Studienzentrum Gerzensee Doctoral Program in Economics: Econometrics, 2022-23, Week 1

(This version: August 26, 2022)

Administrative Details

- Grades
- Exams: midterm and final are closed book, but we allow 1 page (double-sided) of notes.
- Graded take-home problem set after weeks 3 and 4.

Overview of Econometrics Sequence

- Week 1: Basic tools of probability, statistics, and econometrics
- Week 2: Linear model (including IV and linear GMM)
- Weeks 3 and 4: Mostly cross-section (Honore) and time-series (Watson) topics

Readings for Week 1

- Hogg, R.V, J.W. McKean, and A.T. Craig, *Introduction to Mathematical Statistics*, 6^{th} Edition, Prentice Hall, 2005. HMC (or earlier version: Hogg, R.V and A.T. Craig, *Introduction to Mathematical Statistics*, Fifth Edition, 1995, Macmillon Publishing.)
- Rao, C.R., Linear Statistical Inference and Its Applications, Second Edition, 1973, Wiley.
- Many other good books ... if you have a favorite, use it ... here's a nice one:
  - Casella, G. and R.L. Berger (2008), *Statistical Inference*, 2^{nd} Edition, Thompson Press.

1 Some Probability Concepts

1.1 Basic concepts

1. Experiments and Outcomes
   - Uncertain/Not Perfectly Predictable/Random/Stochastic (example: roll of die)
2. Sample Space (denoted by Ω)
   - Set of all possible outcomes (die example: $\{(1),(2),(3),(4),(5),(6)\}$)
   - Points in Ω are denoted by ω
   - More complicated examples ... may contain an infinite number of possible outcomes
     - Countable (Experiments like flipping a coin until a "Tails" appears)
     - Uncountable (growing a tomato and measuring its weight)
3. Events
   - An *event* is a subset of Ω, that is, it is collection of specific outcomes.
     - Die example: Rolling a '3 or 4', that is the subset $A = \{(3),(4)\} \in \Omega = \{(1),(2),(3),(4),(5),(6)\}$ is an event
4. Set Operations
   Let A and B denote two subsets of Ω and let a denote an element of Ω
   - $a \in A$ (a is contained in A)
   - $A \subset B$ (A is a subset of B)
   - $A \cup B$ (the union of A and B)
   - $A \cap B$ (the interection of A and B)
   - A^c (the compliment of A in Ω)
5. σ-Algebra (or σ-field)
   - Let $\mathcal{A}$ be a collection of subsets of Ω that satisfy
     (a) $\Omega \in \mathcal{A}$
     (b) If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$
     (c) If $A_1, A_2, ..., \in \mathcal{A}$ then $(\cup_{i=1}^{\infty} A_i) \in \mathcal{A}$
   - Sometimes (a) is replaced with $\emptyset \in \mathcal{A}$
   - You should show that, because $(\cup_{i=1}^{\infty} A_i^c)^c = \cap_{i=1}^{\infty} A_i$, then (b) and (c) imply that $\cap_{i=1}^{\infty} A_i \in \mathcal{A}$
6. Probability measure
   - A real-valued set function (maps sets into the real line) with the properties:
     (a) If $A \in \mathcal{A}$ then $\mathbb{P}(A) \geq 0$
     (b) $\mathbb{P}(\Omega) = 1$

(c) If $\{A_i\}_{i=1}^{\infty}$ is a countable collection of disjoint sets in $\mathcal{A}$, then $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

7. A Probability Space is the triple $(\Omega, \mathcal{A}, \mathbb{P})$.

8. Facts

- There are many useful facts (see Hogg and Craig, Section 1.3 theorems 1-5), including:
  (a) $0 \leq \mathbb{P}(A) \leq 1$
  (b) $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$
  (c) $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
  (d) If $\{A_i\}_{i=1}^{\infty}$ are a set of mutually exclusive and exhaustive subsets of Ω, then $\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \cap A_i)$.

1.1.1 An Exercise: Bonferroni inequality

Suppose A and B are two events. Let $C = A \cup B$ and $D = A \cap B$.

1. Show $\mathbb{P}(C) \leq \mathbb{P}(A) + \mathbb{P}(B)$
2. Show that $\mathbb{P}(D) \geq 1 - \mathbb{P}(A^c) - \mathbb{P}(B^c)$. (Hint: Show that $A \cap B = (A^c \cup B^c)^c$ and apply (1).
3. Let $A_1, A_2, ..., A_n$ denote n events. Show that

$$\mathbb{P}(A_1 \cup A_2 \cup ... \cup A_n) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2) + ... + \mathbb{P}(A_n).$$

(Hint: Use induction)

1.2 Conditional Probability

Let A and B denote two events in Ω with $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$. Then

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

is called the conditional probability of the event A given B.

- Note: $\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(A \cap B) = \mathbb{P}(B|A)\mathbb{P}(A)$.

Being a bit more careful: Formally, the idea is to construct a new probability space from $(\Omega, \mathcal{A}, \mathbb{P})$ by assigning zero probability to all outcomes that are not in B. The new probability measure, say $\mathbb{P}_o$ is constructed using the restriction that if $\omega_A \in (A \cap B)$ and $\omega_B \in B$ (so that ω_A and ω_B are in B), then

$$\frac{\mathbb{P}_o(\omega_A)}{\mathbb{P}_o(\omega_B)} = \frac{\mathbb{P}(\omega_A)}{\mathbb{P}(\omega_B)}$$

so that the relative odds of events in B remain the same under $\mathbb{P}$ and $\mathbb{P}_o$. If $\omega \notin B$ then $\mathbb{P}_o(\omega) = 0$. These restrictions determine P_o up to a scale factor, which is determined by the restriction that $\mathbb{P}_o(B) = 1$. The notation $\mathbb{P}(A|B)$ is short-hand for $\mathbb{P}_o(A)$, with $\mathbb{P}_o$ constructed in this way.

1.2.1 Independence

- Events A and B are independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$
- $\mathbb{P}(A|B) = \mathbb{P}(A)$ implies $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. This implies that $\mathbb{P}(B|A) = \mathbb{P}(B)$.

1.2.2 Bayes Rule

Suppose we know $\mathbb{P}(B|A)$ but we really want to know $\mathbb{P}(A|B)$. (Example, let B denote to the event that a medical test comes up "positive" and A be the event that a patient has a particular disease.) How can we compute $\mathbb{P}(A|B)$ from $\mathbb{P}(B|A)$ together with some additional information?

We know

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

and we also know that

$$B = (B \cap A) \cup (B \cap A^c)$$

where $(B \cap A)$ and $(B \cap A^c)$ are two disjoint sets. Thus,

$$\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$$

and

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

and

$$\mathbb{P}(B \cap A^c) = \mathbb{P}(B|A^c)\mathbb{P}(A^c)$$

so that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \mathbb{P}(B|A)\frac{\mathbb{P}(A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}$$

which is known as Bayes Rule.

2 Random Variables

Consider a probability space $(\Omega, \mathcal{A}, \mathbb{P})$. (Example: Ω denotes outcomes of 3 flips of a fair coin.) A random variable is a function that maps elements of Ω into the real line. (Examples: (i) $X(\omega)$ = number of heads; (ii) $X(\omega)$ = number of heads in first two flips; (iii) $X(\omega) = 1$ if heads appears on the 1^{st} and 3^{rd} flip and equals 0 otherwise.)

Being a bit more careful: As a technical matter, we must make sure that the sets of events that give rise to particular values of the function X are contained in $\mathcal{A}$. This makes the probability space for X, say $(\Omega_X, \mathcal{A}_X, \mathbb{P}_X)$, consistent with the original probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Such a restriction makes X *measurable* with respect to $\mathcal{A}$. Thus, a random variable $X(\omega)$ is a real valued function that maps $\omega \in \Omega$ into the real line, with the property that for any real x, $\{\omega | X(\omega) = x\} = A(x) \in \mathcal{A}$. Letting $A_X \subset \mathcal{A}_X$ the resulting probability function, $\mathbb{P}_X$, is given by $\mathbb{P}_X(A_X) = \mathbb{P}\{\omega | \omega \in \Omega, X(\omega) \in A_X\}$.

- Example: A fair coin is tossed 3 times.
  - $\Omega = \{(HHH), (THH), (HTH), (HHT), (HTT), (THT), (TTH), (TTT)\}$
  - $X(\omega)$ = number of heads
  - $\Omega_X = \{0, 1, 2, 3)$
  - $\mathcal{A}$ denotes all subsets of Ω, and $\mathcal{A}_X$ denotes all subsets of Ω_X
  - $\mathbb{P}_X(1) = \mathbb{P}[(HTT), (THT), (TTH)] = 3/8$, etc.

2.1 Distribution and Density Functions

Cumulative Distribution Function (CDF): The CDF of a random variable $X(\omega)$ is defined as

$$F_X(x) \stackrel{def}{=} \mathbb{P}(\omega | X(\omega) \leq x)$$

which is often denoted $\mathbb{P}(X \leq x)$ (with a slight abuse of notation). The notation $F_X(x)$ emphasizes that this function is for the random variable X and is evaluated at the point x.

- Some Properties of the CDF:
  1. For $x_2 \geq x_1$, $F_X(x_2) - F_X(x_1) = \mathbb{P}(x_1 < X \leq x_2)$
     (a) $F_X(-\infty) = 0$
     (b) $F_X(\infty) = 1$
     (c) $F_X(\cdot)$ is non-decreasing

Probability Density Function (pdf):

- Suppose X is a *discrete random variable* and can take on only a finite number of values $x_1, x_2, ...x_n$. We can then define
  $$\mathbb{P}(X = x_i) \stackrel{def}{=} p_i \stackrel{def}{=} f_X(x_i)$$
  as the *density function* for X and the resulting CDF is a step function. Again, the notation $f_X(x)$ emphasizes that function is for the random variable X and is evaluated at the point x.
  - For a discrete random variable $F_X(x) = \sum_{x_i \leq x} f_X(x_i)$, so the CDF is a step function.

- For a continuous random variable, define the pdf analogously: $f_X(\ .\ )$ satisfies $F_X(x) = \int_{-\infty}^{x} f_X(s)ds$ so that
$$f_X(x) \stackrel{def}{=} \frac{dF(x)}{dx}.$$
  - Note:
$$\mathbb{P}[x_1 \leq X \leq x_2] = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x)dx$$

- For mixtures of discrete and continuous random variables, $F_X(x)$ is defined by summing over the discrete and continuous components separately.

- Example: Suppose X has pdf
$$f_X(x) = \begin{cases} 1, \text{ for } 0 \leq x \leq 1 \\ 0 \text{ otherwise} \end{cases}$$
The associated CDF is
$$F_X(x) = \begin{cases} x, \text{ for } 0 \leq x \leq 1 \\ 0 \text{ for } x < 0 \\ 1 \text{ for } x > 1 \end{cases}$$
X is said to be *Uniformly Distributed* on $(0, 1)$, sometimes written as $X \sim \mathcal{U}(0, 1)$.

2.2 More than One Random Variable

2.2.1 Two Random Variables

Let X and Y denote two scalar random variables. The joint CDF is defined as

$$F_{X,Y}(x, y) \stackrel{def}{=} \mathbb{P}((X \leq x) \cap (Y \leq y))$$

which is often denoted $\mathbb{P}(X \leq x, Y \leq y)$.

- For a discrete random variable $F_{X,Y}(x, y) = \sum_{y_i \leq y} \sum_{x_i \leq x} f_{X,Y}(x_i, y_i)$, where $f_{X,Y}(x_i, y_j) = P((X = x_i) \text{ and } (Y = y_j))$.

- For a continuous random variable $F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(z_1, z_2)dz_1dz_2$.

- $F_X(x) = \mathbb{P}((X \leq x)) = \mathbb{P}((X \leq x) \cap (Y \leq \infty)) = F_{X,Y}(x, \infty) = \int_{-\infty}^{x} \left[\int_{-\infty}^{\infty} f(z, y)dy\right] dz$ is the CDF of X. Thus (for continuous RVs)
$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)dy$$
with obvious modifications for discrete RVs. In this context, $F_X(x)$ and $f_X(x)$ are sometimes called the *marginal* distribution and *marginal* density of X, respectively.

- The random variables X and Y are independent if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$ for all x and y. (Also, by implication, $f_{X,Y}(x, y) = f_X(x)f_Y(y)$.)

2.2.2 Conditional Distribution Functions

Let X and Y denote two discrete scalar random variables. The conditional pdf is defined as

$$f_{Y|X}(y|X=x) \stackrel{def}{=} \frac{\mathbb{P}[(Y=y) \text{ and } (X=x)]}{\mathbb{P}(X=x)}$$

for values of x with $\mathbb{P}(X=x)>0$. Equivalently

$$f_{Y|X}(y|X=x) = \frac{f_{X,Y}(x,y)}{f_X(x)}$$

for values of x with $f_X(x)>0$. We will use this as the definition of the conditional density in both the discrete and continuous cases. I will write this as $f_{Y|X}(y|x)$.

- The conditional CDF is $\mathbb{P}(Y \leq y|X=x) = \int_{-\infty}^{y} f_{Y|X}(s|x)ds$ (with obvious modification for discrete RVs).
- If X and Y are independent, then $F_{Y|X}(y|X=x) = F_Y(y)$ and $f_{Y|X}(y|x) = f_Y(y)$ for all x and y.

2.2.3 Multivariate Distribution Functions

CDFs and PDFs for the vector of random variables $X=(X_1, X_2, \ldots, X_n)$ are defined analogously to the bivariate case. Similarly for conditional distributions.

2.3 Expectations

- Let X denote a discrete random variable and let $g(X)$ denote a function of X, then

$$\mathbb{E}g(X) \stackrel{def}{=} \sum_i g(x_i)f_X(x_i)$$

- Let X denote a continuous random variable and let $g(X)$ denote a function of X, then

$$\mathbb{E}g(X) \stackrel{def}{=} \int_{-\infty}^{\infty} g(x)f_X(x)dx$$

- A few useful facts. Suppose a and b are constants, and $g_1(x)$ and $g_2(x)$ are two functions:
  - $\mathbb{E}a = \int_{-\infty}^{\infty} af_X(x)dx = a\int_{-\infty}^{\infty} f_X(x)dx = a$
  - $\mathbb{E}[ag(X)] = \int_{-\infty}^{\infty} ag(x)f_X(x)dx = a\int_{-\infty}^{\infty} g(x)f_X(x)dx = a\mathbb{E}g(X)$
  - $\mathbb{E}[g_1(X)+g_2(X)] = \mathbb{E}g_1(X) + \mathbb{E}g_2(X)$
  - Thus, $\mathbb{E}$ is a linear operator. (Note, sometimes I will use the notation $\mathbb{E}_X$ to make it clear that the expectation is taken with respect to f_X.)

- Example: Suppose X is uniformly distributed on $[0,1]$

$$\mathbb{E}(X) = \int_0^1 x dx = \frac{1}{2}x^2 \ \Big|_0^1 = \frac{1}{2}$$

and

$$\mathbb{E}(X^2) = \int_0^1 x dx = \frac{1}{3}x^3 \ \Big|_0^1 = \frac{1}{3}.$$

- Jargon: $\mathbb{E}(X)$ is called the *mean of X*. (More jargon involving expectations will be listed below.)

2.3.1 Functions of more than one random variable

- Suppose X and Y have joint density $f_{X,Y}(x,y)$ and let $g(X,Y)$ be a scalar function of X and Y, then $\mathbb{E}g(X,Y) = \int\int g(x,y)f_{X,Y}(x,y)dxdy$. (Note, following the notation I introduced above, I could have written this as $\mathbb{E}_{X,Y}g(X,Y)$ to make it clear that the expectation is taken with respect to $f_{X,Y}$.)
- Suppose G is a matrix of random variables, then $\mathbb{E}(G)$ is a matrix with $(ij)^{th}$ element equal to $\mathbb{E}(G_{ij})$.
- Exercise: Suppose $g(X,Y) = a(X)b(Y)$ and X and Y are independent. Show $\mathbb{E}g(X,Y) = [\mathbb{E}a(X)]\times[\mathbb{E}b(Y)]$. (Or, perhaps more clearly $\mathbb{E}_{X,Y}g(X,Y) = [\mathbb{E}_X a(X)]\times[\mathbb{E}_Y b(Y)]$.)

2.3.2 Conditional Expectations

The conditional expectation of Y given $X = x$ is the expectation of Y constructed using the probability density $f_{Y|X}(y|x)$. Thus, for a continuous random variable

$$\mathbb{E}(Y|X=x) = \int_{Y(\Omega_x)} y f_{Y|X}(y|x)dy$$

where Ω_x is the restricted sample space associated with the event $X = x$.

- Note that $\mathbb{E}(Y|X=x)$ depends on the particular value of x (obvious, but worth pointing out). The function $\mu_Y(x) = \mathbb{E}(Y|X=x)$ is called a *regression function*. It shows how the conditional mean of Y changes as the realization of X changes.

2.3.3 Law of Iterated expectations

The law of iterated expectations says that the expecation over Y can be computed in two steps: Specifically

$$\mathbb{E}_Y(Y) = \mathbb{E}_X\left[\mathbb{E}_{Y|X}(Y|X)\right]$$

This results follows from

$$\mathbb{E}_X\left[\mathbb{E}_{Y|X}(Y|X)\right] = \int\left[\int y f_{Y|X}(y|x)dy\right] f(x)dx$$

$$= \int \int y f_{Y|X}(y|x) f_X(x) dy dx = \int \int y f_{Y,X}(y,x) dy dx$$
$$= \int y \int f_{Y,X}(y,x) dx dy = \int y f_Y(y) dy = \mathbb{E}_Y(Y)$$

2.4 Application: Optimal Prediction

- **Problem 1:** Find the constant h that minimizes *Mean Squared Error* (mse), $\mathbb{E}\left[(Y-h)^2\right]$.
  - **Solution:** $\mathbb{E}[(Y-h)^2] = \int (y-h)^2 f_Y(y) dy$, which yields the first order conditions:

$$\int y f_Y(y) dy = h \int f_Y(y) dy = h$$

    so the optimal value of h is the mean of Y.

- **Problem 2:** Find the function $h(x)$ that minimizes $\mathbb{E}_{Y|X=x}\left[(Y-h(x))^2\right]$.
  - **Solution:** This is same problem as 1, but using the conditional distribution of $Y|X=x$. Thus, the optimal $h(.)$ is $h(x) = \mathbb{E}(Y|X=x)$.
    - Equivalently: The minimum mean square error predictor is given by regression function.
- **Problem 3:** Find the function $h(x)$ that minimizes $\mathbb{E}_{Y,X}\left[(Y-h(X))^2\right]$.
  - **Solution**: Write

$$\begin{aligned}\mathbb{E}_{Y,X}\left[(Y-h(X))^2\right] &= \int \int [(y-h(x)]^2 f_{X,Y}(x,y) dx dy \\ &= \int \int [(y-h(x)]^2 f_{Y|X}(y|x) f_X(x) dx dy \\ &= \int \left[\int [(y-h(x)]^2 f_{Y|X}(y|x) dy\right] f_X(x) dx\end{aligned}$$

    where the second equality factors the joint density into the conditional times the marginal and the final equality interchanges the order of integration. Note that the term in [] was minimized in Problem 2 with $h(x) = \mathbb{E}(Y|X=x)$. Evidently, this solves Problem 3.

