \ell_0, \text{ and}$$
$$\ell(\mathrm{H}_a, \text{accept } \mathrm{H}_0) = \ell_1.$$

The prior on p is Beta(α, β). Using the results in example 1 on page 100 of these notes, the posterior distribution of p conditional on $\mathbf{x}$ is Beta$(\alpha + y, \beta + n - y)$, where y is the observed number of successes on the n Bernoulli trials.

The Bayes losses are

$$\begin{aligned}
B(\text{reject } \mathrm{H}_0|\mathbf{x}) &= \mathrm{E}_{p|\mathbf{x}}\left[\ell(p, \text{reject } \mathrm{H}_0)\right] \\
&= \int_{\mathrm{H}_0} \ell_0 \frac{p^{\alpha+y-1}(1-p)^{\beta+n-y-1}}{B(\alpha+y, \beta+n-y)}\, dp \\
&= \ell_0 \int_0^{0.5} \frac{p^{\alpha+y-1}(1-p)^{\beta+n-y-1}}{B(\alpha+y, \beta+n-y)}\, dp \\
&= \ell_0 \times P\left(p \leq 0.5|\mathbf{x}\right) \text{ and} \\
B(\text{accept } \mathrm{H}_0|\mathbf{x}) &= \mathrm{E}_{p|\mathbf{x}}\left[\ell(p, \text{accept } \mathrm{H}_0)\right] \\
&= \int_{\mathrm{H}_a} \ell_1 \frac{p^{\alpha+y-1}(1-p)^{\beta+n-y-1}}{B(\alpha+y, \beta+n-y)}\, dp \\
&= \ell_1 \int_{0.5}^1 \frac{p^{\alpha+y-1}(1-p)^{\beta+n-y-1}}{B(\alpha+y, \beta+n-y)}\, dp \\
&= \ell_1 \times P\left(p > 0.5|\mathbf{x}\right).
\end{aligned}$$

The required probabilities can be computed using any computer routine that calculates the CDF of a beta distribution.

If $n = 10$, $y = 3$, $\alpha = 7$, $\beta = 3$, $\ell_0 = 3$, $\ell_1 = 2$, then the posterior distribution of p is Beta$(10, 10)$ and the Bayes Losses are

$$B(\text{reject } \mathrm{H}_0|\mathbf{x}) = 3P(W \leq 0.5) \text{ and } B(\text{accept } \mathrm{H}_0|\mathbf{x}) = 2P(W > 0.5),$$

where $W \sim$ Beta$(10, 10)$. This beta distribution is symmetric around 0.5 and, therefore, each of the above probabilities is $\frac{1}{2}$. The Bayes test is to accept H_0 because the Bayes loss is 1, whereas the Bayes loss for rejection is 1.5.

Appendix A

GREEK ALPHABET

Name	Lower Case	Upper Case
Alpha	α	A
Beta	β	B
Gamma	γ	Γ
Delta	δ	Δ
Epsilon	ϵ or ε	E
Zeta	ζ	Z
Eta	η	H
Theta	θ or ϑ	Θ
Iota	ι	I
Kappa	κ	K
Lambda	λ	Λ
Mu	μ	M
Nu	ν	N
Xi	ξ	Ξ
Omicron	o	O
Pi	π	Π
Rho	ρ or ϱ	P
Sigma	σ or ς	Σ
Tau	τ	T
Upsilon	υ	Υ
Phi	ϕ or φ	Φ
Chi	χ	X
Psi	ψ	Ψ
Omega	ω	Ω

Appendix B

ABBREVIATIONS

- BF: Bayes Factor. If H is a hypothesis and $\mathbf{T}$ is a sufficient statistic, then
$$\text{BF} = \frac{\text{Posterior odds of } H}{\text{Prior odds of } H} = \frac{P(H|\mathbf{T}=\mathbf{t})/P(H^c|\mathbf{T}=\mathbf{t})}{P(H)/P(H^c)} = \frac{f_{\mathbf{T}|H}(\mathbf{t}|H)}{f_{\mathbf{T}|H^c}(\mathbf{t}|H^c)}.$$

- CDF or cdf: Cumulative Distribution Function. If X is a random variable, then
$$F_X(x) = P(X \leq x)$$
is the cdf of X.

- CLT: Central Limit Theorem. If $X_1, X_2, \ldots, X_n$ is a random sample of size n from a population with mean μ_X and variance σ_X^2, then, the distribution of
$$Z_n = \frac{\overline{X} - \mu_X}{\sigma_X/\sqrt{n}}$$
converges to $\text{N}(0,1)$ as $n \to \infty$.

