Studienzentrum Gerzensee Doctoral Program in Economics
Econometrics 2017-18, Week 1 Lecture Notes

Updated September 11, 2017
(typos corrected, CLT discussion extended pages 46-47)

Administrative Details

Grades
Exams (Midterm and Final. Closed book. 1 page of notes)
Exercises (Randomly selected student, chosen in advance)
Graded Problem Set in 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 – Time Series (Watson) and Cross Section (Honore) Topics

Readings for Week 1

Hogg, R.V, J.W. McKean, and A.T. Craig, *Introduction to Mathematical Statistics*, 6th Edition, Prentice Hall, 2005. HMC (or earlier version: Hogg, R.V and A.T. Craig, A *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 … choose the one(s) you like. Here's one:

Casella, G. and R.L. Berger (2008), *Statistical Inference*, 2nd Edition, Thompson Press.

Probability Concepts

Experiment and Outcomes

Uncertain/Not Perfectly Predictable/Random/Stochastic (example: roll of die)

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

Note Ω may contain an infinite number of possible outcomes
Countable (Experiments like flipping a coin until "Tails" appears)
Uncountable (growing a tomato and measuring its weight)

Events

Subset of Ω (die example {(3),(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 intersection of A and B)
A^c (the compliment of A)

σ-Algebra (or σ-field): Let **A** be a collection of subsets of Ω that satisfies

(1) $\Omega \in \mathbf{A}$
(2) $A \in \mathbf{A}$ then $A^c \in \mathbf{A}$
(3) if $A_1, A_2, \ldots \in \mathbf{A}$ then $(\cup_{i=1}^{\infty} A_i) \in \mathbf{A}$

Notes: Sometimes (1) is replaced with $\emptyset \in \mathbf{A}$.
Because $(\cup_{i=1}^{\infty} A_i^c)^c = \cap_{i=1}^{\infty} A_i$ then (2) and (3) imply that $\cap_{i=1}^{\infty} A_i \in \mathbf{A}$.

Probability measure: A real-valued set function (maps sets into the real line) with the properties

(1) $A \in \mathbf{A}$ then $P(A) \geq 0$
(2) $P(\Omega) = 1$
(3) If $\{A_i\}_{i=1}^{\infty}$ is a countable set of disjoint sets in **A**, then
$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Probability Space: The triple (Ω, **A**, P) defines a probability space.

There are many useful facts (see Hogg and Craig, Section 1.3 theorems 1-5), including:

1. $0 \leq P(A) \leq 1$

2. $P(A) = 1 - P(A^c)$

3. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

4. If $\{A_i\}_{i=1}^{\infty}$ are a set of mutually exclusive and exhaustive subsets of Ω, then $P(A) = \sum_{i=1}^{\infty} P(A \cap A_i)$.

Conditional Probability

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

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

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

Note: $P(A \mid B)P(B) = P(A \cap B) = P(B \mid A)P(A)$.

Being a bit more careful:

Formally, the idea is to construct a new probability space from (Ω, **A**, P) by assigning zero probability to all elementary events that are not in B. The new probability measure, say 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

$$\frac{P_o(\omega_A)}{P_o(\omega_B)} = \frac{P(\omega_A)}{P(\omega_B)}$$

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

Independence

Events A and B are independent if $P(A|B) = P(A)$.

Independence implies $P(A \cap B) = P(A)P(B)$ so that $P(B \mid A) = P(B)$.

Bayes Rule

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

We know

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

To compute the pieces, first note that $B = (B \cap A) \cup (B \cap A^c)$ where $(B \cap A)$ and $(B \cap A^c)$ are two disjoint sets.

Thus,

$$P(B) = P(B \cap A) + P(B \cap A^c),$$

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

and

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

so that

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

which is known as Bayes Rule.

Random Variables

Consider a probability space (Ω, **A**, P).