2.5 Transformation of Variables

Let X be a random variable with CDF F_X. Let $Y = h(X)$ where $h(\cdot)$ is 1-to-1 with inverse h^{-1}. What is the distribution of Y?

- *Discrete case*: Suppose that X can take on values x_1, x_2, ..., x_n. Then Y can take on values y_1, y_2, ..., y_n with $y_i = h(x_i)$. Thus, $\mathbb{P}(Y=y_i) = \mathbb{P}(X = h^{-1}(y_i))$, so that

$$f_Y(y) = f_X(h^{-1}(y))$$

- *Continuous case*: We need to consider two cases:
  - Suppose that $h(.)$ is increasing. Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq h^{-1}(y)) = F_X(h^{-1}(y)).$$

    Thus,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(h^{-1}(y))}{dy} = f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}$$

  - Suppose that $h(.)$ is decreasing. Then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \geq h^{-1}(y)) = 1 - F_X(h^{-1}(y)).$$

    Thus,

$$f_Y(y) = \frac{dF_Y(y)}{dy} = -\frac{dF_X(h^{-1}(y))}{dy} = -f_X(h^{-1}(y))\frac{dh^{-1}(y)}{dy}.$$

  The two cases can be combined as:

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

- Example: Suppose X is uniformly distributed on $[0,1]$, and let $Y = X^2$. Then

$$f_Y(y) = f_X(y^{\frac{1}{2}})(\frac{1}{2}y^{-\frac{1}{2}}) = \begin{cases} \frac{1}{2}y^{-\frac{1}{2}} \text{ for } 0 \leq y \leq 1 \\ 0 \text{ otherwise} \end{cases}$$

- *Extension to the case where h is not one-to-one*: Let $U_1, ..., U_m$ be a partition of the real line and suppose that $y = h(x)$ is a one-to-one transformation on each U_i with range R_i and inverse $h_i^{-1}(y)$. The density of $Y = h(X)$ is

$$f_Y(y) = \sum_i f_X(h_i^{-1}(y))\left|\frac{dh_i^{-1}(y)}{dy}\right|$$

  where the summation is over those i for which $y \in R_i$.

- The extension to the multivariate discrete case is straightforward (just book-keeping). The extension to the continuous case requires somewhat more work (see Hogg and Craig Sections 4.3 and 4.5). The result is

$$f_Y(y) = f_X(h^{-1}(y))|J|$$

  where $|J|$ is absolute value of the Jacobian determinant in the inverse transformation – the absolute value of the determinant of the matrix $[\partial x_i/\partial y_j]$ where x_i is the i^{th} component of X and y_j is the j^{th} component of Y.
  - In particular, suppose $Y = HX$ where H is a non-singular matrix. Then $|J| = |H^{-1}| = |H|^{-1}$ and $f_Y(y) = f_X(H^{-1}y)|H|^{-1}$.

2.6 Moments

- The k^{th} moment of X is defined as $\mathbb{E}(X^k)$
  - The *mean* of X is the first moment, $\mathbb{E}(X^1)$. It is denoted as $\mu = \mathbb{E}(X)$
- The k^{th} centered moment of X is defined as $\mathbb{E}((X-\mu)^k)$
  - The second centered moment is called the variance and is denoted σ^2.
  - A straightforward calculation shows

$$\sigma^2 = \mathbb{E}\left((X-\mu)^2\right) = \mathbb{E}(X^2) - \mu^2$$

  - $\sigma \overset{def}{=} \sqrt{\sigma^2}$ is called the *standard deviation of X.*
  - All odd centered moments are equal to zero for symmetric pdfs. (A pdf is symmetric if $f(-x) = f(x)$ for all $x \geq 0$. A pdf is symmetric around a point a if $f(a-x) = f(a+x)$.)
  - The first moment and the second, third and fourth centered moments are used to measure the center (location), spread, skewness and kurtosis of the distribution.
  - Example: Suppose $X \sim \mathcal{U}(0,1)$ (i.e., X is uniformly distributed on $(0,1)$). Then

$$\mu = \mathbb{E}(X) = \frac{1}{2}$$

$$\sigma^2 = \mathbb{E}(X^2) - \mu^2 = \frac{1}{3} - (\frac{1}{2})^2 = \frac{1}{12}$$

  - Example: Suppose X has mean μ_X and variance σ_X^2. Let a and b be constants and $Y = a + bX$. Then Y has mean and variance given by $\mu_Y = a + b\mu_X$ and $\sigma_Y^2 = b^2\sigma_X^2$ (so that $\sigma_Y = |b|\sigma_X$).
  - Example: Suppose X has mean μ_X and variance σ_X^2. Let

$$Z = \frac{X - \mu_X}{\sigma_X}.$$

    Then $\mu_Z = 0$ and $\sigma_Z = 1$. Z is called the *standardized* version of X.

2.6.1 Moment generating function

The Moment Generating Function (MGF) of X is defined as

$$M(t) = \mathbb{E}(e^{tX})$$

Since

$$\mathbb{E}(e^{tX}) = \int e^{tx} f_X(x) dx$$

$$M'(t) = \int x e^{tx} f_X(x) dx \Rightarrow M'(0) = \mathbb{E}(X)$$

$$M''(t) = \int x^2 e^{tx} f_X(x) dx \Rightarrow M''(0) = \mathbb{E}(X^2)$$

and in general

$$M^{(j)}(t) = \int x^j e^{tx} f_X(x)dx \Rightarrow M^{(j)}(0) = \mathbb{E}(X^j)$$

Thus, if you know the MGF of random variable, it is often a straightforward calculation to find its moments.

- The MGF does not exist for all random variables – the relevant integrals may not converge. (e^{tx} can get very large for extreme values of X). The MGF can be modifed to produce a *Characteristic Function* which always exists. (This uses it in place of t with $i = \sqrt{-1}$.)
- The MGF can be inverted to find the density. A calculation shows that $f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} M(it)e^{-itx}dt$ where $i = \sqrt{-1}$.
  - Thus, a moment generating function uniquely characterizes a distribution. Thus, if X and Y have the same MGF, then they have the same CDF.
- Example: Suppose X is uniformly distributed on $[0, 1]$, then

$$M(t) = \int_0^1 e^{tx}dx = \frac{1}{t}[e^t - 1]$$

- Example: Suppose X has MGF $M_X(t)$ and $Y = a + bX$, where a and b are constants. Then $M_Y(t) = \mathbb{E}(e^{t(a+bX)}) = e^{at}\mathbb{E}(e^{tbX}) = e^{at}M_X(bt)$.

2.6.2 Moments for vector-valued random variables

- Suppose that X and Y are two scalar random variables with joint cdf $F_{X,Y}(x, y)$. The covariance between X and Y is defined as

$$\sigma_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)]$$

  - You should show that $\sigma_{XY} = \mathbb{E}(XY) - \mu_X\mu_Y$.
  - You should that $\sigma_{XY} = 0$ when X and Y are independent.
- Let a and b denote two constants, and let $W = aX + bY$. Then
  - $\mu_W = a\mu_X + b\mu_Y$
  - $\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 + 2ab\sigma_{XY}$
- These results can be generalized. Suppose $X = (X_1\ X_2\ \cdots X_n)'$:

$$\mathbb{E}(X) = \mu_X = \begin{bmatrix} \mathbb{E}(X_1) \\ \mathbb{E}(X_2) \\ \vdots \\ \mathbb{E}(X_n) \end{bmatrix}$$

  is the mean vector.
- $\mathbb{E}(XX') = [\mathbb{E}(X_iX_j)]$ is the $n \times n$ second moment matrix.

- $\mathbb{E}[(X-\mu_X)(X-\mu_X)'] = [\mathbb{E}\left((X_i-\mu_i)(X_j-\mu_j)\right)]$ is the $n \times n$ covariance matrix.
  - $\sigma_{ij} = \mathbb{E}\left[(X_i-\mu_i)(X_j-\mu_j)\right]$ is called the covariance between X_i and X_j.
  - The matrix
$$\mathbb{E}[(X-\mu_X)(X-\mu_X)'] = \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$
    is called the covariance matrix.
  - Exercises ... Show:
    - Σ is a symmetric $n \times n$ matrix since $\sigma_{ij} = \sigma_{ji}$.
    - If X_i and X_j are independent, then $\sigma_{ij} = 0$.
    - $\Sigma = \mathbb{E}(XX') - \mu_X\mu_X'$.
- Let α denote a $n \times 1$ non-stochastic vector and let $Y = \alpha'X$. Denote the mean vector of X as μ_X and its covariance matrix as Σ_X. Then
  - $\mu_Y = \alpha'\mu_X$. (Exercise: Show this)
  - The variance of $Y$is $\sigma_Y^2 = \alpha'\Sigma_X\alpha$. (Because σ_Y^2 must be non-negative, this implies Σ is positive semi-definite.) (Exercise: Show this)
- $\rho_{ij} = \sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}$ is the correlation between X_i and X_j.
  - $[\rho_{ij}]$ is called the correlation matrix
  - Since
$$V = \begin{bmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ji} & \sigma_{jj} \end{bmatrix}$$
    is positive semi-definite, then $|V| \geq 0$, which implies $\sigma_{ii}\sigma_{jj} \geq \sigma_{ij}^2$, so that $-1 \leq \rho_{ij} \leq 1$.
  - If X_i and X_j are independent, then $\rho_{ij} = 0$.
- The moment generating function for X is
$$M_X(t) = \mathbb{E}(e^{t'X})$$
  where t is a $n \times 1$ vector.

3 Selected Probability Distributions

- **Bernoulli**: X can take on two values, 0 and 1, $f(1) = p$ and $f(0) = 1-p$. Thus, $f(x) = p^x(1-p)^{1-x}$ for $x = \{0, 1\}$ and $f(x) = 0$ for all other values of x.
  - The parameter p indexes the distribution
  - Exercise: Work out MGF and all moments.

- **Binomial**: Suppose X_i , $i = 1, ..., n$ are Independent and Identically Distributed (written as *i.i.d.* or *iid)* Bernoulli random variables with parameter p. Let $Y = \sum_{i=1}^{n} X_i$. Then Y has a Binomial distribution with parameters n and p. Y can take on values $0, 1, ..., n$. The PDF is

$$f(y) = \left(\begin{array}{c} n \\ y \end{array} \right) p^y (1-p)^{n-y}$$

where

$$\left(\begin{array}{c} n \\ y \end{array} \right) = \frac{n!}{y!(n-y)!}$$

is the number of ways that y successes can occur in n outcomes.

  - Exercise: work out MGF of Y. Use: $M_Y(t) = \Pi_{i=1}^{n} M_{X_i}(t)$ which follows from (i) ($e^{t\sum X_i} = \Pi e^{tX_i}$) and (ii) independence.

- Poisson: X takes on the values $0, 1, 2, ...$ with

$$f_X(x) = \frac{m^x e^{-m}}{x!}$$

This distribution is useful for modeling "successes" that occur over intervals of time. (Customers walking into a store, changes in Fed Funds Rate, etc.). Let $g(x, w)$ denote the probability that x successes occur in a period of length w. Suppose

1. $g(1, h) = \lambda h + o(h)$, where λ is a positive constant, $h > 0$, and $o(h)$ means a term that satisfies $\lim_{h\to 0}[o(h)/h)] = 0$
   (a) $\sum_{x=2}^{\infty} g(x, h) = o(h)$
   (b) The number of successes in non-overlapping periods are independent.

   When these postulates describe an experiment, then you can show (See Hogg and Craig Section 3.2) that the number of successes over a period of time with length w follows a Poisson distribution with parameter $m = \lambda w$.
   - Exercise: You should be able to show the MGF is $e^{m(e^t-1)}$, and that both the mean and variance are equal to m. (Hint: remember that $e^z = \sum_{k=0}^{\infty} \frac{1}{k!} z^k$.)

3.1 Some Continuous Distributions

- **Uniform**: $f(x) = (b-a)^{-1}$ for $a \leq x \leq b$ and 0 elsewhere. This is written as $X \sim \mathcal{U}[a, b]$.
  - MGF is

$$M_X(t) = \frac{e^{bt} - e^{at}}{(b-a)t}$$

- **Univariate Normal**
  - Standard Normal (denoted $\mathcal{N}(0, 1)$): Standard normal variables play an important role in statistics and are often denoted by Z. Thus $Z \sim \mathcal{N}(0, 1)$ means that Z has pdf:

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

- General Normal (denoted $\mathcal{N}(\mu, \sigma^2)$): Let $Y = \mu + \sigma Z$ where Z is standard normal and $\sigma > 0$. Then from the change-of-variables formula

$$f_Y(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}.$$

  We write $Y \sim \mathcal{N}(\mu, \sigma^2)$.

- MGF for standard normal:

$$M_Z(t) = \int_{-\infty}^{\infty} e^{tz} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}z^2 + tz] dz$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}\{z^2 - 2tz\}] dz$$

$$= \exp[\frac{t^2}{2}] \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}(z-t)^2] dz$$

$$= e^{\frac{t^2}{2}}$$

- since the integral term =1 (it is the integral of the density of a random variable distributed $\mathcal{N}(t, 1)$).

- Thus, from the MGF
  - $\mathbb{E}(Z) = 0$
  - $\mathbb{E}(Z^2) = 1$
    - $\sigma_Z^2 = 1$
  - $\mathbb{E}(Z^k) = 0$ for $k = 1, 3, 5, ...$
  - $\mathbb{E}(Z^4) = 3$.

- MGF for General Normal:

$$Y = \mu + \sigma Z \text{ so that } M_Y(t) = e^{\mu t} M_Z(\sigma t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

  - A direct calculations yields
    - $\mathbb{E}(Y) = \mu$
    - $\mathbb{E}(Y^2) = \sigma^2 + \mu^2$, so that $\text{Var}(Y) = \sigma^2$
    - $\mathbb{E}[(Y-\mu)^k] = 0$, for $k = 1, 3, 5, ...$
    - $\mathbb{E}[(Y-\mu)^4] = 3\sigma^4$

- **Chi-Squared Distribution**: Let Z_i, $i = 1, ..., n$ be distributed *iid* $\mathcal{N}(0, 1)$ (where *iid* $\mathcal{N}$ denotes *independent and identically distributed normal*) and let $Y = \sum_{i=1}^n Z_i^2$. Then Y_i is distributed as a χ_n^2 random variable. The parameter n is called the *degrees of freedom* of the distribution.

- **F Distribution**: Let $Y \sim \chi_n^2$, $X \sim \chi_m^2$ and suppose that Y and X are independent. Then

$$Q = \frac{Y/n}{X/m}$$

is distributed $F_{n,m}$. The parameters n and m are called the numerator and denominator degrees of freedom.

- **Students t distribution**: Let $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_n$ and suppose Z and Y are independent. Then,

$$X = \frac{Z}{(Y/n)^{\frac{1}{2}}}$$

is distributed t_n. The parameter n is called the degrees of freedom of the distribution.

3.2 Multivariate Normal Distribution

A short digression: some algebra and notation

- Matrix square roots: Suppose Σ is a positive definite matrix. Then Σ can be factored as

$$\Sigma = AA'$$

where A is a matrix *square root* of Σ. Often the notation $A = \Sigma^{1/2}$ is used to stress this interpretation, so that $\Sigma = \Sigma^{1/2}\Sigma^{1/2\prime}$. Note that $\Sigma^{-1} = (A')^{-1}A^{-1}$ which is sometimes written as $\Sigma^{-1} = \Sigma^{-1/2\prime}\Sigma^{-1/2}$ where $\Sigma^{-1/2} = (\Sigma^{1/2})^{-1}$. (Note: A is not unique. This non-uniqueness matters in some applications, but it will not matter here.)

- Inverse determinants and determinants of inverses, etc.: Recall that $|A|^{-1} = |A^{-1}|$ and that $|A| = |A'|$,
- With $\Sigma = \Sigma^{1/2}\Sigma^{1/2\prime}$ then $|\Sigma| = |\Sigma^{1/2}| \times |\Sigma^{1/2\prime}| = |\Sigma^{1/2}|^2$. Similarly $|\Sigma^{-1/2}|^2 = |\Sigma^{-1}| = |\Sigma|^{-1}$. Thus $|\Sigma^{-1/2}| = |\Sigma|^{-1/2}$.

Definition: A p–dimensional random vector X is p–dimensionally normally distributed if the one–dimensional random variables $a'X$ are normally distributed for all $a \in \mathbb{R}^p$, where a is non-stochastic and non-zero (See Rao page 518). It follows from this definition that if $X = (X_1, X_2, ..., X_p)'$ is normally distributed, then X_i is normally distributed for $i = 1, .., p$. (To see this, choose a so it extracts X_i from the vector X.)

- Let μ and Σ denote the mean vector and covariance matrix of X. The multivariate normal distribution is characterized by μ and Σ. To see this, note that for any $a \in \mathbb{R}^p$:

$$\mathbb{E}[a'X] = a'\mu \qquad \text{and} \qquad \text{Var}[a'X] = a'\Sigma a.$$

- The moment generating function for X evaluated at a is the moment generating function for $a'X$ evaluated at $t = 1$. That is:

$$M_X(a) = \mathbb{E}e^{a'X} = M_{a'X}(1) = e^{a'\mu + \frac{1}{2}a'\Sigma a},$$

and the final equality follows from $a'X \sim N(a'\mu, a'\Sigma a)$. Because the MGF uniquely defines the probability distribution, again we see that the pdf for X must only depend on the parameters μ and Σ.

- This multivariate normal distribution is typically denoted

$$X \sim \mathcal{N}(\mu, \Sigma)$$

or sometimes $X \sim \mathcal{N}_p(\mu, \Sigma)$ to emphasize that X has p elements.

The multivariate normal has a special role in statistics because of the central limit theorem, a result discussed below. In many applications subsets of elements of X and/or a function of X appears, and it is useful to characterize the relevant pdf for this function of X. Here I discuss a few results that will prove useful in our later work. (For a more detailed discussion see C. R. Rao: *Linear Statistical Inference and Its Applications* pp. 185–189 and pp. 519–527, and Sections 3.5, 9.1, 9.8 and 9.9 of HMC.)