- CRLB: Cramér-Rao Lower Bound. The CRLB is the lower bound on the variance of an unbiased estimator of $g(\theta)$. The bound is
$$\text{CRLB } = \frac{\left[\frac{\partial g(\theta)}{\partial \theta}\right]^2}{I_\theta},$$
where I_θ is Fisher's information.

- LR: Likelihood Ratio. When testing a simple null against a simple alternative, the LR is
$$\Lambda = \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}.$$
When testing a composite null against a composite alternative, the LR is
$$\Lambda = \frac{\sup\limits_{\theta \in \Theta_0} f(\mathbf{x}|\theta)}{\sup\limits_{\theta \in \Theta_a} f(\mathbf{x}|\theta)},$$
where Θ_0 and Θ_a are the parameter spaces under H_0 and H_a, respectively.

- LRT: Likelihood Ratio Test. The LRT of H_0 versus H_a is to reject H_0 for small values of the LR. The critical value is chosen so that the size of the test is α.
- MGF or mgf: Moment Generating Function. If X is a random variable, then

$$\psi_X(t) = \mathrm{E}\left(e^{tX}\right)$$

  is the mgf of X.
- MLE: Maximum Likelihood Estimator. Suppose that $X_i, X_2, \ldots, X_n$ is a random sample from $f_X(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a $k \times 1$ vector of parameters. A maximum likelihood estimator of $\boldsymbol{\theta}$ is any value $\widehat{\boldsymbol{\theta}}$ that maximizes the likelihood function and is a point in the parameter space or on the boundary of the parameter space.
- MSE: Mean Square Error. If T is an estimator of a parameter, θ, then

$$MSE_T(\theta) = \mathrm{E}(T-\theta)^2 = \sigma_T^2 + \text{ bias}^2,$$

  where bias $= \mathrm{E}(T-\theta)$.
- PDF or pdf: Probability Density Function. If X is a continuous random variable, then

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

  is the pdf of X.
- PF or pf: Probability Function. If X is a discrete random variable, then

$$P(X = x) = f_X(x)$$

  is the pf of X. The terms pf and pmf are interchangeable.
- PGF or pgf: Probability Generating Function. If X is a random variable, then

$$\eta_X(t) = \mathrm{E}\left(t^X\right)$$

  is the pgf of X. The pgf is most useful for discrete random variables.
- PMF or pmf: Probability Mass Function. If X is a discrete random variable, then

$$P(X = x) = f_X(x)$$

  is the pmf of X. The terms pmf and pf are interchangeable.
- RV or rv: Random Variable.
- UMP Test: Uniformly Most Powerful Test. A UMP test of H_0 against H_a is most powerful regardless of the value of the parameter under H_0 and H_a.

Appendix C

PRACTICE EXAMS

C.1 Equation Sheet

Series and Limits

$$\sum_{i=1}^{n} r^i = \begin{cases} \dfrac{1-r^{n+1}}{1-r} & \text{if } r \neq 1 \\ n+1 & \text{if } r = 1 \end{cases} \qquad \sum_{i=1}^{\infty} r^i = \begin{cases} \dfrac{1-r^{n+1}}{1-r} & \text{if } |r| < 1 \\ \infty & \text{if } r > 1 \\ \text{undefined} & \text{if } r < -1 \end{cases}$$

$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2} \qquad \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$

$$(a+b)^n = \sum_{i=0}^{n} \binom{n}{i} a^i b^{n-i} \qquad \ln(1+\varepsilon) = -\sum_{i=1}^{\infty} \frac{(-\varepsilon)^i}{i} \text{ if } |\varepsilon| < 1$$

$$\ln(1+\varepsilon) = \varepsilon + o(\varepsilon) \text{ if } |\varepsilon| < 1$$

$$\lim_{n\to\infty} \left(1 + \frac{a}{n} + o(n^{-1})\right)^n = e^a \qquad e^a = \sum_{i=0}^{\infty} \frac{a^i}{i!}$$

Distribution of Selected Sums & Expectations

$$X_i \sim \text{iid Bern}(\theta) \Longrightarrow \mathrm{E}(X_i) = \theta; \quad \mathrm{Var}(X_i) = \theta(1-\theta); \text{ and } \sum_{i=1}^{n} X_i \sim \mathrm{Bin}(n, \theta)$$

$$X_i \sim \text{iid Geom}(\theta) \Longrightarrow \mathrm{E}(X_i) = \frac{1}{\theta}; \quad \mathrm{Var}(X_i) = \frac{1-\theta}{\theta^2}; \text{ and } \sum_{i=1}^{n} X_i \sim \mathrm{NegBin}(n, \theta)$$

$$X_i \sim \text{iid Poi}(\lambda) \Longrightarrow \mathrm{E}(X_i) = \lambda; \quad \mathrm{Var}(X_i) = \lambda; \text{ and } \sum_{i=1}^{n} X_i \sim \mathrm{Poi}(n\lambda)$$