(Example: Ω denotes outcomes of 3 tosses 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 tosses; (iii) $X(\omega) = 1$ if heads appears on the 1^{st} and 3^{rd} toss 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 **A**. This makes the probability space for X, say (Ω_X, $\mathbf{A}_X$, P_X), consistent with the original probability space (Ω, **A**, P). Such a restriction makes X *measurable* with respect to **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, $\{\omega \mid X(\omega) = x) = A(x) \in \mathbf{A}$.

Letting $A_X \in \mathbf{A}_X$, the resulting probability function, P_X, is given by $P_X(A_X) = P(\omega \mid \omega \in \Omega, X(\omega) \in A_X)$.

Example:

A fair coin is tossed 3 times.

Ω = [(*HHH*), (*THH*), (*HTH*), (*HHT*), (*HTT*), (*THT*), (*TTH*), (*TTT*)]

$X(\omega)$ = number of heads,

Ω_X = [0, 1, 2, 3],

A denotes all subset of Ω and $\mathbf{A}_X$ denotes all subsets of Ω_X

$P_X(1) = P$[(*HTT*), (*THT*), (*TTH*)] = 3/8, etc.

Distribution Functions

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

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

which is often denoted $P(X \leq x)$ (with a slight abuse of notation).

Some Properties of the CDF.

1. For $x_2 \geq x_1$, $F_X(x_2) - F_X(x_1) = P(x_1 < X \leq x_2)$,

2. $F_X(-\infty) = 0$

3. $F_X(\infty) = 1$

4. $F_X(\cdot)$ is non-decreasing

***Probability Density Function*:** Suppose that X is a *discrete random variable* and can take on only a finite number of values $x_1, x_2, \ldots, x_n$. We can then define

$$P(X = x_i) \overset{def}{=} p_i \overset{def}{=} f_X(x_i)$$

as the *density function* for X and the resulting CDF is a step function.

For a discrete random variable $F_X(x) = \sum_{x_i \leq x} f_X(x_i)$.

For a continuous random variable, define the pdf analogously: $f_X(\,.\,)$ satisfies

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\, ds$$

so that

$$f_X(x) \overset{def}{=} \frac{dF(x)}{dx} \quad .$$

Note:

$$P[x_1 \leq X \leq x_2] = F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(x) dx$$

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 is a randomly chosen point between 0 and 1 with

$$F_X(x) = \left\{ \begin{array}{c} x, \text{ for } 0 \le x \le 1 \\ 0 \text{ for } x < 0 \\ 1 \text{ for } x > 1 \end{array} \right\}$$

Then X is said to be *Uniformly Distributed* on $[0,1]$. The probability density function of X is

$$f_X(x) = \left\{ \begin{array}{c} 1 \text{ for } 0 < x < 1 \\ 0 \text{ elsewhere} \end{array} \right.$$

Bivariate Distribution Functions

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

$$F_{X,Y}(x,y) \overset{def}{=} P((X \le x) \text{ and } (Y \le y))$$

which is often denoted $P(X \le x, Y \le y)$.

For a discrete random variable $F_{X,Y}(x,y) = \sum_{y_i \le y} \sum_{x_i \le x} f_{X,Y}(x_i, y_i)$

For a continuous random variable $F_{X,Y}(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(z_1, z_2) dz_1 dz_2$

$F_X(x) = P(X \le x) = P\left[(X \le x) \cap (Y \le \infty)\right] = F_{X,Y}(x, \infty)$ is the CDF of X and thus

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$$

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 $f_{X,Y}(x,y) = f_X(x) f_Y(y)$.)

Conditional Distribution Functions

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

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

for values of x with $P(X = x) > 0$.

Equivalently

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

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 often write this as $f_{Y|X}(y \mid x)$.

The conditional CDF is $P(Y \le y \mid X = x) = \int_{-\infty}^{y} f_{Y|X}(s \mid x)ds$

which I will write as $F_{Y|X}(y \mid X = x)$ or $F_{Y|X}(y \mid x)$.