We will now state a number of results about multivariate normal distributions. Proofs are sketched. (Complete proofs can be found in C. R. Rao: *Linear Statistical Inference and Its Applications* pp. 185–189 and pp. 519–527 and Section 3.5, 9.1, 9.8 and 9.9 of HMC.)

Theorem A. (Linear functions of X are normally distributed) Let $X \sim \mathcal{N}_p(\mu, \Sigma)$, let B be a $k \times p$ matrix, and let η denote a $k \times 1$ vector, then

$$Y = \eta + BX \sim \mathcal{N}_k(\eta + B\mu, B\Sigma B').$$

Notes: The proof is straightforward.

Theorem B. (The multivariate normal density) Suppose $X \sim \mathcal{N}_p(\mu, \Sigma)$ and Σ has rank p. Then X has density given by

$$f_X(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\}, \quad x \in R^p.$$

Notes: Perhaps the easiest way to see this is as follows. Let Z denote a p-vector of iid $\mathcal{N}(0,1)$ random variables, then you can verify that $M_Z(a) = M_{a'Z}(1) = e^{\frac{1}{2}a'a}$, which implies that Z is multivariate normal. (Remember we worked out the MGF for a multivariate normal above.)

The pdf of Z is $f_Z(z) = \prod_{i=1}^p \left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{1}{2}z_i^2} = \left(\frac{1}{\sqrt{2\pi}}\right)^p e^{-\frac{1}{2}z'z}$, where the first equality follows from the PDF for a standard normal and the independence of the elements of Z. Now let $X = \mu + \Sigma^{1/2}Z$, and from Theorem A, $X \sim \mathcal{N}(\mu, \Sigma)$. From the change-of-variables formula, the density of X is then $f_X(x) = \left|\Sigma^{1/2}\right|^{-1} f_Z\left(\Sigma^{-1/2}(x-\mu)\right)$, and rearranging yields the formula given above.

Theorem C. (Independent normally distributed random variables have a joint normal distribution.) If $X_1 \sim \mathcal{N}_p(\mu_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}_q(\mu_2, \Sigma_2)$, and X_1 and X_2 are independent, then $X = (X_1', X_2')' \sim \mathcal{N}_{p+q}(\mu, \Sigma)$, where

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \qquad \text{and} \qquad \Sigma = \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix}.$$

Proof: Write out the joint density of $X = (X_1' \ X_2')'$ as the product of the densities of X_1 and X_2 and rearrange.

Theorem D. (Conditional normal distribution). Let $X \sim \mathcal{N}_p(\mu, \Sigma)$. Also let $X = (X_1' \ X_2')'$, $\mu = (\mu_1' \ \mu_2')'$, and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, be the partitions of X, μ and Σ.

The conditional distribution of X_1 given $X_2 = x_2$ is given by

$$X_1|X_2 = x_2 \sim \mathcal{N}\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right).$$

Proof: This is a (tedious) calculation applied to the definition of a conditional distribution. Write:

$$f_{X_1|X_2}(x_1) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{(2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\}}{(2\pi)^{-p_2/2}|\Sigma_{22}|^{-1/2}\exp\left\{-\frac{1}{2}(x_2-\mu_2)'\Sigma_{22}^{-1}(x_2-\mu_2)\right\}}$$

where X_2 is $p_2 \times 1$. Using the partitioned inverse formula:

$$\Sigma^{-1} = \begin{bmatrix} V^{-1} & -V^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{21}V^{-1} & \Sigma_{22}^{-1}(I + \Sigma_{21}V^{-1}\Sigma_{12}\Sigma_{22}^{-1}) \end{bmatrix}$$

where $V = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ and $|\Sigma| = |\Sigma_{22}| \times |V|$.

Substituting these expressions into $f_{X_1|X_2}(x_1)$ and rearranging yields:

$$f_{X_1|X_2}(x_1) = |V|^{-1/2}(2\pi)^{-p_1/2} exp\left(-\frac{1}{2}(x_1 - \mu_{1|2})'V^{-1}(x_1 - \mu_{1|2})\right)$$

with $\mu_{1|2} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)$. The result then follows immediately.

Theorem E. Suppose $X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$ and $X_1|X_2 = x_2 \sim \mathcal{N}(A + Bx_2, \Omega)$ for all values of x_2 and where A, B, and Ω are constants. Then $X = (X_1' \ X_2')'$ has a multivariate normal distribution.

Proof: The joint distribution is

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1|X_2=x_2}(x_1|x_2)f_{X_2}(x_2).$$

Write these out, carry out the multiplication and you will discover that

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} A + B\mu_2 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} B\Sigma_{22}B' + \Omega & B\Sigma_{22} \\ \Sigma_{22}B' & \Sigma_{22} \end{pmatrix}\right)$$

Theorem F. (Sums of independent normals) If $X_1 \sim \mathcal{N}_p(\mu_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}_p(\mu_2, \Sigma_2)$, and X_1 and X_2 are independent, then

$$X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2).$$

Proof: Immediate.

Let $X \sim \mathcal{N}_p(\mu, \Sigma)$. Also let $X = (X_1', X_2')'$, $\mu = (\mu_1', \mu_2')'$, and

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

be the partitions of X, μ and Σ such that X_1 and μ_1 are k–dimensional and Σ_{11} is a $k \times k$ matrix, then Theorems G and H hold:

Theorem G. The marginal distribution of X_1 is $\mathcal{N}_k(\mu_1, \Sigma_{11})$.

Theorem H. (For a normal, a zero correlation implies independence) If $\Sigma_{12} = 0$ then X_1 and X_2 are independent.

Proof: Write out the joint distribution and you can see that it factors appropriately.

Theorem I. (Characterizing independence of linear combinations of normal variables) If $X \sim \mathcal{N}_p(\mu, \Sigma)$, B is a $p \times k$ matrix, and C is a $p \times m$ matrix, then $B'X$ and $C'X$ are independent if and only if $B'\Sigma C = 0$.

Proof: Note the $B'X$ and $C'X$ are jointly normally distributed with covariance $B'\Sigma C = 0$.

Quadratic forms of normal random vectors:

The quantity $Y'AY$ is called a *quadratic form*. Without loss of generality, assume that A is symmetric. (This follows since $Y'AY = Y'A'Y$ so that $Y'AY = Y'BY$ with $B = \frac{1}{2}(A + A')$.) In all of the quadratic forms discussed below, assume that the matrix in the middle is symmetric.

There are many useful theorems on the distribution of quadratic forms. We will discuss a few. They rely on the following sets of results: Suppose $Z_i \sim iid\ \mathcal{N}(0,1)$ random variables for $i = 1, ..., n$. Then we know a few things:

- **First:** $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$. This can be written in another way: let Z denote the $n \times 1$ vector $(Z_1, Z_2, ..Z_n)'$, then $\sum_{i=1}^n Z_i^2 = Z'Z \sim \chi_n^2$.
- **Second:** Suppose we partition Z into its first n_1 elements and last n_2 elements: $Z = \begin{bmatrix} Z_{1:n_1} \\ Z_{n_1+1:n} \end{bmatrix}$. Then we know

  - $Z_{1:n_1}$ is independent of $Z_{n_1+1:n}$
  - $AZ_{1:n_1}$ is normally distributed
  - $Z'_{n_1+1:n}Z_{n_1+1:n} \sim \chi^2_{n_2}$
  - $AZ_{1:n_1}$ and $Z'_{n_1+1:n}Z_{n_1+1:n}$ are independent.

- **Third:** Suppose P is a $n \times m$ matrix with orthonormal columns, that is $P'P = I_m$. Then
  - $P'Z$ is normally distributed
  - $P'Z \sim \mathcal{N}(0, I_m)$ so that the elements of $P'Z$ are m *iid* $\mathcal{N}(0,1)$ random variables
  - Letting $Y = P'Z$, then $Y'Y \sim \chi^2_m$ and rewriting this $Y'Y = Z'PP'Z \sim \chi^2_m$

Now, a few results:

Theorem J. If $X \sim \mathcal{N}_p(\mu, \Sigma)$ where Σ has rank p, then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_p$.

Notes: Write $Z = \Sigma^{-1/2}(X - \mu)$. Note that $Z \sim \mathcal{N}(0, I_p)$. Then $(X-\mu)'\Sigma^{-1}(X-\mu) = Z'Z \sim \chi^2_p$ follows directly.

Theorem K. (Quadratic forms around idempotent matrices) Let M denote an idempotent $p \times p$ matrix with rank k, then $Z\,'MZ \sim \chi^2_k$.

Notes: Write $M = P\Lambda P'$, where Λ contains the eigenvalues of M on the diagonal and the rows of P are the orthonormal eigenvectors. Because M is idempotent we can write

$$M = \begin{bmatrix} P_1 & P_2 \end{bmatrix} \begin{bmatrix} I_k & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} P_1' \\ P_2' \end{bmatrix} = P_1 P_1'$$

Thus $Z'MZ = Y'Y$ where $Y = P_1'Z$, and the result follows from $Y \sim N(0, P_1'P_1)$, where $P_1'P_1 = I_k$.

Theorem L: Let $X = PZ$ and $Q = Z'AZ$, where $PA = 0$, then X and Q are independent.

Notes: Suppose Z is $p \times 1$, A is $p \times p$ with rank m, and P is $q \times p$. Because A is symmetric, it can be decomposed as $A = G\Lambda G'$, where G is a $p \times m$ matrix with full column rank and Λ is a diagonal matrix with the non-zero eigenvalues of A on the diagonal. Let $Y = G'Z$. Then $Z'AZ = Y'\Lambda Y$. Note that $(Y'\ X')'$ are multivariate normal, with $\text{Cov}(X, Y) = PG$. But $PA = PG\Lambda G' = 0$. This implies $PG\Lambda G'G = 0$, so $PG = 0$ (because $\Lambda G'G$ is non-singular). Because X and Y have covariance zero, they are independent.

Theorem M: Let $Q_1 = Z\,'A_1Z$ and $Q_2 = Z\,'A_2Z$, where $A_1A_2 = 0$. Then Q_1 and Q_2 are independent.

Notes: Same idea as Theorem K.

Exercise: Let Y_i, for $i = 1, ..., n$ be distributed *iid* $\mathcal{N}(\mu, \sigma^2)$. Let $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^n (Y_i - \overline{Y})^2$. Show

1. (i) $\frac{(\overline{Y}-\mu)}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(0,1)$
2. (ii) $(n\text{–}1)s^2/\sigma^2 \sim \chi^2_{n-1}$
3. (iii) $\overline{Y}$ and s^2 are independent.
4. (iv) $\left[\frac{(\overline{Y}-\mu)}{\sqrt{\sigma^2/n}}\right] / \left[\left(\frac{(n-1)s^2}{\sigma^2}\right)/(n-1)\right]^{1/2} = \frac{(\overline{Y}-\mu)}{\sqrt{s^2/n}} \sim t_{n-1}$

where t_{n-1} is the Student's t distribution with $n-1$ degrees of freedom.

Solution: Let

- $Y_{1:n}$ denote the $n\times1$ vector $(Y_1, Y_2, ..., Y_n)'$
- l denote an $n\times1$ vector of 1s.

Note:

- From Theorem C: $Y_{1:n} \sim \mathcal{N}_n(l\mu, \sigma^2 I)$.
- Let $Z_{1:n} = (Y_{1:n} - \mu l)/\sigma$. Then $Z_{1:n} \sim \mathcal{N}_n(0, I)$ follows from Theorem A. Rearranging yields $Y_{1:n} = l\mu + \sigma Z_{1:n}$.

To show (1):

- Write $\overline{Y} = AY_{1:n}$ with $A = (l'l)^{-1}l'$.
- From Theorem A: $\overline{Y} \sim \mathcal{N}(Al\mu, \sigma^2 AIA')$. But $Al = 1$ and $AA' = n^{-1}$, so $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$

To show (2):

- $\sum_{i=1}^n (Y_i - \overline{Y})^2 = (Y_{1:n} - \overline{Y}l)'(Y_{1:n} - \overline{Y}l)$.
- $Y_{1:n} - l\overline{Y} = MY_{1:n}$ with $M = I\text{–}l(l'l)^{\ 1}l'$.
- $MY_{1:n} = M(\mu l + \sigma Z_{1:n}) = \sigma M Z_{1:n}$ because $Ml = 0$.
- $\sigma^{-2}(Y_{1:n} - l\overline{Y})'(Y_{1:n} - l\overline{Y}) = Z_{1:n}MZ_{1:n} \sim \chi^2_{rank(M)}$ where the results follows from Theorem K.
- The result (2) follows by noting that rank(M) = trace(M) = trace$(I_n\text{–}l(l'l)^{-1}l) = n -$ trace$[l(l'l)^{\ 1}l] = n -$ trace$[(l'l)^{-1}l'l] = n - 1$.

To show (3);

- $s^2 = \sigma^2(n-1)^{-1}Z_{1:n}MZ_{1:n}$
- $\overline{Y} = AY_{1:n} = \mu l + \sigma AZ_{1:n}$
- The results follows from Theorem L after noting the $MA = 0$.

Result (4) follows directly from (1)-(3).

4 Some Useful Inequalities

4.1 Jensen's inequality

Jensen's inequality: Let $h(\bullet)$ be a convex function and X a random variable. Then $\mathbb{E}[h(X)] \geq h(\mathbb{E}(X))$.

Proof: Recall that if the function h is convex, then for any value x_0, there is a line through $(h(x_0), x_0)$ such that $h(x)$ is never below the line. Equivalently, for any x_0, there is a constant a, such that $h(x) \geq h(x_0) + a(x - x_0)$ for all x.

- Set $x_0 = \mathbb{E}(X)$, then
  - $\mathbb{E}[h(X)] \geq \mathbb{E}\left(h(x_0) + a(X - x_0)\right) = h(\mathbb{E}(X)) + a\mathbb{E}(X - \mathbb{E}(X)) = h(\mathbb{E}(X))$.

If h is concave, $\mathbb{E}[h(X)] \leq h(\mathbb{E}(X))$, by an analogous argument.

Example: $\mathbb{E}(Y^4) \geq [\mathbb{E}(Y^2)]^2$, so that $\mathbb{E}(Y^4) < \infty$ implies that $\mathbb{E}(Y^2) < \infty$. (This follows from Jensen's inequality with $X = Y^2$ and $h(X) = X^2$).

4.2 Chebyshev's inequality

Chebyshev's inequality: Let $\epsilon > 0$, then $\mathbb{P}(|X| \geq \epsilon) \leq \mathbb{E}(X^2)/\epsilon^2$.

Proof:

$$\begin{aligned}
\mathbb{E}(X^2) &= \int_{-\infty}^{\infty} x^2 f(x)dx \\
&= \int_{-\infty}^{-\epsilon} x^2 f(x)dx + \int_{-\epsilon}^{\epsilon} x^2 f(x)dx + \int_{\epsilon}^{\infty} x^2 f(x)dx \\
&\geq \int_{-\infty}^{-\epsilon} x^2 f(x)dx + \int_{\epsilon}^{\infty} x^2 f(x)dx \\
&\geq \int_{-\infty}^{-\epsilon} \epsilon^2 f(x)dx + \int_{\epsilon}^{\infty} \epsilon^2 f(x)dx \\
&= \epsilon^2 \mathbb{P}(|X| \geq \epsilon)
\end{aligned}$$

and rearranging yields the result.

- The following result, called *Markov's inequality*, follows from an analogous argument. Let $\epsilon > 0$, then

$$\mathbb{P}(|X| \geq \epsilon) \leq \mathbb{E}(|X|^p)/\epsilon^p.$$

5 Large Sample Theory

5.1 Convergent sequences of non-stochastic variables

- Let $\{X_n\}$ denote a sequence of non-stochastic variables. Then we say X_n has a limit X if, for any $\epsilon > 0$ there is a number N (which may depend on ϵ, so write it as $N(\epsilon)$), such that $|X_n - X| < \epsilon$ for all $n > N(\epsilon)$. This is written as

$$\lim_{n\to\infty} X_n = X$$

or $X_n \to X$.

5.2 Convergent sequences of random variables

We need to discuss convergence of random sequences, $\{X_n\}$ to the random limit X. With random sequences, one can't be sure that $|X_n - X| < \epsilon$ because of randomness in X_n and X. There are a variety of ways to handle this randomness, leading to different notions of convergence for random sequences.

Recall that a random variable is a mapping from the sample space, Ω to the real line. So, we can write the random variable as $X_n(\omega)$, where the randomness depends on the outcome, ω. Write the sequence as $\{X_n(\omega)\}$ and we want to know if this sequence of random variables converges to a random variable, say $X(\omega)$. There are a variety of notions of convergence, and they are discussed in turn.

5.2.1 Almost sure convergence

For a given ω we can ask whether $\lim_{n\to\infty} X(\omega) = X(\omega)$ using the standard definition of a limit. If the set of ω for which this limit obtains has probability 1 then we say $X_n(\omega)$ converges to $X(\omega)$ almost surely (or with probability 1). This is written as

$$X_n \xrightarrow{as} X \text{ if } \mathbb{P}\{\omega | \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1.$$

5.2.2 Convergence in Probability

For any $\epsilon > 0$ we can calculate $p_n(\epsilon) = \mathbb{P}(|X_n - X| > \epsilon)$. If for any value of $\epsilon > 0$, the associated $\{p_n(\epsilon)\}$ sequence converges to 0, then we say that X_n converges in probability to X.

$$X_n \xrightarrow{p} X \text{ if for any } \epsilon > 0, \lim_{n\to\infty} p_n(\epsilon) = 0.$$

This is sometimes written as $plim X_n = X$.

5.2.3 Mean Square convergence

Let $ms_n = \mathbb{E}(X_n - X)^2$ denote the mean squared deviation of X_n from X. Then X_n converges to X in *mean square* if $\lim_{n\to\infty} ms_n = 0$. This sometimes written as

$$X_n \xrightarrow{ms} X \text{ if } \lim_{n\to\infty} \mathbb{E}\left[(X_n - X)^2\right] = 0.$$

5.2.4 Weak Convergence (Convergence in Distribution)

Suppose $F_{X_n}(x)$ is the CDF for X_n and $F_X(x)$ is the CDF for X, both evaluated at x. Then, in the limit X_n will have the same CDF as X if the function F_{X_n} converges to F_X. This notion of convergence is called *weak convergence*, or *convergence in Distribution* or *convergence in Law*. It is written as

$$X_n \Rightarrow X \text{ (or } X_n \xrightarrow{d} X) \text{ if } \lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all values of x where $F_X(.)$ is continuous.