$$X_i \sim \text{iid Expon}(\lambda) \Longrightarrow \text{E}(X_i) = \frac{1}{\lambda}; \quad \text{Var}(X_i) = \frac{1}{\lambda^2}; \text{ and } \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, \lambda)$$

$$X_i \sim \text{iid NegBin}(k, \theta) \Longrightarrow \text{E}(X_i) = \frac{k}{\theta}; \quad \text{Var}(X_i) = \frac{k(1-\theta)}{\theta^2}; \text{ and}$$

$$\sum_{i=1}^{n} X_i \sim \text{NegBin}(nk, \theta)$$

C.2 Exam 1

1. Suppose $X \sim \text{Gam}(\alpha, \lambda)$;

$$f_X(x) = \frac{x^{\alpha-1}\lambda^{\alpha}e^{-\lambda x}}{\Gamma(\alpha)} I_{(0,\infty)}(x),$$

where $\alpha > 0$ and $\lambda > 0$.

   (a) Verify that the mgf of X is

$$\psi_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}.$$

   (b) For what values of t does the mgf exist?

2. Suppose that $W_1, \ldots, W_n$ is a random sample of size n from Expon(λ);

$$f_W(w) = \lambda e^{-\lambda w} I_{(0,\infty)}(w),$$

where $\lambda > 0$. Use mgfs to obtain the distribution of $Y = \sum_{i=1}^{n} W_i$. Hint: The mgf of W can be obtained from question #1 because the exponential distribution is a special case of the gamma distribution.

3. Suppose that X is a random variable with mgf

$$\psi_X(t) = \frac{1}{1-t}.$$

   (a) Give the pdf of X.

   (b) Derive an expression for $\text{E}(X^r)$; $r = 0, 1, 2, \ldots$.

4. Suppose that $X \sim \text{N}(\mu_X, \sigma_X^2)$; $Y \sim \text{N}(\mu_Y, \sigma_Y^2)$; and that $X \perp\!\!\!\perp Y$. The mgf of X is

$$\psi_X(t) = \exp\left\{t\mu_X + \frac{t^2\sigma_X^2}{2}\right\}.$$

(a) Verify that $\mathrm{E}(X) = \mu_X$ and that $\mathrm{Var}(X) = \sigma_X^2$.

(b) Prove that $X - Y \sim \mathrm{N}(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2)$.

5. Suppose that $\mathrm{X} \sim \mathrm{LogN}(\mu, \sigma^2)$. Compute

$$\Pr\left(e^{\mu} \leq X \leq e^{\mu+\sigma}\right).$$

6. Let W_i for $i = 1, \ldots, n$ and X_i for $i = 1, \ldots, m$ be iid random variables, each with distribution $\mathrm{N}(0, \sigma^2)$.

(a) Give the distribution of

$$U = \sum_{i=1}^{n} \left(\frac{W_i}{\sigma}\right)^2.$$

Justify your answer. Hint: First give the distribution of W_i/σ.

(b) Give the distribution of

$$V = \left(\frac{m}{n}\right) \left(\frac{\sum_{i=1}^{n} W_i^2}{\sum_{i=1}^{m} X_i^2}\right).$$

Justify your answer.

7. Suppose that X_i is a random sample of size n from an infinite sized population having mean μ and variance σ^2. Let $\overline{X}$ be the sample mean.

(a) Verify that $\mathrm{E}(\overline{X}) = \mu$

(b) Verify that $\mathrm{Var}(\overline{X}) = \sigma^2/n$.

(c) Let S^2 be the sample variance;

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

Verify that $\mathrm{E}(S^2) = \sigma^2$.

C.3 Exam 2

1. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from $f_X(x|\alpha, \beta)$, where

$$f_X(x|\alpha, \beta) = \frac{\alpha\beta^{\alpha}}{x^{\alpha+1}} I_{(\beta,\infty)}(x),$$

where $\alpha > 0$ and $\beta > 0$ are unknown parameters. This distribution is called the Pareto(α, β) distribution.

(a) Find a two dimensional sufficient statistic.

(b) Verify that the pdf of $X_{(1)}$ is Pareto$(n\alpha, \beta)$. That is,

$$f_{X_{(1)}}(x|\alpha,\beta) = \frac{n\alpha\beta^{n\alpha}}{x^{n\alpha+1}} I_{(\beta,\infty)}(x).$$

(c) The joint sampling distribution of the sufficient statistics can be studied using simulation. Let $U_1, U_2, \ldots, U_n$ be a random sample from Unif$(0,1)$. Show how U_i can be transformed into a random variable having a Pareto(α, β) distribution.