If X and Y are independent, then $F_{Y|X}(y \mid X = x) = F_Y(y)$ and $f_{Y|X}(y \mid x) = f_Y(y)$ for all x and y.

Multivariate Distribution Functions

CDFs for the vector of random variables $X = (X_1, X_2, \ldots, X_n)$ are defined analogously to the bivariate case. Conditional distributions are defined similarly, and so forth.

Expectations

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

$$\mathbf{E}g(X) \overset{def}{=} \sum_i g(x_i) f_X(x_i)$$

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

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

A few useful facts. Suppose a and b are constants, and $g_1(x)$ and $g_2(x)$ are two functions:

$$\mathbf{E}a = \int_{-\infty}^{\infty} a f_X(x) dx = a \int_{-\infty}^{\infty} f_X(x) dx = a$$

$$\mathbf{E}[ag(X)] = \int_{-\infty}^{\infty} ag(x) f_X(x) dx = a \int_{-\infty}^{\infty} g(x) f_X(x) dx = a\mathbf{E}g(X)$$

$$\mathbf{E}[g_1(X) + g_2(X)] = \mathbf{E}g_1(X) + \mathbf{E}g_2(X)$$

(Said differently, **E** is a linear operator.)

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

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

and

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

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 $\mathbf{E}g(X,Y) = \int\int g(x,y) f_{X,Y}(x,y) dxdy$

Suppose G is a matrix of random variables, then $\mathbf{E}(G)$ is a matrix with $ij'th$ element equal to $\mathbf{E}(G_{ij})$.

Suppose $g(X,Y) = a(X)b(Y)$ and X and Y are independent.
Exercise: Show $\mathbf{E}g(X,Y) = \mathbf{E}a(X) \times \mathbf{E}b(Y)$.

Conditional Expectations

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

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

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

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

The Law of Iterated Expectations: $\mathbf{E}_Y(Y) = \mathbf{E}_X[\mathbf{E}_{Y|X}(Y|X)]$

Proof:

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

$$= \int \int y f_{Y|X}(y \mid x) f_X(x) dy dx = \int \int y f_{Y,X}(y, x) dy dx$$

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

Optimal Forecasting:

Problem 1: Find the constant h that minimizes "Mean Squared Error", $E[(Y - h)^2]$.

$E[(Y - h)^2] = \int (y-h)^2 f_Y(y)\,dy$ yielding 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 h.

Problem 2: Find the function $h(x)$ that minimizes $E_{Y|X=z}[(Y - h(x))^2]$.
Same problem as 1, but using the conditional distribution of $Y|X$=x. Thus, the optimal $h(.)$ is $h(x) = E(Y \mid X = x) = \mu_Y(x)$.

"The minimum mean square error forecast is given by regression function."

Here's an alternative derivation, yielding the result by inspection.

Optimal Forecasting: You observe $X=x$, and you want to forecast Y.

Suppose $h(x) = \mu_Y(x) + g(x)$.

Then:

$$(Y-h(x))^2 = (Y-\mu_Y(x))^2 - 2g(x)(Y-\mu_Y(x)) + g(x)^2.$$

But

$$\mathbf{E}_{Y|X}\left\{g(x)(Y-\mu_Y(x))\right\} = g(x)\left\{\mathbf{E}_{Y|X}Y - \mu_Y(x)\right\} = 0,$$

so $E_{Y|X}[(y - h(x))^2] = E_{Y|X}[(y - \mu_Y(x))^2] + g(x)^2$

which is minimized by setting $g(x) = 0$.

Transformations 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 $x_1,\ x_2, \ldots,\ x_n$. Then Y can take on values $y_1,\ y_2, \ldots,\ y_n$ with $y_i = h(x_i)$.

Thus, $P(Y = y_i) = P(X = h^{-1}(y_i))$, so that $f_Y(y) = f_X(h^{-1}(y))$

Continuous case: We need to consider two cases.