To see the implication of restricting the definition to points of continuity of F_X, consider the following example. Suppose $X_n = 1/n$ with probability 1 and $X = 0$ with probability 1. Then $F_{X_n}(x) = \mathbf{1}\left(x \geq \frac{1}{n}\right)$, while $F_X(x) = \mathbf{1}(x \geq 0)$. Thus $F_{X_n}(0) = 0$ for all n, while $F_X(0) = 1$. Yet, X_n is getting close to X, so that for all probability statements about values other than $x = 0$, $F_{X_n}(x)$ is well approximated by $F_X(x)$ when n is large. In this sense, $X_n \Rightarrow X$.

The following are alternative equivalent ways to characterize weak convergence:

$$X_n \Rightarrow X \text{ if}$$

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all values of x where $F_X(.)$ is continuous, or

$$\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$$

for any continuous bounded function g.

5.2.5 Vector-valued random variables

If X_n is a vector, then $X_n \xrightarrow{as} X$ if each element of X_n converges *a.s.* to the corresponding element of X. Convergence in probability and mean square convergence is defined analogously. $X_n \Rightarrow X$ if the joint CDF of X_n converges to the joint CDF of X.

As it turns out, convergence in distribution obtains when $a'X_n \Rightarrow a'X$ for arbitrary non-stochastic vector a. This result is known as the Cramèr-Wold device. We'll see a version of this following the univariate central limit theorem shown below.

5.2.6 Relationship between the modes of convergence

- If $X_n \xrightarrow{ms} X$ then $X_n \xrightarrow{p} X$.
  - Using the fact that probabilities are non-negative, Chebychev's inequality, and convergence in mean square: $0 \leq \mathbb{P}(|X_n - X| \geq \epsilon) \leq \mathbb{E}(X_n - X)^2/\epsilon^2 \to 0$. So $\mathbb{E}(X_n - X)^2 \to 0$ implies that $\mathbb{P}(|X_n - X| \geq \epsilon) \to 0$.
- If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{p} X$
  To prove this we need to show that for any $\varepsilon > 0$ and $\delta > 0$ $\exists$ $N(\varepsilon, \delta)$ such that $\mathbb{P}(\omega \mid |X_n(\omega) - X(\omega)| > \varepsilon) < \delta$ for $n > N$. For each ω with $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ we can find a $N(\varepsilon, \omega)$ such that $|X_n(\omega) - X(\omega)| < \varepsilon$ for all $n > N(\varepsilon, \omega)$. Let $N(\varepsilon, \delta)$ be the largest of these values such that $\mathbb{P}(\omega \mid |X_n(\omega) - X(\omega)| < \varepsilon) > 1 - \delta$, for all $n > N(\varepsilon, \delta)$. (The existence of this value of N is guaranteed by the condition that $P\{\omega | \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1$). Then $\mathbb{P}(\omega \mid |X_n(\omega) - X(\omega)| > \varepsilon) < \delta$ for all $n > N$ as required.
- If $X_n \xrightarrow{p} X$ does not imply that $X_n \xrightarrow{a.s} X$. (See Amemiya, page 88 for a counterexample)
- If $X_n \xrightarrow{p} X$ then $X_n \Rightarrow X$. (The proof is in Rao, page 122, result ix)
- Note: $X_n \Rightarrow X$ does not imply that $X_n \xrightarrow{p} X$. (Think about the definitions.)

5.3 Slutsky's Theorem and the Continuous Mapping Theorem

- Slutsky's theorem (Rao page 122)
  - $X_n \Rightarrow X$ and $Y_n \xrightarrow{p} 0$ implies $X_n Y_n \xrightarrow{p} 0$
  - Let c be a constant and suppose $X_n \Rightarrow X$ and $Y_n \xrightarrow{p} c$
    - $X_n + Y_n \Rightarrow X + c$
    - $X_n Y_n \Rightarrow Xc$
    - $X_n / Y_n \Rightarrow X/c$, if $c \neq 0$
  - If $(X_n - Y_n) \xrightarrow{p} 0$ and $Y_n \Rightarrow Y$, then $X_n \Rightarrow Y$
- Continuous Mapping Theorem (Rao, page 124)

- Let $g(.)$ be a continuous function, then
  - $X_n \Rightarrow X$ implies that $g(X_n) \Rightarrow g(X)$
  - $X_n \xrightarrow{p} X$ implies that $g(X_n) \xrightarrow{p} g(X)$
  - If $(X_n - Y_n) \xrightarrow{p} 0$ and $Y_n \Rightarrow Y$, then $g(X_n) - g(Y_n) \xrightarrow{p} 0$.

5.4 O_p and o_p notation

Sometimes expressions contain many different sequences of random variables and *order of magnitude* notation is useful to keep track of which terms are most important.

A quick review from your first calculus course: Let $\{a_n\}_{n=1}^{\infty}$ and $\{g_n\}_{n=1}^{\infty}$ denote two sequences of real numbers. Recall that

$$a_n = o(g_n) \text{ if } \lim_{n\to\infty} \frac{a_n}{g_n} = 0$$

and

$$a_n = O(g_n) \text{ if } \exists \text{ a number } M < \infty \text{ such that } \left|\frac{a_n}{g_n}\right| < M \text{ for all } n$$

Similar notation is used for sequences of random variables. Suppose that $\{a_n\}_{n=1}^{\infty}$ is a sequence of random variables, then

$$a_n = o_p(g_n) \text{ if } \frac{a_n}{g_n} \xrightarrow{p} 0$$

and

$$a_n = O_p(g_n) \text{ if for any } \epsilon > 0, \exists \text{ a number } M \text{ such that } \mathbb{P}\left(\left|\frac{a_n}{g_n}\right| < M\right) > 1 - \epsilon \text{ for all } n.$$

Let $\{f_n\}$ and $\{g_n\}$ be sequences of real numbers and let $\{X_n\}$ and $\{Y_n\}$ be sequences of random variables. You can verify the following:

- If $X_n = o_p(f_n)$ and $Y_n = o_p(g_n)$, then
  - $X_n Y_n = o_p(f_n g_n)$
    - $|X_n|^s = o_p(f_n^s)$ for $s > 0$
    - $X_n + Y_n = o_p(\max\{f_n, g_n\})$
  - If $X_n = O_p(f_n)$ and $Y_n = O_p(g_n)$, then
    - $X_n Y_n = O_p(f_n g_n)$
    - $|X_n|^s = O_p(f_n^s)$ for $s > 0$
    - $X_n + Y_n = O_p(\max\{f_n, g_n\})$
  - If $X_n = o_p(f_n)$ and $Y_n = O_p(g_n)$, then
    - $X_n Y_n = o_p(f_n g_n)$

5.5 Law of Large Numbers

5.5.1 The sample mean

Let $X_1, X_2, \ldots$ be a sequence of random variables with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$, and $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$. The sample mean is $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$. Note that

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

If X_i are *i.i.d.*$N(\mu, \sigma^2)$, then $\overline{X} \sim N(\mu, \sigma^2/n)$. If X_i are not normally distributed, then the distribution of $\overline{X}$ is not normally distributed.

Laws of large numbers (LLN) and central limit theorems (CLT) yield large-sample approximations to the behavior of $\overline{X}$. Specifically, LLN yields the approximation

$$\overline{X} \approx \mu$$

and CLT yields the approximation

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

even when the X_i random variables are not normally distributed. The notation "$\overset{a}{\sim}$" means "approximately distributed".

5.5.2 A weak law of large numbers

Let $X_1, X_2, \ldots$ be a sequence of random variables with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$, and $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$. Then $\overline{X} \xrightarrow{p} \mu$.

Proof:

$$\mathbb{P}(|\overline{X} - \mu| > \epsilon) \leq \frac{\mathbb{E}[(\overline{X} - \mu)^2]}{\epsilon^2} = \frac{n^{-1}\sigma^2}{\epsilon^2} \to 0$$

where the first inequality follows from Chebyshev's inequality.

Exercise: (A straightforward extension) Let $X_1, X_2, \ldots$ be a sequence of random variables with $\mathbb{E}(X_i) = \mu_i$ and $\text{Var}(X_i) = \sigma_i^2$, and $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$. . Let $\overline{\sigma}_n^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$ and $\overline{\mu} = n^{-1}\sum_{i=1}^n \mu_i$ with $\lim_{n\to\infty} n^{-1}\overline{\sigma}_n^2 = 0$. Then $\overline{X} - \bar{\mu} \xrightarrow{p} 0$.

5.5.3 A strong law of large numbers

If $X_1, X_2, \ldots$ are *i.i.d.* with $\mathbb{E}(X) = \mu < \infty$, then $\overline{X} \xrightarrow{as} \mu$. (Proof: Rao, pages 114-115)

5.6 Central Limit Theorem

5.6.1 A CLT based on the moment generating function

Lemma (Continuity Theorem): Let Y_n be a sequence of random variables with moment generating function $M_n(t)$ that exist for some interval $t \in (-h, h)$, with $h>0$. Let $M(t)$ be the moment generating function of the random variable Y, which also exists on $(-h, h)$. If $\lim_{n\to\infty} M_n(t) = M(t)$ for all $t\in(-h, h)$, then $Y_n \Rightarrow Y$.

A Central Limit Theorem: Let $X_1, X_2, \ldots$ denote a sequence of *i.i.d.* random variables with $\mathbb{E}(X_i) = 0$ and $\text{Var}(X_i) = 1$ and MGF $M_X(t)$ that exists for $t\in(-h, h)$ for some $h>0$. Then $\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i \Rightarrow Z \sim \mathcal{N}(0, 1)$.

(*Note: This theorem assumes the existence of the MGF. A related CLT, but without this restriction, is presented in the next subsection.)*

Proof: Note, the first two moments of X imply that $M_X'(0) = \mathbb{E}(X) = 0$ and $M_X''(0) = \mathbb{E}(X^2) = 1$. And, using a mean value expansion:

$$M_X(t) = M_X(0) + tM_X'(0) + \frac{1}{2}t^2 M_X''(\tau) = 1 + \frac{1}{2}t^2 M_X''(\tau)$$

where τ is between 0 and t. Because $M_X''(t)$ is continuous, $\lim_{t\to 0} M_X''(\tau) = M_X''(0) = 1$.

Let $Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$, then from the properties of MGFs:

$$\begin{aligned} M_{Z_n}(t) &= [M_X(t/\sqrt{n})]^n \\ &= \left[1 + \frac{1}{2}\frac{t^2}{n}M_X''(\tau_n)\right]^n \end{aligned}$$

where τ_n is between 0 and $t/\sqrt{n}$.

Recall that if $\lim_{n\to\infty} a_n = a$, then $\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a$.

Thus

$$\lim_{n\to\infty} M_{Z_n}(t) = \lim_{n\to\infty}\left[1 + \frac{1}{2}\frac{t^2}{n}M_Y''(\tau_n)\right]^n = e^{\frac{1}{2}t^2}.$$

(Here, $\frac{1}{2}t^2 M_Y''(\tau_n) = a_n$ in the expression above and $lim_{n\to\infty}\frac{1}{2}t^2 M_Y''(\tau_n) = \frac{1}{2}t^2$.)

The result of the theorem follows from noting that $e^{\frac{1}{2}t^2}$ is the mgf of a standard normal.

Corollary: Let $Y_1, Y_2, ...$ denote a sequence of *i.i.d.* random with mean μ, variance σ^2 and MGF that exists for $t\in(-h, h)$ for some $h>0$. Then $\sqrt{n}(\overline{Y} - \mu) \Rightarrow \sigma Z\sim\mathcal{N}(0, \sigma^2)$.

Proof: Let $X_i = \frac{Y_i - \mu}{\sigma}$, and note that X_i satisfies the assumption of the CLT. Note $\sqrt{n}(\overline{Y} - \mu) = \sigma\frac{1}{\sqrt{n}}\sum_{i=1}^n X_i$ and the result follows directly.

5.6.2 A CLT based on the characteristic function

Characteristic Function (Rao, page 99-108):

Consider a random variable X with CDF $F(x)$. The characteristic function of X, denoted $C(t)$, is given by

$$C(t) = \mathbb{E}(e^{itX}) = \int e^{itx} dF(x)$$

where $i = \sqrt{-1}$. Thus, $C(t) = M(it)$. This change is useful because

$$e^{ix} = \cos(x) + i\sin(x)$$

so that $|e^{ix}| = 1$ for all x. This means that $C(t)$ will always exists, while $M(t)$ exists only for certain distributions.

Some useful results:

1. Let $\alpha_r = \mathbb{E}(X^r)$, which is assumed to exist. Then

$$\frac{d^r C(t)}{dt^r} = i^r \int x^r e^{itx} dF(x) \text{ exists}$$

2. Suppose α_r exists, then expanding $C(t)$ in a Taylor Series expansion about $C(0)$ yields:

$$C(t) = C(0) + \sum_{j=1}^{r} \alpha_j [\frac{(it)^j}{j!}] + O(t^{r+1})$$

   and $C(0) = 1$

3. Let $\phi(t) = \ln(C(t))$, then if α_r exists

$$\phi(t) = \sum_{j=0}^{r} \kappa_j [\frac{(it)^j}{j!}] + O(t^{r+1})$$

   where κ_j is called the $k'th$ cummulant. A direct calculation shows

   (a) $\kappa_0 = 0$

   (b) $\kappa_1 = \alpha_1 = \mu$

   (c) $\kappa_2 = \alpha_2 - \alpha_1^2 = \sigma^2$

4. If $X \sim N(\mu, \sigma^2)$, then $C(t) = M(it) = \exp[it\mu - \frac{t^2\sigma^2}{2}]$, and thus $\kappa_0 = 0$, $\kappa_1 = \mu$, $\kappa_2 = \sigma^2$, $\kappa_j = 0$ for $j > 2$.

5. Let $Z = X + Y$, where X and Y are independent, then $C_Z(t) = C_X(t)C_Y(t)$ and $\phi_Z(t) = \phi_X(t) + \phi_Y(t)$.

6. Let $Z = \delta X$, where δ is a constant. Then $C_Z(t) = C_X(\delta t)$

7. There is a 1-to-1 relation between $F(x)$ and $C(t)$

8. Let $C_n(t)$ denote the CF of X_n and $C(t)$ denote the CF of X. If $X_n \Rightarrow X$, then $C_n(t) \to C(t)$ for all t. Moreover, if $C_n(t) \to C(t)$ for all t and if $C(t)$ is continuous at $t = 0$, then $X_n \Rightarrow X$.

(Lindberg-Levy CLT): Let $X_1, X_2, ...$ denote a sequence of *iid* random variables with $\mathbb{E}(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2 \neq 0$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then

$$\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) \Rightarrow \mathcal{N}(0,1)$$

Proof (this parallels the proof using the MGF):

Let $Z_n = \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) = \sum_{i=1}^n (\frac{X_i - \mu}{\sqrt{n}\sigma})$. Since $\mathbb{E}(\frac{X_i-\mu}{\sigma}) = 0$ and $\text{Var}(\frac{X_i-\mu}{\sigma}) = 1$, the log-CF of $\frac{X_i-\mu}{\sigma}$ is

$$\phi(t) = -\frac{1}{2}t^2 + O(t^3)$$

so that Z_n has log-CF

$$\begin{aligned}\phi_{Z_n}(t) &= \sum_{i=1}^n \left[-(\frac{1}{2})(\frac{t}{\sqrt{n}})^2 + O[(\frac{t}{\sqrt{n}})^3\right] \\ &= -\frac{1}{2}t^2 + n \times O(\frac{t^3}{n^{3/2}}) \to -\frac{1}{2}t^2\end{aligned}$$

which is the log-CF of a $N(0,1)$ random variable. Since $-\frac{1}{2}t^2$ is continuous at $t = 0$, $Z_n \Rightarrow Z \sim \mathcal{N}(0,1)$.

5.6.3 Numerical Example

Suppose that X_i are *iid* Bernoulli random variables with parameter p. Then $\mathbb{E}(X) = p$ and $\text{Var}(X) = p(1-p)$. The CLT says that

$$\sqrt{n}\frac{(\overline{X} - p)}{(p(1-p))^{1/2}} \Rightarrow \mathcal{N}(0,1)$$

which implies that for large n

$$\sqrt{n}\frac{(\overline{X} - p)}{(p(1-p))^{1/2}} \overset{a}{\sim} \mathcal{N}(0,1)$$

where "$\overset{a}{\sim}$" means "approximately distributed as". Thus

$$\overline{X} \overset{a}{\sim} \mathcal{N}(p, \frac{p(1-p)}{n})$$

Suppose $p = 0.25$ and $n = 100$ then $\mathbb{P}(\overline{X} \leq .20) = \mathbb{P}(Y \leq 20)$ where $Y = \sum_{i=1}^{100} X_i$ is distributed binomial with $n = 100$ and $p = .25$. A direct calculation shows: $\mathbb{P}(Y \leq 20) = .14$. The normal approximation gives

$$\begin{aligned}\mathbb{P}(\overline{X} &\leq .20) = P\left(\frac{\overline{X} - .25}{(\frac{.25\times.75}{100})^{1/2}} \leq \frac{.20 - .25}{(\frac{.25\times.75}{100})^{1/2}}\right) \\ &\overset{a}{=} \mathbb{P}(Z \leq -1.155) = .12;\end{aligned}$$

5.6.4 Multivariate CLT

Suppose $X_i \sim iid(\mu, \Sigma)$. Then $\sqrt{n}(\bar{X} - \mu) \Rightarrow \mathcal{N}(0, \Sigma)$.

Sketch of proof: (We'll use the *Cramer-Wold device* discussed earlier).

Let $Y_n = \sqrt{n}\left(\overline{X} - \mu\right)$, with MGF, $M_{Y_n}(t)$. The goal is the show that $M_{Y_n}(t) \to e^{t'\Sigma t/2}$, the MGF for a $\mathcal{N}(0, \Sigma)$ random variable.

Note that $M_{Y_n}(t) = \mathbb{E}(e^{t'Y_n}) = \mathbb{E}(e^{n^{-1/2}\sum_{i=1}^{n}(t'X_i - t'\mu)})$.

But $(t'X_i - t'\mu) \sim iid(0, t'\Sigma t)$, so the proof to the univariate CLT shows $M_{Y_n}(t) = M_{t'Y_n}(1) \to e^{t'\Sigma t/2}$.