2. Suppose that $X \sim \text{Gamma}(\alpha, \lambda)$, where λ is known.

(a) Verify that the distribution of X belongs to the exponential family.

(b) Let $X_1, X_2, \ldots, X_n$ be a random sample from the Gamma(α, λ) distribution, where λ is known. Use the results from part (a) to find a sufficient statistic.

(c) Give the likelihood function that corresponds to part (b).

3. Consider the problem of making inferences about θ, the parameter of a geometric distribution. Let $X_1, X_2, \ldots, X_n$ be a random sample from $f_{X|\Theta}(x|\theta)$, where

$$f_{X|\Theta}(x|\theta) = \theta(1-\theta)^{x-1} I_{\{1,2,\ldots\}}(x).$$

(a) Verify that $T = \sum_{i=1}^n X_i$ is a sufficient statistic.

(b) Verify that the conditional distribution $P(\mathbf{X} = \mathbf{x}|T = t)$ does not depend on θ.

(c) Suppose that the investigator's prior beliefs about Θ can be summarized as $\Theta \sim \text{Beta}(\alpha, \beta)$. Find the posterior distribution of Θ and find the expectation of Θ conditional on $T = t$.

(d) Let $Z_1, Z_2, \ldots, Z_k$ be a sequence of future Geom(θ) random variables and let $Y = \sum_{i=1}^k Z_i$. Find the posterior predictive distribution of Y given T. That is, find $f_{Y|T}(y|t)$.

4. Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a distribution having mean μ, variance σ^2. Define Z_n as

$$Z_n = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

(a) State the central limit theorem.

(b) Verify that

$$Z_n = \sum_{i=1}^n U_i, \text{ where } U_i = \frac{Z_i^*}{\sqrt{n}} \text{ and } Z_i^* = \frac{X_i - \mu}{\sigma}.$$

(c) Assume that X has a moment generating function. Verify that

$$\psi_{Z_n}(t) = [\psi_{U_i}(t)]^n .$$

(d) Verify that the mean and variance of U_i are 0 and n^{-1}, respectively.

(e) Complete the proof of the central limit theorem.

C.4 Exam 3

1. Let X be a random variable; let $h(X)$ be a non-negative function whose expectation exists; and let k be any positive number. Chebyshev's inequality reveals that

$$P\left[h(X) \geq k\right] \leq \frac{\mathrm{E}\left[h(X)\right]}{k}$$

or, equivalently, that

$$P\left[h(X) < k\right] \geq 1 - \frac{\mathrm{E}\left[h(X)\right]}{k}.$$

(a) Define what it means for an estimator T_n to be consistent for a parameter θ.

(b) Use Chebyshev's inequality to verify that $\lim_{n\to\infty} MSE_{T_n}(\theta) = 0 \Longrightarrow T_n \xrightarrow{\text{prob}} \theta$.

2. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from Bern(θ).

(a) Give the likelihood function.

(b) Find a sufficient statistic.

(c) Verify that the score function is

$$S(\theta|\mathbf{X}) = \frac{\sum_{i=1}^n X_i - n\theta}{\theta(1-\theta)}.$$

(d) Derive the MLE of θ.

(e) Derive the MLE of $\dfrac{1}{\theta}$.

(f) Derive Fisher's information.

(g) Verify or refute the claim that the MLE of θ is the minimum variance unbiased estimator of θ.

3. Suppose that $X_i \sim$ iid Expon(λ) for $i = 1, \ldots, n$. It can be shown that $Y = \sum_{i=1}^n X_i$ is sufficient and that $Y \sim$ Gamma(n, λ).

(a) Derive the moment generating function of $Q = 2\lambda \sum_{i=1}^n X_i$ and verify that Q is a pivotal quantity. Use the moment generating function of Q to determine its distribution.

(b) Use Q to find a $100(1-\alpha)\%$ confidence interval for λ.

4. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from $f_X(x|\theta)$, where

$$f_X(x|\theta) = \theta x^{\theta-1} I_{(0,1)}(x) \text{ and } \theta > 0.$$

(a) Verify or refute the claim that the distribution of X belongs to the exponential class.

(b) Find the most powerful test of $\mathrm{H}_0\colon \theta = \theta_0$ versus $\mathrm{H}_a\colon \theta = \theta_a$, where $\theta_a > \theta_0$.

(c) Find the most uniformly powerful test of $\mathrm{H}_0\colon \theta = \theta_0$ versus $\mathrm{H}_a\colon \theta > \theta_0$.

(d) Suppose that the investigator's prior beliefs about θ can be summarized as $\Theta \sim \text{Gamma}(\alpha, \lambda)$. Find the posterior distribution of Θ. Hint: write x_i as $x_i = e^{\ln(x_i)}$.

(e) Find the Bayes estimator of θ based on a squared error loss function.