1. $h(\cdot)$ is increasing. Then $F_Y(y) = P(Y \le y) = P(X \le h^{-1}(y)) = F_X(h^{-1}(y))$. Thus

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

2. $h(\cdot)$ is decreasing. Then $F_Y(y) = P(Y \le y) = P(X \ge h^{-1}(y)) = 1 - F_X(h^{-1}(y))$. 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

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

Extension to the case where h is not one-to-one: Let $U_1, \ldots, U_m$ 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 $y \in R_i$.

Multivariate case: 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:

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

where $|J|$ is the absolute value of the Jacobian determinant of 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 $f_Y(y) = f_X(H^{-1}y)\,|H|^{-1}$.

Moments

The k^{th} moment of X is defined as $E(X^k)$.

The **mean** of X is the first moment, $E(X^1)$. It is denoted as $\mu = E(X)$

The k^{th} centered moment of X is defined as $E[(X-\mu)^k]$

The second centered moment is called the **variance** and is denoted σ^2.

A straightforward calculation shows

$$\sigma^2 = E[(X-\mu)^2] = E(X^2) - \mu^2$$

$\sigma \overset{def}{=} \sqrt{\sigma^2}$ is the **standard deviation**.

All odd centered moments are equal to zero for symmetric distributions. (A distribution is symmetric if $f(x) = f(-x)$. A distribution is symmetric around a point a if $f(x-a) = f(a-x)$.) The mean is the point of symmetry. (You should prove these results as an exercise) .

The first moment and 2-4th centered moments are used to measure the center (location), spread, skewness and kurtosis of the distribution.

Examples:

(1) Suppose X is uniformly distributed on $[0,1]$

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

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

(2) 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 $\mu_Y = a + b\mu_X$ and variance $\sigma_Y^2 = b^2\sigma_X^2$

Moment Generating Function

The Moment Generating Function of X is defined as $M(t) = \mathbf{E}(e^{tX})$

Since

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

$$M'(t) = \int xe^{tx} f(x)dx, \text{ so that } M'(0) = \mathbf{E}(X^1)$$

$$M''(t) = \int x^2 e^{tx} f(x)dx, \text{ so that } M''(0) = \mathbf{E}(X^2)$$

and in general

$$M^{(j)}(t) = \int x^j e^{tx} f(x)ds, \text{ so that } M^{(j)}(0) = \mathbf{E}(X^j)$$

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). We will modify the MGF shortly to produce a ***Characteristic*** *Function* which always exists.

A moment generating function uniquely characterizes a distribution. Thus, if X and Y have the same MGF, then they have the same CDF . (We will discuss this later in the context of characteristic functions.)

Examples:

(1) Suppose X is uniformly distributed on $[0,1]$, then

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

(2) Suppose X has MGF $M_X(t)$, and $Y = a + bX$, where a and b are constants. Then $M_Y(t) = e^{at}M_X(bt)$.

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

You should show that $\sigma_{XY} = E(XY) - \mu_X\mu_Y$.

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 $X = (X_1\ X_2\ \ldots\ X_n)'$.
Then

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

is the mean vector.

$\mathbf{E}(XX') = [\mathbf{E}(X_iX_j)]$ is the $n \times n$ second moment matrix

$\mathbf{E}[(X - \mu_X)(X - \mu_X)'] = [\mathbf{E}\{(X_i - \mu_i)(X_j - \mu_j)\}]$ is the $n \times n$ covariance matrix.

$\sigma_{ij} = \mathbf{E}\{(X_i - \mu_i)(X_j - \mu_j)\}$ is called the covariance between X_i and X_j.

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix}$$

is called the covariance matrix.

Notes: (Exercise: Show these results):

Σ is a symmetric $n \times n$ matrix since $\sigma_{ij} = \sigma_{ji}$.

$\Sigma = \mathrm{E}(XX') - \mu\mu'$.

If X_i and X_j are independent, then $\sigma_{ij} = 0$.