5.7 The δ-method

Let U_n denote a sequence of scalar random variables, and let $V_n = \sqrt{n}(U_n - a)$, where a is a constant. Let $g(.)$ be a continuously differentiable function. Suppose $V_n \Rightarrow V \sim N(0, \sigma^2)$. Then

$$\sqrt{n}(g(U_n) - g(a)) \Rightarrow \frac{dg(a)}{da}V \sim \mathcal{N}\left(0, \left[\frac{dg(a)}{da}\right]^2 \sigma^2\right).$$

Proof: By the mean value theorem

$$g(U_n) = g(a) + (U_n - a)\frac{dg(\widetilde{U}_n)}{da}$$

where $\widetilde{U}_n$ is between a and U_n. Since $V_n \Rightarrow V \sim \mathcal{N}(0, \sigma^2)$, then $n^{-\frac{1}{2}}V_n \xrightarrow{p} 0$ (by Slutsky's them) so that $U_n \xrightarrow{p} a$. Since $dg(.)/da$ is continuous $dg(\widetilde{U}_n)/da \xrightarrow{p} dg(a)/da$ (by the continuous mapping theorem). Thus

$$\sqrt{n}(g(U_n) - g(a)) = [\sqrt{n}(U_n - a)]\left[dg(\widetilde{U}_n)/da\right] \Rightarrow V\,[dg(a)/da] \sim \mathcal{N}\left(0, \left[\frac{dg(a)}{da}\right]^2 \sigma^2\right)$$

by Slutsky's theorem.

- A similar result holds for vectors. Let U_n denote a sequence of random vectors, and let $V_n = \sqrt{n}(U_n - a)$, where a is a constant vector. Let $g(.)$ be a continuously differentiable function. Suppose $V_n \Rightarrow V \sim \mathcal{N}(0, \Sigma)$. Then

$$\sqrt{n}(g(U_n) - g(a)) \Rightarrow \frac{dg(a)}{da}V \sim \mathcal{N}\left(0, \left[\frac{dg(a)}{da}\right]\Sigma\left[\frac{dg(a)}{da}\right]'\right).$$

An example: Suppose X_i is distributed $iidN(2, 1)$ for $i = 1, ..., n$. Then we know

$$\overline{X} \sim \mathcal{N}(2, \frac{1}{n})$$

or

$$\sqrt{n}(\overline{X} - 2) \sim \mathcal{N}(0, 1)$$

Let $Y = \overline{X}^2$.

In the δ-method, let $U_n = \overline{X}$, $a = 2$, $g(U_n) = Y = \overline{X}^2$, and $g(a) = a^2 = 4$. Then the delta-method implies

$$\sqrt{n}(Y - 2^2) \overset{a}{\sim} \mathcal{N}(0, (4)^2)$$

Suppose $n = 100$ and we want to know $\mathbb{P}(Y \leq 3.7)$. Using the properties of normal random variables, an exact calculation yields

$$\mathbb{P}(Y \leq 3.7) = .22.$$

The delta-method approximation yields

$$\begin{aligned} \mathbb{P}(Y &\leq 3.7) = \mathbb{P}\left(\frac{Y-4}{(16/100)^{1/2}} \leq \frac{3.7-4}{(16/100)^{1/2}}\right) \\ &\overset{a}{=} \mathbb{P}(Z \leq -.75) = .23 \end{aligned}$$

6 Estimators

6.1 Background and Examples

Let Y denote an $n \times 1$ vector of observations randomly sampled from a population. Because of random sampling, Y is a random variable. Denote the CDF of Y by $F(y, \theta)$, where θ is a parameter that characterizes the CDF. An *estimator* is a procedure for using the observed Y to guess the value of θ. Let $g(.)$ denote the "guessing" function.

That is, let $\hat{\theta} = g(Y)$ denote an *estimator* of θ. The realization of an estimator is an *estimate.*

Example: ***Method of Moments Estimators*** find $\hat{\theta}$ so that sample moments of Y match the population moments of Y.

Let Y_i , $i = 1, 2, \ldots, n$ be scalar *i.i.d.*$N(\mu, \sigma^2)$ random variables. Then $\mathbb{E}(Y_i) = \mu$ and $\mathbb{E}[(Y_i - \mu)^2] = \sigma^2$. The method-of-moment estimators of μ and σ^2 are

$$\hat{\mu} = n^{-1}\sum_{i=1}^{n} Y_i \text{ and } \hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{\mu})^2$$

which use sample moments as estimators of corresponding population moments.

Because Y is random, the value of $\hat{\theta}(Y)$ is random. The mean of $\hat{\theta}$ is $\mathbb{E}(\hat{\theta})$ and the variance is $\text{Var}(\hat{\theta}) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}(\hat{\theta}))^2\right]$, where the expectation is computed over Y, or equivalently over the values of $\hat{\theta}$.

An estimator is said to be *unbiased* if $E(\hat{\theta}) = \theta$. The *bias* in the estimator is defined as

$$\text{Bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta.$$

A natural question is what constitutes a "good" estimator. One way to answer this question is to define a ***Loss Function***, say $L(\hat{\theta}, \theta)$ which shows the loss that occurs when $\hat{\theta}$ is used when taking some action, when, in fact, the true value of the parameter is θ. (Think of Loss as the negative of utility). Good estimators are estimators that make expected loss as small as possible. (This is analogue of choosing the estimator that maximizes expected utility.) We will discuss this more formally in a later section (as an introduction to our discussion of Bayes methods), for now we focus on a special case:

$$L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2 = \text{quadratic loss}$$

which is called quadratic loss. Expected quadratic loss is *mean squared error:*

$$\mathbb{E}\left(L(\hat{\theta}, \theta)\right) = \mathbb{E}\left((\hat{\theta} - \theta)^2\right) = \text{mse}(\hat{\theta})$$

where the expectation is taken over Y.

A convenient decomposition for *mse* is obtained by writing $(\hat{\theta} - \theta) = [\hat{\theta} - \mathbb{E}(\hat{\theta})] + [\mathbb{E}(\hat{\theta}) - \theta)]$, so that

$$\begin{aligned}\mathbb{E}\left((\hat{\theta} - \theta)^2\right) &= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right)^2\right] + \left(\mathbb{E}(\hat{\theta}) - \theta\right)^2 + 2 \times \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}(\hat{\theta})\right)\left(\mathbb{E}(\hat{\theta}) - \theta\right)\right] \\ &= \text{Var}(\hat{\theta}) + \left[\text{Bias}(\hat{\theta})\right]^2.\end{aligned}$$

noting that the final term vanishes because $\mathbb{E}[\hat{\theta} - \mathbb{E}(\hat{\theta})] = 0$.

Thus for an unbiased estimator, $\text{mse}(\hat{\theta}) = \text{Var}(\hat{\theta})$.

- **Exercise***:* Y_i is *i.i.d.* $(\mu, 1)$. Consider the estimator $\hat{\mu}_1 = \overline{Y}$ and $\hat{\mu}_2 = \frac{1}{2}\overline{Y}$.
  - Derive $\text{mse}(\hat{\mu}_1)$ and $\text{mse}(\hat{\mu}_2)$.
  - Which is estimator has the lowest mse?

A final bit of jargon: An estimator is said to be *consistent* if

$$\hat{\theta} \xrightarrow{p} \theta.$$

6.2 The Likelihood Function

An important tool for statistical inference is the *Likelihood Function.* Suppose that the random variable Y has probability density function$f(y|\theta)$, where the notation makes it clear that the pdf depends on θ. The likelihood function is

$$\mathcal{L}(\theta, Y) = f(Y|\theta).$$

This is the density of Y evaluated at $y = Y$. Because the density is evaluated at a random point, $\mathcal{L}(\theta, Y)$ is random. When viewed as a function of θ, $\mathcal{L}(\theta, Y)$ is a random function.

It is useful to derive a few identities involving the likelihood function. For simplicity, suppose θ is a scalar.

- Because $f(y|\theta)$ is a pdf
$$1 = \int f(y|\theta)dy$$
  - Differentiating both sides, and assuming the support of Y does not depend on θ
$$0 = \int \frac{\partial f(y|\theta)}{\partial \theta}dy$$
  - Let
$$S(\theta, y) = \frac{\partial \ln f(y|\theta)}{\partial \theta} = \frac{1}{f(y|\theta)}\frac{\partial f(y|\theta)}{\partial \theta}$$
  which is called a ***Score function***. (When I want to emphasize dependence of this function on θ I will write the function as $S(\theta)$.)
  - Note
$$\frac{\partial f(y|\theta)}{\partial \theta} = S(\theta, y) \times f(y|\theta)$$
  so that
$$0 = \int \frac{\partial f(y|\theta)}{\partial \theta}dy = \int S(\theta, y)f(y|\theta)dy = \mathbb{E}[S(\theta, Y)].$$
  Evidently the Score function has an expected value of 0. (Note the randomness in the score function comes from evaluating the function at the random value Y.)

- Differentiating again, yields:
$$0 = \int \frac{\partial S(\theta, y)}{\partial \theta}f(y|\theta)dy + \int S(\theta, y)^2 f(y|\theta)dy$$
  - So that
$$-\mathbb{E}\left[\frac{\partial S(\theta, Y)}{\partial \theta}\right] = \mathbb{E}\left[S(\theta, Y)^2\right] = \text{Var}\left[S(\theta, Y)\right]$$
  - Let
$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial S(\theta, Y)}{\partial \theta}\right] = -\mathbb{E}\left(\frac{\partial^2 \ln[f(Y|\theta)]}{\partial \theta^2}\right) = \mathbb{E}\left[S(\theta, Y)^2\right] = \text{Var}\left[S(\theta, Y)\right]$$
  which is called the ***Information***.

6.2.1 The Cramer-Rao inequality and unbiased estimators

Let $\hat{\theta} = g(Y)$ denote an *unbiased* estimator of θ. Then

$$\theta = \int g(y)f(y|\theta)dy$$

Differentiating both sides with respect to θ yields:

$$1 = \int g(y) \frac{\partial f(y|\theta)}{\partial \theta} dy = \int g(y) S(\theta, y) f(y|\theta) dy$$

with $\hat{\theta} = g(Y)$ this implies

$$\mathbb{E}\left[\hat{\theta} \times S(\theta, Y)\right] = \text{Cov}[\hat{\theta}, S(\theta, Y)] = 1$$

where the first equality holds because $\mathbb{E}(S(Y, \theta)) = 0$.

But, for any two random variables, say U and V, $\text{Var}(U) \times \text{Var}(V) \geq [\text{Cov}(U, V)]^2$. Thus

$$\text{Var}(\hat{\theta}) \times \text{Var}\left(S(\theta, Y)\right) \geq 1$$

This yields

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\text{Var}\left(S(\theta, Y)\right)} = \mathcal{I}(\theta)^{-1}$$

which is the Cramer-Rao inequality.

6.2.2 Extensions to vector valued θ

Analogous results obtain when θ is a $k \times 1$ vector:

- $S(\theta, Y)$ is a $k \times 1$ score vector with $\mathbb{E}(S(\theta, Y)) = 0$
  - $\text{Var}(S(\theta, Y)) = \mathbb{E}(S(\theta, Y)S(\theta, Y)') = -\mathbb{E}(\frac{\partial S(\theta, Y)}{\partial \theta'}) = \mathcal{I}(\theta)$ which is the $k \times k$ Information matrix
  - If $\widehat{\theta}$ is an unbiased estimator, then $\mathbb{E}[(\theta - \widehat{\theta})(\theta - \widehat{\theta})'] \geq \mathcal{I}(\theta)^{-1}$. (Where $\geq$ means that the difference between the lhs and rhs matrices is positive semi-definite.)

6.3 Maximum Likelihood Estimators

- Let $Y_1, Y_2, ..., Y_n$ be *iid*, each with density $f(y|\theta)$. Then write the likelihood as

  $$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(Y_i|\theta)$$

  where I supressed its dependence on Y for convenience.

- Let

  $$L_n(\theta) = \ln(\mathcal{L}_n(\theta))$$

  denote the log-likelihood function.

- Suppose that θ is a $k \times 1$ vector and let

$$s_i(\theta) = \frac{\partial \ln f(Y_i|\theta)}{\partial \theta}$$

and

$$S_n(\theta) = \sum_{i=1}^{n} s_i(\theta)$$

denote the Score. (Note that these functions are evaluated at the random value Y. For notational simplicity I write $s_i(\theta)$ instead of $s_i(\theta, Y_i)$, etc.)

- Let

$$\mathcal{I}(\theta) = -\mathbb{E}(\frac{\partial s_i(\theta)}{\partial \theta'}) = \mathbb{E}(s_i(\theta)s_i(\theta)'),$$

denote the information in the i^{th} observation. Note that this does not depend on i when Y_i are *iid*. Also,

$$\mathbb{E}(S_n(\theta)S_n(\theta)') = \text{var}\, S_n(\theta) = n\mathcal{I}(\theta)$$

because the observations are *iid*.

- Let $\widehat{\theta}_{mle}$ solve

$$\max_{\theta} L_n(\theta)$$

or equivalently solve $\max_\theta \mathcal{L}_n(\theta)$.

- Notation: In this formulation θ is the argument of the function. To highlight the 'true' value of θ that describes the pdf of Y, denote this true value by θ_0. (Note, the arguments above about the properties of $s(\theta)$, $\mathcal{I}(\theta)$ were for $\theta = \theta_0$.)

- Some asymptotic properties of MLEs
  Given a set of "regularity" conditions:

$$\widehat{\theta}_{mle} \xrightarrow{p} \theta_0$$

and

$$\mathcal{I}(\theta_o)^{1/2}\sqrt{n}(\widehat{\theta}_{mle} - \theta_0) \Rightarrow N(0, I)$$

so that

$$\widehat{\theta}_{mle} \overset{a}{\sim} N(\theta_0, n^{-1}\mathcal{I}(\theta_0)^{-1}).$$

- These results say that the MLE is
  - Consistent
  - Approximately normal
  - Achieve Cramer-Rao lower bound.

Let me now sketch the proof to these results: (for simplicity assuming θ is a scalar)

***Sketch of consistency proof under** iid **sampling**:*

Let

$$C(\theta) = \mathbb{E}_{\theta_0}[\ln(f(Y|\theta)) - \ln(f(Y|\theta_0))]$$

where θ_0 is the true value of θ and $\mathbb{E}_{\theta_0}$ means taking the expected value using the density $f(y|\theta_0)$.

Let's show that $C(\theta) \leq 0$.

To see this, note that

$$\mathbb{E}_{\theta_0} \ln[\frac{f(Y|\theta)}{f(Y,|\theta_0)}] \leq \ln \mathbb{E}_{\theta_0}[\frac{f(Y|\theta)}{f(Y|\theta_0)}] = \ln(1) = 0$$

where the first inequality follows from Jensen's inequality since the log function is concave.

Clearly then $C(\theta)$ is maximized at $\theta = \theta_o$. When this maximum is unique, that is, when $C(\theta) < 0$ for $\theta \neq \theta_o$ then we say that the parameter θ is *identified* when $\theta = \theta_0$ (or, more simply θ_o is identified).

To complete the discussion of consistency, assume that $C(\theta) < 0$ for $\theta \neq \theta_o$ and that

$$C_n(\theta) = n^{-1} \sum \{\ln(f(Y_i|\theta)) - \ln(f(Y_i|\theta_0))\} \xrightarrow{p} C(\theta)$$

uniformly in θ. (This is *Uniform* LLN result —- see Gallant, A. R. (1997), *An Introduction to Econometric Theory*, Princeton University Press., page 135). Thus the maximizer of $C_n(\theta)$ converges to the maximizer of $C(\theta)$, which we just showed was θ_o. Thus the maximizer of $n^{-1} \sum \ln(f(Y_i, \theta)) = n^{-1} L_n(\theta)$ converges to θ_0. This means that the MLE is consistent.

Sketch of Asymptotic Normality:

- First

$$\frac{1}{\sqrt{n}} S_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} s_i(\theta_0) \Rightarrow \mathcal{N}(0, \mathcal{I}(\theta_0))$$

  follows immediately from applying the CLT to $n^{-1/2} \sum s_i(\theta_0)$.

  - Next

$$S_n(\widehat{\theta}_{mle}) = S_n(\theta_0) + \frac{\partial S_n(\widetilde{\theta})}{\partial \theta}(\widehat{\theta}_{mle} - \theta_0)$$

    where $\widetilde{\theta}$ is between θ_0 and $\widehat{\theta}_{mle}$. Since $S_n(\widehat{\theta}_{mle}) = 0$,

$$\sqrt{n}(\widehat{\theta}_{mle} - \theta_o) = \left[-\frac{1}{n} \frac{\partial S_n(\widetilde{\theta})}{\partial \theta} \right]^{-1} [\frac{1}{\sqrt{n}} S_n(\theta_0)].$$

    Also,

$$\left[-\frac{1}{n} \frac{\partial S_n(\widetilde{\theta})}{\partial \theta} = -\frac{1}{n} \sum_{i=1}^{n} \frac{\partial s_i(\widetilde{\theta})}{\partial \theta} \right] \xrightarrow{p} \mathbb{E} \left[-\frac{\partial s_i(\theta_0)}{\partial \theta} \right] = \mathcal{I}(\theta_0)$$

    (uniform LLN, CMT, Consistency of $\widehat{\theta}_{mle}$).

– Thus

$$\sqrt{n}(\widehat{\theta}_{mle} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$$

by Slutsky's Theorem.

– These results also hold for vector $\widehat{\theta}_{mle}$ and vector values $S_n(\theta_0)$, etc.

6.3.1 Examples (to be worked out in class)

- $iid\ \mathcal{N}(\mu, \sigma^2)$
- Binomial (n, p)
- Uniform $[0, \theta]$

6.4 Method of Moment Estimators

Suppose $Y_i, i = 1, ..., n$ is a sequence of $iid(\mu, \Sigma)$ random $l \times 1$ vectors.

- The method of moments estimator of μ is

$$\widehat{\mu}_{mm} = n^{-1} \sum Y_i.$$

From the LLN and CLT, we have

$$\widehat{\mu}_{mm} \xrightarrow{as} \mu$$

and

$$\sqrt{n}(\widehat{\mu}_{mm} - \mu) \Rightarrow \mathcal{N}(0, \Sigma).$$

Notice that the estimator can be constructed and these properties obtained without knowing very much about the probability distribution of Y.

- Now suppose that $\mu = h(\theta_o)$ where μ is $l \times 1$, θ_o is $k \times 1$ with $k \leq l$. Our goal is the estimate θ_o. A Method of Moments estimator can be obtained by solving

$$\min_{\theta} J_n(\theta)$$

where

$$\begin{aligned} J_n(\theta) &= [\frac{1}{n}\sum_{i-1}^{n}(Y_i - h(\theta))]'[\frac{1}{n}\sum_{i-1}^{n}(Y_i - h(\theta))] \\ &= (\overline{Y} - h(\theta))'(\overline{Y} - h(\theta)) \end{aligned}$$

Let $\widehat{\theta}_{mm}$ denote the method of moments estimator. The properties of $\widehat{\theta}_{mm}$ can be derived in a way that parallels the discussion of the maximum likelihood estimator.

– Consistency follows by arguing that $J_n(\theta) \to J(\theta)$ and that $J(\theta)$ is minimized at $\theta = \theta_o$.