Let α denote a $n\times 1$ non-stochastic vector and let $Y = \alpha' X$. Then

$\mu_Y = \alpha'\mu_X$.

The variance of Y is $\Sigma_Y = \alpha'\Sigma_X\alpha$. Because $\Sigma_Y \geq 0$, Σ_X is positive semi-definite.

$\rho_{ij} = \sigma_{ij}/(\sigma_{ii}\sigma_{jj})^{\frac{1}{2}}$ 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 psd, 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

$$M_X(t) = \mathbf{E}(e^{t'X})$$

where t is a $n\times 1$ vector.

Selected Probability Distributions

Some Discrete Distributions

Bernoulli: X can take on two values, 0 and 1, $f(1)=p$ and $f(0)=1-p$. Thus, $f(x)=p^x(1-p)^{1-x}$

The parameter p indexes the distribution

Exercise: Work out MGF and all moments.

Binomial: Suppose X_i, $i = 1, \ldots, n$ are Independent and Identically Distributed (*i.i.d.)* 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.

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

where

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

is the number of ways that y 1s can occur in n (0,1) 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, central bank changes in the policy interest rate, etc.).

Let $g(x,w)$ denote the probability that x successes occur in a period of length w. Suppose

(i) $g(1,h)=\lambda h+o(h)$, where λ is a positive constant, $h>0$ ($o(h)$ means a term that satisfies $\lim_{h\to 0}[o(h)/h)]=0$),

(ii) $\sum_{x=2}^{\infty} g(x,h)=o(h)$

(iii) 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 $e^z = \sum_{k=0}^{\infty} \frac{1}{k!} z^k$.)

Some Continuous Distributions

Uniform: $f(x) = (b-a)^{-1}$ for $a \le x \le b$ and 0 elsewhere.

The MGF is

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

Univariate Normal **:**

Standard Normal (denoted $N(0,1)$ and often represented by Z):

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

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

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

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 $N(t,1)$). Thus, from the MGF:

$E(Z) = 0$

$E(Z^2) = \sigma_Z^2 = 1$

$E(Z^k) = 0$ for $k = 1, 3, 5, \ldots$

$E(Z^4) = 3$

MGF for General Normal:

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

and a direct calculation shows:

$E(Y) = \mu$

$E(Y^2) = \sigma^2 + \mu^2$, so that $Var(Y) = \sigma^2$

$E[(Y-\mu)^k] = 0$, for $k = 1, 3, 5, \ldots$

$E[(Y-\mu)^4] = 3\sigma^4$

Chi-Squared Distribution: Let Z_i, $i = 1, \ldots, n$ be distributed *i.i.d.* $N(0,1)$.

(where *i.i.d.* denotes *Independent and Identically Distributed*).

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 $Z \sim N(0,1)$ and $Y \sim \chi_n^2$ 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.

Multivariate Normal Distribution

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 $a \in \mathbb{R}^p$. (See Rao page 518)

It follows from this definition that if X is p-dimensionally normally distributed, then X_i is normally distributed.

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 $a \in \mathbb{R}^p$, we have

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

Note also that the moment generating function for X evaluated at a is the moment for the scalar $a'X$ evaluated at 1. That is:

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

and the final equality follows because $a'X \sim \mathrm{N}(a'\mu, a'\Sigma a)$. Because the MGF uniquely defines the probability distribution, the PDF for X must depend on the parameters μ and Σ.

This distribution is typically denoted: $X \sim N(\mu,\Sigma)$ (or sometimes $X \sim N_p(\mu,\Sigma)$, where the subscript p emphasizes 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 function of X appear, and it is useful to characterize the relevant PDF. 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.)

Theorem A. (Linear functions of X are normally distributed). Let $X \sim N_p(\mu,\Sigma)$, let B be a $k \times p$ matrix, let η denote a $k \times 1$ vector, and $Y = \eta + BX$

$$Y \sim N_k(\eta + B\mu, B\Sigma B')$$

Notes: The proof is straightforward.