– Asymptotic normality is proved using the following steps

1. (*Asymptotic normality of gradient*) Show the gradient evaluated at θ_o satisfies a CLT. The gradient is
$$g_n(\theta) = \frac{\partial J_n(\theta)}{\partial \theta} = -2[\frac{\partial h(\theta)}{\partial \theta'}]'(\overline{Y} - h(\theta))$$
so that
$$\sqrt{n}g_n(\theta_o) = -2[\frac{\partial h(\theta_o)}{\partial \theta'}]'[\sqrt{n}(\overline{Y} - h(\theta_o))] \Rightarrow \mathcal{N}\left(0, 4\left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]'\Sigma\left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]\right)$$
2. (*Mean value expansion*) Linearize $g_n(\widehat{\theta}_{mm})$ around $g_n(\theta_o)$ and solve for $\widehat{\theta}_{mm}$.
$$g_n(\widehat{\theta}_{mm}) = g_n(\theta_o) + \frac{\partial g_n(\widetilde{\theta})}{\partial \theta'}(\widehat{\theta}_{mm} - \theta_o)$$
where $\widetilde{\theta}$ is between θ_o and $\widehat{\theta}$.
3. (*Asymptotic behavior of Hessian*) Show
$$\frac{\partial g_n(\widetilde{\theta})}{\partial \theta'} \xrightarrow{p} 2H$$
where
$$H = \left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]'\left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]$$
a constant, non-singular matrix.
$$\frac{\partial g_n(\theta)}{\partial \theta'} = 2\left[\frac{\partial h(\theta)}{\partial \theta'}\right]'\left[\frac{\partial h(\theta)}{\partial \theta'}\right] + m_n(\theta)(\overline{Y} - h(\theta))$$
where $m_n(\theta)$ denotes the derivatives of $\partial h(\theta)/\partial\theta'$ with respect to θ. Evaluating this expression at $\theta = \theta_o$, the second term vanishes in probability and the first term is $2H$.
4. (*Rearrange and use Slutsky's theorem*) Write
$$\sqrt{n}(\widehat{\theta}_{mm} - \theta_o) = \left[\frac{\partial g_n(\widetilde{\theta})}{\partial \theta'}\right]^{-1}[\sqrt{n}g_n(\theta_o)] \Rightarrow \mathcal{N}\left(0, H^{-1}\left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]'\Sigma\left[\frac{\partial h(\theta)}{\partial \theta'}\right]H^{-1}\right)$$
so that
$$\widehat{\theta}_{mm} \overset{a}{\sim} \mathcal{N}(\theta_o, V_n)$$
where
$$V_n = (1/n)H^{-1}\left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]'\Sigma\left[\frac{\partial h(\theta)}{\partial \theta'}\right]H^{-1}$$

7 Sufficient Statistics

A key task in statistics is data reduction, by which I mean summarizing the information in a large data set using a small number of "statistics" (functions of the data). A useful concept in this regard is a sufficient statistic. Loosely speaking, if θ is an unknown parameter characterizing the probability density of

$(Y_1, Y_2, \ldots, Y_n)$, then a statistic $S(Y_1, Y_2, \ldots, Y_n)$ is sufficient for θ, if $S(Y_1, Y_2, \ldots, Y_n)$ summarizes all of the information in $(Y_1, Y_2, \ldots, Y_n)$ about θ. Thus, if interest focuses on the value of θ, one only needs to retain the statistic $S(Y_1, Y_2, \ldots, Y_n)$, and the rest of the data can be discarded.

To formalize this, let $Y = (Y_1, Y_2, \ldots, Y_n)'$ denote the vector of random variables under study, and $S(Y)$ denote a statistic. Write the pdf of Y as$f_Y(y|\theta)$, the pdf of S as $f_S(s|\theta)$ and the conditional pdf of Y given $S = s$ as $f_{Y|S}(y|s, \theta)$, where each has been written to emphasize that the density depends on θ. The statistic S is *sufficient* if $f_{Y|S}(y|s, \theta) = f_{Y|S}(y|s)$, that is the conditional density of Y given S does not depend on θ.

Two examples:

- Suppose Y_1 and Y_2 are *iid* Bernoulli random variables with $\mathbb{P}(Y_i = 1) = \theta$. Let $S = Y_1 + Y_2$, and note that S can take on the values 0, 1, or 2. If $S = 0$, then $Y_1 = Y_2 = 0$, so $\mathbb{P}(\{0, 0\}|S = 0) = 1$; similarly if $S = 2$, then $Y_1 = Y_2 = 1$, so $\mathbb{P}(\{1, 1\}|S = 2) = 1$. If $S = 1$, then one of Y_1 or Y_2 is equal to 1 and the other is equal to 0, with both events being equally likely, thus $\mathbb{P}(\{0, 1\}|S = 1) = \mathbb{P}(\{1, 0\}|S = 1) = 0.5$. In all of these cases $\mathbb{P}(y|S = s)$ does not depend on the value of θ, so S is a sufficient statistic.
- Suppose $Y_i \sim iid\ N(\mu, 1)$, for $i = 1, \ldots, n$. Equivalently $Y \sim \mathcal{N}(l\mu, I_n)$, where l is an $n \times 1$ vector of 1's. Let $S(Y) = \overline{Y} = n^{-1} \sum_{i=1}^n Y_i$ denote the sample mean. Using the conditional normal formula, the pdf of $Y|S$ is normal with mean vector $l\mu + l \times (n^{-1}/n^{-1})(\overline{Y} - \mu) = l\overline{Y}$, and covariance matrix $I_n - n^{-1}ll'$. Because this conditional distribution does not depend on μ, $\overline{Y}$ is sufficient for μ.

7.1 2 useful results for Sufficient Statistics

Factorization Theorem: Let $f_Y(y|\theta)$ denote the density of Y. Then S is sufficient for θ if and only if $f_Y(y|\theta)$ can be factored as $f_Y(y|\theta) = h(y)g(s|\theta)$, where $h(.)$ does not depend on θ. (The proof is straightforward, and you can see it in the HCM textbook).

This theorem is useful for two reasons:

- As a mechanical matter it provides another way to check that a candidate S is sufficient. For example, in the $Y_i \sim iid N(\mu, 1)$ example, the pdf $f_Y(y|\mu)$ is proportional to $\exp\left[-\frac{1}{2}\sum_{i=1}^n (y_i - \mu)^2\right]$, but $\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n \left[(y_i - \bar{Y}) + (\bar{Y} - \mu)\right]^2 = \sum_{i=1}^n (y_i - \bar{Y})^2 + n(\bar{Y} - \mu)^2$ so that the pdf factors as required, thus showing that $\overline{Y}$ is sufficient for μ.
- Because $f_Y(Y|\theta) = h(Y)g(S|\theta)$ is the likelihood, any likelihood inference will be based on $g(S|\theta)$ and only involve the data through the sufficient statistic. Thus, for example, the MLE of θ is

$$\hat{\theta}(Y) = \arg\max_\theta f(Y|\theta) = \arg\max_\theta g(S|\theta) = \hat{\theta}(S).$$

Rao-Blackwell Theorem:

Background: Suppose Y is a random variable with $\mathbb{E}(Y) = \mu$ and variance σ_Y^2. Let X denote another random variable and let $\mu(x) = \mathbb{E}(Y|X = x)$. From the law of iterated expectations we know that $\mathbb{E}(\mu(X)) = \mu$. Further, writing $Y = \mu(X) + (Y - \mu(X))$, recognize that two terms on the rhs of this expression are uncorrelated (from the law of iterated expectations), so that

$$\sigma_Y^2 = \text{Var}(\mu(X)) + \text{Var}(Y - \mu(X)).$$

This implies that $\text{Var}(\mu(X)) \leq \sigma_Y^2$.

Application: Suppose $\hat{\theta}(Y)$ is an unbiased estimator of θ, so that $\theta = \mathbb{E}[\hat{\theta}(Y)]$ and let S be a sufficient statistic for θ. Using the law of iterated expectations:

$$\theta = \mathbb{E}[\hat{\theta}(Y)] = \mathbb{E}[\mathbb{E}[\hat{\theta}(Y)|S]] = \mathbb{E}[\tilde{\theta}(S)],$$

where $\tilde{\theta}(S) = \mathbb{E}[\hat{\theta}(Y)|S]$.

Note, while $\tilde{\theta}(S) = \mathbb{E}[\hat{\theta}(Y)|S]$ is a function of S, it is not a function of θ, because the conditional distribution of Y given S does not depend on θ. Thus, $\tilde{\theta}(S)$ is an estimator in the sense that it depends on the data (S) but not the unknown value of θ.

Now, $\mathbb{E}(\hat{\theta}(Y)) = \mathbb{E}(\mathbb{E}(\hat{\theta}(Y)|S)) = \mathbb{E}(\tilde{\theta}(S))$ from the law of iterated expectations, so $\tilde{\theta}(S)$ is unbiased, and from our result above, it has a variance that is weakly smaller than the variance of $\hat{\theta}(Y)$. Thus, the MSE of an unbiased estimator, $\hat{\theta}(Y)$, can be reduced, by computing the expected value of the estimator conditional on a sufficient statistic; that is by computing $\tilde{\theta}(S)$.

Example: $Y_i \sim i.i.d.N(\mu, 1)$. Let $S = \overline{Y}$. Let $\hat{\mu} = Y_1$. This estimator is unbiased and has a variance equal to 1. Now $E(\hat{\mu}|S) = \tilde{\mu} = E(Y_1|\overline{Y}) = \mu + \frac{1/n}{1/n}\left(\overline{Y} - \mu\right) = \overline{Y}$, so that the variance of the estimator $\tilde{\mu}$ is $1/n$.

8 Hypothesis Tests

We will first cover hypothesis testing in a specific (important) example. We'll then move on to a more general discussion of the hypothesis testing problem.

8.1 Wald Tests

Suppose we are interested in a $k \times 1$ vector of parameters, say θ, that characterizes the probability distribution of Y. Also, suppose we have an estimator of θ, say $\hat{\theta}$, where (perhaps based on an asymptotic approximation)

we have $\hat{\theta}\sim N_k(\theta,\Omega)$ where we know Ω but we don't know the value of θ. Suppose there are two competing hypotheses:

$$H_o:\theta=\theta_0 \ \text{ (where } \theta_0 \text{ is a known value of } \theta)$$

and

$$H_a:\theta\neq\theta_0.$$

In the jargon of hypothesis testing, H_o is called the null hypothesis and H_a is called the alternative hypothesis.

How might we decide between H_o and H_a? The standard procedure is based on the following logic:

If $\theta=\theta_0$, then $\hat{\theta}$ should be close to θ_0, that is $\left\|\hat{\theta}-\theta_0\right\|$ is likely to be small.

But if $\theta\neq\theta_0$, then $\left\|\hat{\theta}-\theta_0\right\|$ is likely to be large.

We are helped with "likely" and "small" and "large" because we know that

$\hat{\theta}\sim\mathcal{N}(\theta,\Omega)$. Thus, we can form a *test-statistic*, say ξ, as

$$\xi=(\hat{\theta}-\theta_0)'\Omega^{-1}(\hat{\theta}-\theta_0).$$

Under $H_o:\xi\sim\chi^2_k$, where k is the number of elements in θ.

Under H_a: ξ will have a distribution that puts more mass on larger values than under the χ^2_k distribution because the wrong mean has been used for the distribution of $\hat{\theta}$.

This gives rise to a decision rule of the form:

$$\text{Choose } H_o \text{ if } \xi\leq \text{cv and Choose } H_a \text{ if } \xi>\text{cv}\,.$$

where the number cv is called the ***critical value*** of the test.

The critical value is chosen so the probability of incorrectly choosing H_a (that is incorrectly "rejecting" H_o) is set equal to a pre-specified value (typically 1%, 5%, or 10%). This probability is called the ***size*** of the test.

Suppose we want the size of the test to be α, how do we choose cv? That's easy: cv solves

$$\mathbb{P}(\xi>\text{cv}\,|\theta=\theta_o\,)=\alpha$$

so that cv is the $1-\alpha$ percentile of the χ^2_k distribution.

The ***power*** of the test is defined as $\mathbb{P}(\xi > \text{cv} \,|H_a \text{ is true})$. But because H_a includes many values of θ (the alternative hypothesis is said to be ***composite***), the power will be different for the different values of θ included in H_a.

We can say a few general things, however. Suppose the normal distribution for $\hat{\theta}$ was based on a CLT argument, say $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, V)$. In this case we know that $\Omega = n^{-1}V$, and $\Omega^{-1} = nV^{-1}$. In this case the test statistic is

$$\xi = n(\hat{\theta} - \theta_0)' V^{-1} (\hat{\theta} - \theta_0)$$

so that if the mean of $\hat{\theta}$ is equal to a fixed constant that differs from θ_0, then $\xi \to \infty$. In this case $\mathbb{P}(\xi > \text{cv}) \to 1$ for any fixed value of cv. The test therefore has power $= 1$ for any (fixed) value of θ under the alternative.

When power $\to 1$, a test is said to be ***consistent***.

A test of the form ξ is called a ***Wald test***. It's basic form can generalized in several ways.

8.1.1 Hypotheses involving linear functions of θ

Suppose the null does not involve restrict all the elements of θ, but rather the linear combinations, say $R\theta$, where R is a $j \times k$ matrix with rank j. Suppose the null and alternative are:

$$H_o : R\theta = r_0 \text{ where } r_0 \text{ is a known value and } H_a : R\theta \neq r_0.$$

If $\hat{\theta} \sim N(\theta, \Omega)$, then $R\hat{\theta} \sim N(R\theta, R\Omega R')$, so the hypotheses can be tested using the Wald statistic

$$\xi = (R\hat{\theta} - r_0)' (R\Omega R')^{-1} (R\hat{\theta} - r_0)$$

which will be distributed as a χ^2_j random variable under the null. (Thus, the critical value will be the $1 - \alpha$ percentile of the χ^2_j distribution.)

8.1.2 Hypotheses involving nonlinear functions of θ:

When the normal approximation is motivated by the CLT: $\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, V)$, then nonlinear functions can be handled via the δ-method. Thus, consider the j non-linear functions $R(\theta)$, with null and alternative:

$$H_o : R(\theta) = r_0 \text{ where } r_0 \text{ is a known value and } H_a : R(\theta) \neq r_0.$$

The delta-method implies

$\sqrt{n}(R(\hat{\theta}) - R(\theta)) \Rightarrow \mathcal{N}(0, HVH')$ where $H = \frac{\partial R(\theta)}{\partial \theta'}$

so that $R(\hat{\theta}) \overset{a}{\sim} \mathcal{N}(R(\theta), \tilde{\Omega})$, where $\tilde{\Omega} = n^{-1}HVH'$.

The Wald statistic becomes $\xi = (R(\hat{\theta}) - r_0)'\tilde{\Omega}^{-1}(R(\hat{\theta}) - r_0)$.

8.2 Neyman-Pearson Tests

8.2.1 A more general framework

Suppose that we have two competing hypotheses about the distribution of a random variable Y.

Hypothesis 1 will be called the ***Null*** and is written as

$$H_o : Y \sim F_o$$

Hypothesis 2 will be called the ***Alternative*** and is written as

$$H_a : Y \sim F_a$$

It is useful to categorize the errors in inference that we can make

We can say that H_a is true when H_o is true. This is called ***Type 1 Error***

We can say that H_o is true when H_a is true. This is called ***Type 2 Error***

We will consider tests based on realizations of the random variable Y. Specifically, we will define a region of the sample space, say W, and reject H_o (Accept H_a) if $Y \in W$, and otherwise reject H_a (Accept H_o). W is called a ***critical region*** .

Our goal is to find procedures for choosing W to minimize the probability of making errors. However, we can also always make the probability of type 1 error smaller by making W smaller, and make the probability of type 2 error smaller by making W larger. A standard procedure in test design (procedures for choosing W) is to fix the probability of type 1 error at some pre-specified value, and choose the critical region to minimize the probability of type 2 error.

The pre-chosen probability of type 1 error is called the ***size*** of the test.

The probability of accepting H_a when H_a is true is called the ***power*** of the test: $power = 1 - \mathbb{P}(type\ 2\ error)$.

The hypothesis testing design problem is: Choose a test to maximize *power* subject to a pre-specified *size*.

8.2.2 Likelihood Ratio Tests and the Neyman-Pearson Lemma

The Neyman-Pearson Lemma says that power is maximized, subject to a size constraint, by choosing the critical region based on the likelihood ratio

$$LR(Y) = \frac{\mathcal{L}_a(Y)}{\mathcal{L}_o(Y)}$$

where $\mathcal{L}_a(Y)$ and $\mathcal{L}_o(Y)$ are the likelihoods under the alternative and null, respectively. The critical region for a test with size α is

$$W = \{y \,|LR(y) > \text{cv}\,\}$$

where cv is chosen so that

$$\mathbb{P}[LR(Y) > \text{cv}\,|Y \sim F_o] = \alpha$$

The proof of this remarkable result is pretty easy.

Suppose the random variables have a continuous distribution with density f_a and f_o under the alternative and null. Then $\mathcal{L}_o(Y) = f_o$ and $\mathcal{L}_a(Y) = f_a$.

Let W denote the NP critical region. Let X denote any other critical region with size α. Decompose W and X as

$$W = (W \cap X) \cup (W \cap X_\alpha^c)$$

and

$$X = (X \cap W) \cup (X \cap W_\alpha^c)$$

Now (because tests have size α):

$$\alpha = \int_W f_o(y)dy = \int_X f(y)dy$$

so that

$$\alpha = \int_{W\cap X} f_o(y)dy + \int_{W\cap X^c} f_o(y)dy = \int_{X\cap W} f_o(y)dy + \int_{X\cap W^c} f_o(y)dy$$

which implies

$$\int_{W\cap X^c} f_o(y)dy = \int_{X\cap W^c} f_o(y)dy$$

But, for any $Y \in W$ (and hence for any $Y \in (W \cap X^c)$), $f_a(Y) > \text{cv}\, f_o(Y)$, and for any $Y \in W^c$ (and hence for any $Y \in (X \cap W^c)$) $f_a(Y) \leq \text{cv}\, f(Y)$. Thus

$$\int_{W_\alpha \cap X^c} f_a(y)dy \geq \int_{X_\alpha \cap W^c} f_a(y)dy.$$

Adding $\int_{X_\alpha \cap W_\alpha} f_a(y)dy$ to both sides of this inequality yields

$$\int_W f_a(y) = \int_{W\cap X} f_a(y) + \int_{W\cap X^c} f_a(y)dy \geq \int_{X\cap W} f_a(y)dy + \int_{X\cap W^c} f(y)dy = \int_X f_a(y)dy$$

or

$$\mathbb{P}(Y \in W | Y \sim F_a) \geq \mathbb{P}(Y \in X | Y \sim F_a).$$

8.2.3 Two examples

Example 1: $Y_i \sim iid\ \mathcal{N}(\mu, 1)$, for $i = 1, \ldots, n$.

$$H_o : \mu = \mu_0$$

$$H_a : \mu = \mu_a$$

with $\mu_0 \neq \mu_a$. Note

$$f(y|\mu) = (2\pi)^{-\frac{n}{2}} \exp[-\frac{1}{2}\sum_{i=1}^{n}(y_i - \mu)^2]$$

and thus

$$\begin{aligned} lr(Y) &= ln(LR(Y)) \\ &= \frac{1}{2}\left(\sum (Y_i - \mu_0)^2 - \sum (Y_i - \mu_a)^2\right) \\ &= a(\mu_0, \mu_a) + \sum Y_i(\mu_a - \mu_0) \end{aligned}$$

Evidently, when $\mu_a > \mu_0$, the LR test rejects for large values of $\sum Y_i$, or equivalently large values of $\overline{Y} = n^{-1}\sum Y_i$.

This means that we can write the LR testing procedure as

Reject H_0 when $\overline{Y} >$ cv where cv is chosen so that

$$\mathbb{P}(\overline{Y} > \text{cv} \,|\overline{Y} \sim N(\mu_0, \frac{1}{n}) = \alpha$$

where the notation makes clear that the probability is computed under the assumption that the sample was drawn from the null distribution.

Notice that the critical region is the same for any H_a with $\mu_a > \mu_0$. That is, we use the same critical region for

$H_o : \mu = \mu_0$

$H_a : \mu > \mu_o$

Since the LR critical regions are the same for all of the simple hypotheses making up H_a and each is most powerful, then the LR procedure is said to be ***Uniformly Most Powerful*** **(UMP)** for H_o vs. H_a in this instance.

As a general matter, UMP tests don't exist. That is, there is no *single* test (critical region) that maximizes power for all values of the parameter under the alternative. What can be done in this case? One approach is to use a weighting function to capture the tradeoff between the various values under the alternation and then to construct a test that maximizes weighted average power.

8.3 Maximizing Weighted Average Power

Consider the *simple* null and *composite* alternative hypotheses:

$$H_o : \theta = \theta_o \text{ and } H_a : \theta \in \Theta_a.$$

Where *simple* means the hypothesis includes only one pdf (that is, one value of θ) and *composite* means the hypothesis contains more than one pdf (multiple values of θ).

Suppose you want to construct a test that maximizes *weighted average power* using the weight function $w(\theta)$ for values of $\theta \in \Theta_a$. Write the density of y, conditional on a particular value of θ as $f(y|\theta)$. For critical region W, the power of the test for a particular θ is $\int_W f(y|\theta)dy$, so that weighted average power is

$$\text{WAP} = \int_{\Theta_a} \left[\int_W f(y|\theta)dy\right] w(\theta)d\theta.$$

Interchanging the order of integration yields

$$\text{WAP} = \int_W \left[\int_{\Theta_a} f(y|\theta)w(\theta)d\theta\right] dy = \int_W g(y)dy, \text{ where } g(y) = \int_{\Theta_a} f(y|\theta)w(\theta)d\theta.$$

Notice that $g(y)$ is the density of Y under the assumption that θ is a random variable with density $w(\theta)$ and $f(y|\theta)$ is the density of Y conditional on θ. Thus, the problem is equivalent to the testing problem with a simple alternative:

$$\tilde{H}_a : y \sim g(y).$$

The best test is given by the Neyman-Pearson test, that is the null is rejected for large values of

$$LR(Y) = \frac{g(Y)}{f(Y|\theta_o)} = \frac{\int_{\Theta_a} f(Y|\theta)w(\theta)d\theta}{f(Y|\theta_o)}.$$

Example 1: $Y_i \sim iid N(\mu, 1)$, where Y is a scalar. We are interested in $H_o : \mu = \mu_0$ versus $H_a : \mu \neq \mu_o$. Without loss of generality, set $\mu_0 = 0$. Suppose we put weight of $\frac{1}{2}$ on each of $\mu = 1$ and $\mu = -1$. One shortcut to constructing the test is to note that $\overline{Y}$ is sufficient for μ, so we need only consider the scalar random variable $\overline{Y} \sim N(\mu, 1/n)$. A calculation shows that the WAP test rejects for large values of $\zeta = e^{n\overline{Y}} + e^{-n\overline{Y}} = e^{n|\overline{Y}|} + e^{-n|\overline{Y}|} = \zeta(|\overline{Y}|)$. You can verify that $\zeta(|\overline{Y}|)$ is increasing in the value of $|\overline{Y}|$. Thus, the test rejects for large values of $|\overline{Y}|$.

Example 2: $Y_i \sim iid\, \mathcal{N}_k(\mu, \Sigma)$ where Y is a $k \times 1$ vector. We are interested in $H_o : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$. Suppose we use a weight function with $\mu \sim \mathcal{N}(\mu_0, \omega^2\Sigma)$. In this case we can see that the distribution of Y under this weighted average alternative is

$$H_{a,weighted} : Y \sim \mathcal{N}(\mu_0, (1 + \omega^2)\Sigma).$$

Using sufficient statistics, the null and weighted-average alternative are:

$$H_0 : \overline{Y} \sim \mathcal{N}(\mu_0, n^{-1}\Sigma) \text{ versus } H_{a,weights} : \overline{Y} \sim \mathcal{N}(\mu_0, (1 + n\omega^2)n^{-1}\Sigma)$$

The WAP test is then the LR, which is

$$\zeta = e^{0.5\left(n(\overline{Y}-\mu_0)'\Sigma^{-1}(\overline{Y}-\mu_0) - n(\overline{Y}-\mu_0)'\Sigma^{-1}(\overline{Y}-\mu_0)/(1+n\omega^2)\right)},$$

so the test rejects for large values of the exponent. Notice that the exponent can be written as

$$\frac{n\omega^2}{1 + n\omega^2} n(\overline{Y} - \mu_0)'\Sigma^{-1}(\overline{Y} - \mu_0)$$

So the test rejects for large values of

$$\xi = n(\overline{Y} - \mu_0)'\Sigma^{-1}(\overline{Y} - \mu_0)$$

which is the Wald statistic that we studied earlier as an *ad hoc* test.

8.4 Tests based on the maximized value of the likelihood ratio.

Another way to accommodate a composite alternative is to use the largest value of the LR statistic under all values of $\theta \in \Theta_a$. For testing

$$H_o : \theta = \theta_0 \text{ versus } H_a : \theta \neq \theta_o$$

this yields:

$$\max_{\theta \neq \theta_0} \left(\frac{f(Y_{1:n}|\theta)}{f(Y_{1:n}|\theta_0)} \right) = \frac{f(Y_{1:n}|\hat{\theta})}{f(Y_{1:n}|\theta_0)} = \zeta(\hat{\theta})$$

where $Y_{1:n}$ denotes the $n \times 1$ data vector, $\hat{\theta}$ is the MLE (and I have assumed that $\hat{\theta} \neq \theta_0$). The *Likelihood Ratio Statistic* is defined as

$$\xi_{LR} = 2(ln(\zeta(\hat{\theta}))$$

and the null hypothesis is rejected for large values of ξ_{LR}, that is for $\xi_{LR} >$ cv, where cv is a critical value that satisfies $\mathbb{P}_{H_o}(\xi_{LR} > \text{cv}) = \alpha$, where α is the size of the test. To find cv we need the distribution of ξ_{LR} under the null.

Let $L_n(\theta) = \ln(f(Y_{1:n}|\theta))$, so that $\ln(\zeta(\hat{\theta})) = L_n(\hat{\theta}) - L_n(\theta_0)$

Write

$$L_n(\theta_o) = L_n(\hat{\theta}) + (\theta_o - \hat{\theta})' \frac{\partial L_n(\hat{\theta})}{\partial \theta} + \frac{1}{2}(\theta_o - \hat{\theta})' \frac{\partial^2 L_n(\tilde{\theta})}{\partial \theta \partial \theta'} (\theta_o - \hat{\theta})$$

where $\tilde{\theta}$ is between θ_o and $\hat{\theta}$.

Since $\frac{\partial L_n(\hat{\theta})}{\partial \theta} = 0$, one sees that

$$\begin{aligned} \xi_{LR} &= -(\hat{\theta} - \theta_0)' \frac{\partial^2 L_n(\tilde{\theta})}{\partial \theta \partial \theta'} (\hat{\theta} - \theta_0) \\ &= [\sqrt{n}(\hat{\theta} - \theta_0)' \left[-\frac{1}{n} \frac{\partial^2 L_n(\tilde{\theta})}{\partial \theta \partial \theta'} \right] [\sqrt{n}(\hat{\theta} - \theta_0)] \end{aligned}$$

From our earlier results

$$\sqrt{n}(\hat{\theta} - \theta_0)] \Rightarrow N(0, \mathcal{I}(\theta_0)^{-1})$$

under H_o. And

$$-\frac{1}{n} \frac{\partial^2 L_n(\tilde{\theta})}{\partial \theta \partial \theta'} = \frac{1}{n} \sum \frac{\partial^2 \ln(f(Y_i, \tilde{\theta})}{\partial \theta \partial \theta'} \xrightarrow{p, H_o} \mathcal{I}(\theta_0)$$

Thus, under H_o

$$\xi_{LR} \Rightarrow \xi \sim \chi^2_k$$

This final result follows from noting that ξ_{LR} is asymptotically a quadratic form of a $N(0, \mathcal{I})$ variable around the inverse of its covariance matrix.

Thus, we see that ξ_{LR} is (essentially) the same as the Wald statistic ξ_W that we discussed earlier, using the MLE of θ. They differ only to the extent that they use different estimators of the covariance matrix.

More generally, as we discussed previoulsy, if $\sqrt{n}(\hat{\theta} - \theta) \Rightarrow \mathcal{N}(0, V)$

We could form a statistic:

$\xi = n(\hat{\theta} - \theta_0)'\hat{V}^{-1}(\hat{\theta} - \theta_0)$ where $\hat{V}$ is consistent for V. Then $\xi \Rightarrow \chi^2_k$ random variable under the null.

For MLEs we know that V is the information matrix. We could estimate it in a variety of ways. Here are a few

$$\hat{V} = \left[\frac{1}{n}\sum s_i(\hat{\theta})s_i(\hat{\theta})'\right]^{-1}$$

$$\hat{V} = \left[\frac{1}{n}\sum s_i(\theta_o)s_i(\theta_o)'\right]^{-1}$$

$$\hat{V} = \left[\frac{1}{n}\sum s_i(\tilde{\theta})s_i(\tilde{\theta})'\right]^{-1}$$

where $\tilde{\theta}$ is between θ and $\hat{\theta}$.

Score tests (sometimes called Lagrange multiplier tests) are based on $\frac{1}{\sqrt{n}}\sum_{i=1}^n s_i(\theta_0)$, as we will discuss in class.

8.5 Odds and ends: what is a p-value?

Let ξ denote a test statistic. Suppose a test rejects when $\xi >$ cv, where cv is the critical value. Suppose you collect some data and find that $\xi = \xi^{Observed}$, the observed value of your statistic. The p-value associated with this statistics is

$$p - value = \mathbb{P}_{H_o}(\xi(Y) > \xi^{Observed})$$

which is the probability of a value of the test statistic that is at least as large as the value you observed. Note that $p - value$ represents smallest size for which you would reject H_0.

(Exercise: Suppose that the null hypothesis is true. Show that the $p - value$ is distributed $\mathcal{U}(0, 1)$.)

9 Confidence Sets

A $(1 - \alpha) \times 100\%$ confidence set for θ is a (random) set of values of θ that contains θ_0, the true value of θ, with probability 1–α. Let $C(Y)$ denote such a set. That is, suppose that $\mathbb{P}[\theta_0 \in C(Y)] =$ 1–α. (Note that the randomness comes from $Y \sim F_Y(y|\theta_0)$, so that the set $C(Y)$ is random.)

9.1 Inverting test statistics

An easy way to construct such a confidence set is to "invert" hypothesis tests.

Here's the approach: Let Θ denote a set that contains the true value of θ. Consider carrying out hypothesis tests using every value of θ in Θ as a null hypothesis and use a size α test for each of these θ values. Let $C(Y)$ denote the set of values of θ for which the test does not reject the null.

Note that, since $\theta_0 \in \Theta$, the null $\theta = \theta_0$ was one of the tests constructed. This test rejected the null with probability α, hence did not reject with probability $1-\alpha$. Thus, with $C(Y)$ constructed in this way, $\mathbb{P}[\theta_0 \in C(Y)] = 1$–$\alpha$.

When the hypothesis test is carried out using a Wald-statistic with a limiting χ^2 distribution, the confidence set is particularly easy to construct. Thus, suppose $\hat{\theta} \overset{a}{\sim} \mathcal{N}\left(\theta, \frac{1}{n}V\right)$, and $\hat{V}$ is a consistent estimator for V. Then $H_o : \theta = \theta_0$ is not rejected using a test of size α if

$$\xi = (\widehat{\theta} - \theta_0)' \left[\frac{1}{n}\widehat{V}\right]^{-1} (\widehat{\theta} - \theta_0) \leq \chi^2_{\kappa,1-\alpha}$$

where $\chi^2_{k,1-\alpha}$ denotes the $1-\alpha$ quantile of the χ^2_k distribution. The confidence set is therefore

$$C(Y) = \left\{ \theta \middle| (\widehat{\theta} - \theta)' \left[\frac{1}{n}\widehat{V}\right]^{-1} (\widehat{\theta} - \theta) \leq \chi^2_{\kappa,1-\alpha} \right\}$$

which is recognized as the interior of an ellipse centered at $\theta = \hat{\theta}$

In the one dimensional case ($k = 1$), the normal distribution can be used in the place of the χ^2 yielding

$$C(Y) = \left\{ \theta | \widehat{\theta} - Z_{1-\alpha/2} \times \sqrt{\frac{1}{n}\widehat{V}} \leq \theta \leq \widehat{\theta} + Z_{1-\alpha/2} \times \sqrt{\frac{1}{n}\widehat{V}} \right\}$$

where $Z_{1-\alpha/2}$ denotes the $1-\alpha/2$ ordinate of the $N(0,1)$ distribution.

9.2 Efficient confidence sets

We have discussed constructing confidence sets by "inverting" test statistics, but this is not the only way to form such sets. Perhaps the easiest is to simply flip a coin for each value of θ and include that value if a "heads" appears, and otherwise exclude that value. Such a confidence set will contain the true value with probability 0.50 (the probability of a head appearing). Of course, this is a silly way to form a confidence set because it ignores the information in Y, yielding a confidence set that is "larger" than it needs to be.

Pratt (1961)[1] discusses efficient confidence sets and shows how these are related to most powerful (i.e., efficient) tests. Here is a version of his insight.

Let $C(Y)$ denote a confidence set for a parameter θ. The "volume" of $C(Y)$ can be expressed as

$$V_C(Y) = \int_\Theta \mathbf{1}[\theta \in C(Y)] d\theta$$

where $\mathbf{1}[]$ is the indicator function. ($\mathbf{1}[x] = 1$ if x is 'true' and $\mathbf{1}[x] = 0$ if x is 'false'.)

Because the set $C(Y)$ depends on Y, the set is random, and so is its volume. Suppose $Y \sim f$. Then the expected volume is:

$$R_C = \mathbb{E}[V_C(Y)] = \int V_C(y) f(y) dy.$$

which serves as a criterion for evaluating confidence sets: $C_1(Y)$ is preferred to $C_2(Y)$ if $R_{C_1} < R_{C_2}$.

Now, consider the testing problem: $H_o : Y \sim f_\theta$ versus $H_a : Y \sim f$. If we have a $1-\alpha$ confidence set $C(Y)$, an α-level test can be constructed as: "accept H_o if $\theta \in C(Y)$, and otherwise reject H_o." The probability of Type 2 error (i.e., 1-Power) for this test is

$$\mathbb{P}(\text{accept } H_0 | H_a \text{ is true}) = \mathbb{P}[(\theta \in C(Y) | Y \sim f] = \int \mathbf{1}[\theta \in C(y)] f(y) dy.$$

Now, rewrite the expression for expected volume:

$$\begin{aligned} R_C &= \mathbb{E}[V_C(Y)] \\ &= \int V_C(y) f(y) dy \\ &= \int \left[\int_\Theta \mathbf{1}(\theta \in C(y) d\theta \right] f(y) dy \\ &= \int_\Theta \left[\int \mathbf{1}(\theta \in C(y) f(y) dy \right] d\theta \end{aligned}$$

Thus, R_C can be minimized by making the final term in [ ]'s as small as possible for each value of θ. But, as shown above, this term is the probability of Type 2 error (i.e., 1-Power) for $H_0 : Y \sim f_\theta$. So $\left[\int \mathbf{1}(\theta \in C(y) f(y) dy\right]$ can be minimized by choosing the most power test for $H_0 : Y \sim f_\theta$. This is achieved using a Neyman-Pearson test.

Example: Above we considered the testing problem with $Y_i \sim i.i.d. \mathcal{N}_k(\mu, \Sigma)$ and

$$H_0 : \mu = \mu_o \text{ versus } H_a : \mu \sim \mathcal{N}(0, \omega^2 \Sigma).$$

[1] Pratt, John W. (1961), "Length of Confidence Intervals," *Journal of the American Statistical Association*, 56 (295), pp. 549-567.

We showed that the optimal test was the Wald test

$$\xi = (\overline{Y} - \mu_o)'(n^{-1}\Sigma)^{-1}(\overline{Y} - \mu_o).$$

Pratt's results show that a confidence interval formed by inverting this test will have the smallest expected volume, with the expectation computed using $Y|\mu \sim \mathcal{N}(\mu, \Sigma)$ with $\mu \sim \mathcal{N}(0, \omega^2\Sigma)$.

10 The Bayes Approach to Estimation and Inference

10.1 Basic concepts and some jargon

We have been concerned with learning about a parameter θ from data Y where the *pdf* of Y depends on θ, that is $Y \sim f(.|\theta)$. We have studied the behavior of procedures (e.g., estimators, tests, etc.) over possible random draws of Y from this *pdf*. Estimators with small expected loss over this *pdf* are good, as are tests that have a low probability of type 1 and type 2 error. This approach to inference is called *frequentist* (or *classical).* It treats θ as a constant and the data (Y) as a random draw from $f(.|\theta)$.

An alternative framework, called *Bayes*, treats both Y and θ as random. The data come $f(Y|\theta)$, the probability distribution of Y for a particular θ, and the goal is to learn about θ from the Y you actually observe, say $Y = y$. This is the same goal as frequentist inference, but with θ random, we can use the rules of probability to deduce that the information about θ contained in $Y$is summarized in the conditional distribution $f(\theta|Y)$. Bayes analysis focuses on computing this conditional distribution.