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

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

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

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

Theorem C. (Independent normally distributed random variables have a joint normal distribution.) If $X_1 \sim N_p(\mu_1, \Sigma_1)$ and $X_2 \sim N_q(\mu_2, \Sigma_2)$, and X_1 and X_2 are independent, then $X = \left(X_1^{'}\ X_2^{'}\right)' \sim N_{p+q}(\mu, \Sigma)$, 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: Just write the joint density of $(X_1^{'}, X_2^{'})^{'}$ as a product of the densities of X_1 and X_2, and rearrange.

Theorem D. (Conditional normal distribution) Let $X \sim N_p(\mu,\Sigma)$. 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

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

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

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

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 $|\Sigma|=|\Sigma_{22}||V|$.

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. (Sums of independent normals) If $X_1 \sim N_p(\mu_1, \Sigma_1)$ and $X_2 \sim N_p(\mu_2, \Sigma_2)$, and X_1 and X_2 are independent,, then

$$X_1 + X_2 \sim N_p(\mu_1 + \mu_2\ , \Sigma_1 + \Sigma_2).$$

Notes: Immediate

Theorem F. (Marginal distributions from the multivariate normal) The marginal distribution of X_1 is $N_k(\mu_1, \Sigma_{11})$.

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

Notes: Write out the joint distribution to see that it factors.

Theorem H. (Characterizing independence of linear combinations of normal variables) If $X \sim N_p(\mu, \Sigma)$, 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 $B'\Sigma C = 0$.

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

Quadratic forms of normal random vectors:

We will call a quantity of the form $Y'AY$ 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 $B = ½ (A + A')$.) In all of the quadratic forms discussed below, we 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 $z_i \sim iid\text{N}(0,1)$ random variables for $i = 1, \ldots, n$. 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: letting Z denote the $n\times1$ 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: $Z = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$. Then we know (i) Z_1 is independent of Z_2 (ii) AZ_1 is normally distributed, (iii) $Z_2'Z_2 \sim \chi_{n_2}^2$, and (iv) AZ_1 and $Z_2'Z_2$ are independent.

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

Now, a few results:

Theorem I. (Quadratic form of centered X around inverse of covariance matrix) If $X \sim N_p(\mu, \Sigma)$ where Σ has rank p, then $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2_p$

Notes: Write $Z = \Sigma^{-1/2}(X - \mu)$, so that $(X-\mu)'\Sigma^{-1}(X-\mu) = Z'Z$, and the result follows by noting that $Z \sim N(0, I_p)$.

Theorem J. (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 K**:** 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 $PA = PG\Delta G' = 0$. 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 *L*: 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 1: Let Y_i, $i = 1, \ldots, n$ be distributed *i.i.d.* $N(\mu, \sigma^2)$. Let $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Show

(i) $\frac{(\bar{Y} - \mu)}{\sqrt{\sigma^2 / n}} \sim N(0,1)$

(ii) $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$

(iii) $\bar{Y}$ and s^2 are independent.

(iv) $\left[\frac{(\bar{Y} - \mu)}{\sqrt{\sigma^2 / n}}\right] / \left[\left(\frac{(n-1)s^2}{\sigma^2}\right)/(n-1)\right]^{1/2} = \frac{(\bar{Y} - \mu)}{\sqrt{s^2 / n}} \sim t_{n-1}$

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

Proof:

Notation: Let $Y_{1:n}$ denote the $n \times 1$ vector $(Y_1\ Y_2\ \ldots\ Y_n)'$. Let l denote an $n \times 1$ vector of 1s.

From Theorem C: $Y_{1:n} \sim \mathrm{N}_n(\mu l, \sigma^2 I)$. Letting $Z_{1:n} = (Y_{1:n} - \mu l)/\sigma$ then $Z_{1:n} \sim \mathrm{N}_n(0,\mathrm{I})$ (Theorem A) and $Y_{1:n} = \mu l + \sigma Z_{1:n}$ (rearranging the definition of $Z_{1:n}$).