Bayes inference uses some new jargon.

- The marginal pdf for θ, say $f(\theta)$, is called the *prior.*
- The conditional pdf of Y given θ, $f(Y|\theta)$, is called the *likelihood.* This is same definition of the likelihood used in above from frequentist inference.
- The conditional pdf of θ given Y, $f(\theta|Y)$, is called the *posterior.*
- The marginal pdf of Y, say $f(Y)$, is called the *marginal likelihood.*

Being a little clearer with notation:

- $f_\theta(\tilde{\theta})$ is the pdf for θ evaluated at $\tilde{\theta}$. (the prior)
- $f_{Y|\theta}(y|\tilde{\theta})$ is the pdf for Y conditional on $\theta = \tilde{\theta}$ evaluated at y. (the likelihood)
- $f_Y(y)$ is the pdf for Y evaluated at y. (the marginal likelihood)
- $f_{\theta|Y}(\tilde{\theta}|y)$ is the the pdf for θ conditional on $Y = y$ evaluated at $\widetilde{\theta}$. (the posterior)

- $f_{Y,\theta}(y,\tilde{\theta})$ is the joint pdf for (Y,θ) evaluated at $(y,\tilde{\theta})$.

The relationship between these densities follows directly from what we know about marginal, joint and conditional pdfs:

1. The joint pdf for (Y,θ) can be written as

$$f_{Y,\theta}(y,\tilde{\theta}) = f_{Y|\theta}(y|\tilde{\theta})f_{\theta}(\tilde{\theta})$$

2. The marginal density of Y is the joint density, integrated with respect to θ. That is

$$f_Y(y) = \int f_{Y,\theta}(y,\tilde{\theta})d\tilde{\theta} = \int f_{Y|\theta}(y|\tilde{\theta})f_{\theta}(\tilde{\theta})d\tilde{\theta}.$$

3. The conditional density of θ given Y=y and evaluated at $\theta = \tilde{\theta}$ is given by

$$f_{\theta|Y}(\tilde{\theta}|y) = \frac{f_{Y,\theta}(y,\tilde{\theta})}{f_Y(y)} = \frac{f_{Y|\theta}(y|\tilde{\theta})f_{\theta}(\tilde{\theta})}{\int f_{Y,\theta}(y,\tilde{\theta})d\tilde{\theta}} = \frac{f_{Y|\theta}(y|\tilde{\theta})f_{\theta}(\tilde{\theta})}{\int f_{Y|\theta}(y|\tilde{\theta})f_{\theta}(\tilde{\theta})d\tilde{\theta}}\,.$$

4. Thus

$$Posterior = \frac{Likelihood \times Prior}{Marginal\ Likelihood}$$

5. And noting that the Marginal likelihood does not depend on θ:

$$Posterior(\theta) \propto (Likelihood(\theta) \times Prior(\theta))$$

10.2 Examples

Example 1:

$Y|\mu \sim \mathcal{N}(\mu,1)$. Y is a scalar. Prior $\mu = 2$ w.p. $1/3$ and $\mu = 4$ w.p. $2/3$. You observe $Y = 2.7$. Derive posterior for μ.

First, note that the posterior is (Likelihood×prior)/(Marginal likelihood). Thus, if prior = 0, then so will be the posterior. Thus, the posterior will only have mass at μ =2 and $\mu = 4$.

$$\mathbb{P}(\mu = 2|Y = 2.7) = \frac{f_{Y|\mu}(2.7|\mu = 2)\mathbb{P}(\mu = 2)}{f_{Y|\mu}(2.7|\mu = 2)\mathbb{P}(\mu = 2) + f_{Y|\mu}(2.7|\mu = 4)\mathbb{P}(\mu = 4)}$$

where $f_{Y|\mu}(2.7|\mu = 2) = \frac{1}{\sqrt{2\pi}}e^{-(1/2)\times(2.7-2)^2}$, and similarly for $\mu = 4$.

Plugging in the numbers we find:

$\mathbb{P}(\mu = 2|Y = 2.7) = \frac{e^{-(1/2)(2.7-2)^2}(1/3)}{e^{-(1/2)(2.7-2)^2}(1/3)+e^{-(1/2)(2.7-4)^2}(2/3)} \approx 0.65$

and because μ can take on only two values, 2 and 4, then

$\mathbb{P}(\mu = 4|Y = 2.7) \approx 1 - 0.65 = 0.35.$

Example 2:

Now suppose the prior is $\mu \sim \mathcal{N}(1, 4)$. We carry out the same calculations

$$f_{\mu|Y}(\mu = m|Y = 2.7) = \frac{f_{Y|\mu}(2.7|\mu = m)f_{\mu}(m)}{\int f_{Y|\mu}(2.7|\mu = u)f_{\mu}(u)du}$$

We can solve this directly. Alternatively, from our work on the multivariate normal we know that if $f_{Y|\mu}(y|\mu = m) \sim N(m, 1)$ and $f\mu(m) \sim \mathcal{N}(1, 4)$, then

$$\begin{bmatrix} Y \\ \mu \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 5 & 4 \\ 4 & 4 \end{bmatrix}\right)$$

and $f_{\mu|Y}(m|Y = 2.7) \sim \mathcal{N}(1 + (4/5)(2.7 - 1), 4 - 4^2/5)$.

Example 3:

Suppose $Y_i|\mu$ are iid $\mathcal{N}(\mu, 1)$ and $\mu \sim \mathcal{N}(\tau, \omega^2)$, where τ and ω are known constants. Let $Y = (Y_1, \ldots, Y_n)$'. $f_{\mu}(m)$ is therefore the $\mathcal{N}(\tau, \omega^2)$ density and $f_{Y|\mu=m}$ is the $\mathcal{N}(lm, I_n)$ density where l is an $n \times 1$ vector of 1s. The normal-normal densities imply that $(Y' \ \mu)'$ has a joint normal density, and you can verify that

$$\begin{bmatrix} Y \\ \mu \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} l\tau \\ \tau \end{bmatrix}, \begin{bmatrix} \omega^2 ll' + I_n & \omega^2 l \\ \omega^2 l' & \omega^2 \end{bmatrix}\right)$$

so that $\mu|Y = y \ \sim \mathcal{N}\left(\tau + \omega^2 l'(\omega^2 ll' + I_n)^{-1}(y - l\tau), \omega^2 - \omega^2 l'(\omega^2 ll' + I_n)^{-1}\omega^2 l\right)$.

Note that $(\omega^2 ll' + I_n)^{-1} = (I_n - \kappa ll')$ where $\kappa = n\omega^2/(n + \omega^2)$. Plugging this in and simplifying yields:

$\mu|Y = y \sim \mathcal{N}(\lambda\tau + (1 - \lambda)\overline{y}, (1 - \lambda)^2/n)$, where $\lambda = 1/(1 + n\omega^2)$ and $\overline{y} = n^{-1}\sum_{i=1}^{n} y_i$.

This is the posterior for μ.

There is a simplification that can be exploited here. Recall that $\overline{Y}$ is sufficient for μ. Thus, for the purpose of conducting inference about μ, the data are completely summarized by $\overline{Y}$. Note $\overline{Y}|\mu = m \sim N(m, 1/n)$. Following the same steps as above

$$\begin{bmatrix} \overline{Y} \\ \mu \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \tau \\ \tau \end{bmatrix}, \begin{bmatrix} \omega^2 + 1/n & \omega^2 \\ \omega^2 & \omega^2 \end{bmatrix}\right)$$

so that $\mu|\overline{Y} = \overline{y}$ is normally distributed with mean $\tau + \left[\frac{\omega^2}{\omega^2+1/n}\right](\overline{y}-\tau)$ and variance $\omega^2 - \frac{\omega^4}{\omega^2+1/n}$. Rearranging these expressions yields the same expression that were derived above.

Jargon: Note that the posterior has the same form as the prior − it is normal − but with different parameter. When the posterior and prior have the same form, the prior is said the be ***conjugate***.

10.3 Bayes Estimators

Armed with the posterior, constructing Bayes estimators is straightforward. Suppose loss is denoted by $L(\hat{\theta}, \theta)$. You observe $Y = y$. The problem is then

$$\min_{\hat{\theta}} \mathbb{E}_{\theta|Y=y}\left[L(\hat{\theta}, \theta)\right],$$

that is, $\hat{\theta}$ is chosen to minimize expected loss, where the loss is averaged over the values of θ using the knowledge that $Y = y$. (Jargon: $\mathbb{E}_{\theta|Y=y}\left[L(\hat{\theta}, \theta)\right]$ is called *posterior risk*.)

When loss is quadratic, i.e., $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, then risk is *mean squared error*, and we know (from the results in Section 2.4) that the *mse*-minimizing value of $\hat{\theta}$ is

$$\hat{\theta} = \mathbb{E}_{\theta|Y=y}(\theta),$$

that is, the Bayes estimator is the posterior mean.

Example: In example 3 in the last section, the posterior was

$$\mu|Y = y \sim \mathcal{N}(\lambda\tau + (1-\lambda)\overline{y}, (1-\lambda)^2/n)$$

where $\lambda = 1/(1 + n\omega^2)$. Thus, the Bayes estimator for μ is

$$\hat{\mu}^{Bayes} = \lambda\tau + (1-\lambda)\bar{y}.$$

10.4 Bayes Credible Sets

Bayes credible sets are the analogue of frequentist confidence sets. Specifically, a $100 \times (1-\alpha)\%$ Bayes credible set, say $C(y)$, is a set that satisfies

$$\mathbb{P}_{\theta|Y=y}(\theta \in C(y)) = 1 - \alpha$$

where the notation emphasizes that the probability is computed using the posterior for θ given $Y = y$.

Because this holds for all y, we also have

$$\mathbb{P}_{\theta,Y}(\theta \in C(Y)) = 1 - \alpha$$

where now the probability is computed over both θ and Y.

Bayes credible sets are easily computed from the posterior. Examples will be discussed in class.

11 More on Bayes methods

As time allows we will cover more topics in Bayes methods.

11.1 More on Risk and Bayes and Frequentist estimators

Let $L(\hat{\theta}, \theta)$ denote loss. Then, there are various versions of *Risk*.

- *Frequentist Risk* uses $\hat{\theta} = \hat{\theta}(Y)$ and is

$$R(\hat{\theta}, \theta) = \mathbb{E}_{Y|\theta}\left[L(\hat{\theta}(Y), \theta)\right].$$

  That is, $R(\hat{\theta}, \theta)$ holds θ fixed and computes the risk over the possible values of Y that might be drawn.

  - ***Admissability:*** $\hat{\theta}$ is said to be ***inadmissable*** if there exists another estimator, say $\tilde{\theta}$, such that $R(\tilde{\theta}, \theta) \leq R(\hat{\theta}, \theta)$ for all θ, and where the inequality is strict for some θ. Thus, inadmissable estimators are dominated.

- *Posterior Risk* conditions on $Y = y$ (the sample values of Y) and averages the values of θ given $Y = y$ yielding the exected loss

$$\mathbb{E}_{\theta|Y=y}\left[L(\hat{\theta}(y), \theta)\right]$$

  which is called posterior risk.

- *Bayes Risk* averages over both Y and θ using their joint pdf:

$$r(\hat{\theta}) = \mathbb{E}_{Y,\theta}[L(\hat{\theta}(Y), \theta)].$$

  - Notice that (from the law of iterated expectations)

$$r(\hat{\theta}) = \mathbb{E}_{Y,\theta}[L(\hat{\theta}(Y), \theta)] = \mathbb{E}_{Y}[\mathbb{E}_{\theta|Y}[L(\hat{\theta}(Y), \theta)]]$$

    so that $r(\hat{\theta})$ averages posterior risk over all possible values of Y.
  - Because Bayes estimators minimize $\mathbb{E}_{\theta|Y=y}[L(\hat{\theta}(y), \theta)]$ for each value of y, they also minimize the average value of $\mathbb{E}_{\theta|Y}[L(\hat{\theta}(Y), \theta)]$ over Y. Thus Bayes estimators minimize $\mathbb{E}_{Y}[\mathbb{E}_{\theta|Y}[L(\hat{\theta}(Y), \theta)]] = r(\hat{\theta})$.

- Note, you can also write

$$r(\hat{\theta}) = \mathbb{E}_{Y,\theta}[L(\hat{\theta}(Y), \theta)] = \mathbb{E}_{\theta}[\mathbb{E}_{Y|\theta}[L(\hat{\theta}(Y, \theta)]] = \mathbb{E}_{\theta}[R(\hat{\theta}, \theta)]$$

  so that Bayes risk averages frequentist risk over the values of θ using the prior for θ.

– Because Bayes estimators minimize $r(\hat{\theta})$ they must be admissable. (Exercise: Show this.)

11.1.1 Some properties of Bayes Estimators (discussed as time allows)

1. In general, Bayes estimators are biased, conditional on the value of μ.
   (a) Recall in example 3, the Bayes estimator was $\hat{\mu}^{Bayes} = \lambda\tau + (1-\lambda)\overline{Y}$. Thus $\mathbb{E}_{Y|\mu=m}(\hat{\mu}^{Bayes}) = \lambda\tau + (1-\lambda)m = m + \lambda(\tau - m) \neq m$.
2. Bayes estimators are admissable (see the last subsection)
3. Bayes estimators are normally distributed in large samples, centered at the MLE. The general result that the posterior is approximately normal, centered at the MLE with variance given by the inverse of the information is called the "Bernstein – von Mises theorem," which says (loosely) that (under a set of regularity conditions) that the posterior distribution of θ is well approximated in large samples by the $N(\hat{\theta}^{MLE}, n^{-1}I(\theta_0)^{-1})$, where $I(\theta_0)$ is the information.
   (a) You can see this in the example:

$$\sqrt{n}(\hat{\mu}^{Bayes} - \overline{Y}) \quad = \quad \sqrt{n}\lambda(\tau - \overline{Y}) \xrightarrow{p} 0$$

   recognizing that $\sqrt{n}\lambda = \frac{\sqrt{n}}{1+\omega^2 n} \to 0$.
   (b) The more general argument is more involved. I sketch in here:

Let $Y_{1:n}$ denote the sample of size n of *i.i.d.* observations. The likelihood is $f(Y_{1:n}|\theta)$, and the log-likelihood is $L_n(\theta) = \ln(f(Y_{1:n}|\theta))$. Use $w(\theta)$ to denote the prior density for θ.

The posterior for $\theta|Y_{1:n}$ is $f(\theta|Y_{1:n}) = \frac{f(Y_{1:n}|\theta)w(\theta)}{\int f(Y_{1:n}|\theta)w(\theta)d\theta}$.

Let $\gamma = \sqrt{n}(\theta - \hat{\theta}^{MLE})$, so that $\theta = \hat{\theta}^{MLE} + \gamma/\sqrt{n}$. Then, from the change of variables formula:

$$\begin{aligned} f(\gamma|Y_{1:n}) &= \frac{f(Y_{1:n}|\hat{\theta}^{MLE} + \gamma/\sqrt{n})w(\hat{\theta}^{MLE} + \gamma/\sqrt{n})}{\int f(Y_{1:n}|\hat{\theta}^{MLE} + \gamma/\sqrt{n})w(\hat{\theta}^{MLE} + \gamma/\sqrt{n})d\gamma} \frac{n^{-1/2}}{n^{-1/2}} \\ &= \frac{[f(Y_{1:n}|\hat{\theta}^{MLE} + \gamma/\sqrt{n})/f(Y_{1:n}|\hat{\theta}^{MLE})]w(\hat{\theta}^{MLE} + \gamma/\sqrt{n})}{\int [f(Y_{1:n}|\hat{\theta}^{MLE} + \gamma/\sqrt{n})/f(Y_{1:n}|\hat{\theta}^{MLE})]w(\hat{\theta}^{MLE} + \gamma/\sqrt{n})d\gamma} \\ &= \frac{e^{L_n(\hat{\theta}^{MLE}+\gamma/\sqrt{n}) - L_n(\hat{\theta}^{MLE})}w(\hat{\theta}^{MLE} + \gamma/\sqrt{n})}{\int e^{L_n(\hat{\theta}^{MLE}+\gamma/\sqrt{n}) - L_n(\hat{\theta}^{MLE})}w(\hat{\theta}^{MLE} + \gamma/\sqrt{n})d\gamma} \end{aligned}$$

Now as n grows large $w(\hat{\theta}^{MLE} + \gamma/\sqrt{n}) \xrightarrow{p} w(\theta_0)$ and

$L_n(\hat{\theta}^{MLE} + \gamma/\sqrt{n}) - L_n(\hat{\theta}^{MLE}) = \frac{1}{2}\frac{\partial^2 L_n(\tilde{\theta})}{\partial\theta^2}\frac{\gamma^2}{n}$, where $\tilde{\theta}$ is between $\hat{\theta}^{MLE}$ and $\hat{\theta}^{MLE} + \gamma/\sqrt{n}$.

Thus $L_n(\hat{\theta}^{MLE} + \gamma/\sqrt{n}) - L_n(\hat{\theta}^{MLE}) \xrightarrow{p} -\frac{1}{2}\mathcal{I}(\theta_0)\gamma^2$.

Using these approximations

$$\begin{aligned} f(\gamma|Y_{1:n}) &= \frac{e^{-0.5\mathcal{I}(\theta_0)\gamma^2}}{\int e^{-0.5\mathcal{I}(\theta_0)\gamma^2}d\gamma}\frac{w(\theta_0)}{w(\theta_0)} \\ &= \frac{1/\sqrt{2\pi\mathcal{I}(\theta_0)^{-1}}}{1/\sqrt{2\pi\mathcal{I}(\theta_0)^{-1}}}\frac{e^{-0.5\mathcal{I}(\theta_0)\gamma^2}}{\int e^{-0.5\mathcal{I}(\theta_0)\gamma^2}d\gamma} \\ &= 1/\sqrt{2\pi\mathcal{I}(\theta_0)^{-1}}e^{-0.5\mathcal{I}(\theta_0)\gamma^2} \end{aligned}$$

while the final equality follows by noting that the denominator of the preceding expression is 1. (It the pdf of normal integrated over all value .. so it integrates to 1.)

The results then follows by noting that $1/\sqrt{2\pi\mathcal{I}(\theta_0)^{-1}}e^{-0.5\mathcal{I}(\theta_0)\gamma^2}$ is the normal density with $mean = 0$ and $variance = \mathcal{I}(\theta_0)^{-1}$. That is, for large n, $\gamma|Y_{1:n} \sim \mathcal{N}(0, \mathcal{I}(\theta_0)^{-1})$. Because $\theta = \hat{\theta}^{MLE} + \gamma/\sqrt{n}$ the posterior of θ is approximately $\mathcal{N}(\hat{\theta}^{MLE}, n^{-1}\mathcal{I}(\theta_0)^{-1})$.