Write $\bar{Y} = AY_{1:n}$ with $A = (l'l)^{-1}l'$

From Theorem A: $\bar{Y} \sim \mathrm{N}(\mu Al, \sigma^2 AIA')$. But $Al = 1$ and $AA' = n^{-1}$, so $\bar{Y} \sim \mathrm{N}(\mu, \sigma^2/n)$ (i) then follows from Theorem A.

For (ii), write $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = (Y_{1:n} - \bar{Y}l)'(Y_{1:n} - \bar{Y}l)$. Note that $Y_{1:n} - l\bar{Y} = MY_{1:n}$ with $M = I - l(l'l)^{-1}l'$. Note $MY_{1:n} = M(\mu l + \sigma Z_{1:n}) = \sigma MZ_{1:n}$ because $Ml = 0$. Thus $\sigma^{-2}(Y_{1:n} - l\bar{Y})'(Y_{1:n} - l\bar{Y}) = Z_{1:n}MZ_{1:n} \sim \chi^2_{rank(M)}$ from Theorem J.

The results follows by noting that rank(M) = trace(M) = trace($I_n - l(l'l)^{-1}l$) = n − trace[$l(l'l)^{-1}l$] = n − trace[$(l'l)^{-1}l'l$] = n-1.

For (iii) and using the results above, $s^2 = \sigma^2(n\text{-}1)^{-1} Z_{1:n}MZ_{1:n}$ and $\bar{Y} = AY_{1:n} = \mu l + \sigma AZ_{1:n}$, and the results follows from theorem K after noting the $MA = 0$.

(iv) follows from (i)-(iii).

Exercise 2: A generalization of exercise 1 uses Y_i each as a $p\times1$ vector, with $Y_i \sim$ i.i.d. $N_p(\mu, \Sigma)$ where now μ is an $p\times1$ vector, and considers the distribution of

$$T^2 = (\overline{Y} - \mu)' \left(\frac{1}{n-1} \sum_{i=1}^{n} (Y_i - \overline{Y})(Y_i - \overline{Y})' \right)^{-1} (\overline{Y} - \mu) .$$

In this exercise you will show that $[(n-p)/((n\text{-}1)p)]T^2 \sim F_{p,n-p}$.

You should carry out this exercise AFTER you have taken Week 2 with Bo Honore.

Carry out the following preliminary exercises:

1. Let $Z_0, Z_1, \ldots Z_q$, be *i.i.d.* $k\times1$ random vectors, with $Z_i \sim N_k(0,I)$. Let $A = \sum_{i=1}^{q} Z_i Z_i'$. Let $B = Z_0'(q^{-1}A)^{-1}Z_0$.

(a) Let $\iota = (1\ 0\ 0\ ..\ 0)'$ denote a $k\times1$ vector. Show that $\iota'\iota/(\iota'A^{-1}\iota) \sim \chi^2_{q-(k-1)}$, using the following steps:

(a.i) Write $Z_i = (Z_{i1}\ \ldots\ Z_{ik})'$. Consider the least squares problem:

$$SSR = \min_b \sum_{i=1}^{q} (Z_{i,1} - b_1 Z_{i,2} - b_2 Z_{i,2} - b_{k-1} Z_{i,k})^2$$

Show that SSR $\sim \chi^2_{q-(k-1)}$. (Hint: Review Hayashi Chapter 1 and or Honore's notes from one his early lectures.)

(a.2) Show that *SSR* from (a.i) satisfies $SSR = \iota'\iota/(\iota'A^{-1}\iota)$. (Hint: use the partioned inverse formula.)

(b) Generalize (a). Let L be an arbitrary $k\times1$ non-random vector with $L \neq 0$. Show that
$L'L/(L'A^{-1}L) \sim \chi^2_{q-(k-1)}$.

(c) Show that B, defined in part 1, can be written as $B = q\left[\frac{Z_0'Z_0}{Z_0'A^{-1}Z_0}\right]^{-1}(Z_0'Z_0)$.

(d)

(d.i) Use the result in (b) to show that $\left[\frac{Z_0'Z_0}{Z_0'A^{-1}Z_0}\right]^{-1} \Big| Z_0 = z_0 \sim \chi^2_{q-(k-1)}$.

(d.2) Does the distribution of $\left[\frac{Z_0'Z_0}{Z_0'A^{-1}Z_0}\right]^{-1}$ conditional on $Z_0 = z_0$ depend on

the value of z_0? Use your answer to show that $\left[\frac{Z_0'Z_0}{Z_0'A^{-1}Z_0}\right]^{-1}$ and Z_0 are independent.

(e) Show that the distribution of $Z_0'Z_0 \sim \chi^2_k$.

(f) Show that $\frac{\left(Z_0'Z_0\right)/k}{\left[\frac{Z_0'Z_0}{Z_0'A^{-1}Z_0}\right]/(q-(k-1))} \sim F_{k,q-(k-1)}$.

(g) Show that $\frac{q-(k-1)}{qk}B \sim F_{k,q-(k-1)}$.

(h) Let $X_0, X_1, \ldots X_q$, be *i.i.d.* $k\times1$ random vectors, with $X_i \sim N_k(0,\Sigma)$, where Σ is non-singular. Let

$C = X_0'\left(\frac{1}{q}\sum_{i=1}^{q} X_i X_i'\right)^{-1} X_0$. Show that $\frac{q-(k-1)}{qk}C \sim F_{k,q-(k-1)}$.

(i) With this background, prove the result stated in Exercise 2.

Some Useful Inequalities:

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

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 $x_0 = E(X)$, thus

$$E[h(X)\,] \geq h(x_0) + a\,E(X - x_0)$$

$$= h(E(X)) + a\,E(X - E(X)) = h(E(X)).$$

Note: If h is concave, $E[h(X)] \leq hE(X))$, by an analogous argument.

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

***Chebyshev's inequality*:** Let $\varepsilon > 0$, then $P(|X| \geq \varepsilon) \leq E(X^2)/\varepsilon^2$

Proof:

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

And rearranging yields the result.

***Markov's inequality*:** The following result is called Markov's inequality and is proved in analogous fashion: Let $\varepsilon > 0$, then $P(|X| \geq \varepsilon) \leq \frac{\mathbf{E}(|X|^p)}{|\varepsilon|^p}$

Large Sample Theory

Convergence of Sequences of Random Variables

Let $\{X_n\}$ denote a sequence of non-stochastic variables.
Recall $\lim_{n\to\infty} X_n = X$ if for any $\varepsilon > 0$, there exists a number N (which may depend on ε, so write it as $N(\varepsilon)$) with $|X_n - X| < \varepsilon$ for all $n > N(\varepsilon)$. We need to discuss convergence of random sequences $\{X_n(\omega)\}$ to random variables $X(\omega)$. There are a variety of notions of convergence:

Almost Sure Convergence: For a given ω we can ask whether $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ 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

$$X_n \xrightarrow{as} X \text{ if } P[\omega \mid \lim_{n\to\infty} X_n(\omega) = X(\omega)] = 1$$

***Convergence in Probability*:** For any $\varepsilon > 0$ we can calculate $p_n(\varepsilon) = \mathrm{P}(|X_n - X| > \varepsilon)$. If for any value of $\varepsilon > 0$, this sequence converges to 0, then we say that X_n converges in probability to X.

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

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

***Mean Square convergence*:** Let $ms_n = \mathbf{E}(X_n - X)^2$ 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

$$X_n \xrightarrow{ms} X \quad \text{if} \quad lim_{n\to\infty} \mathbf{E}(X_n - X)^2 = 0$$

***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*.

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

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 $F_{X_n}(x) = 1\left(x \geq \frac{1}{n}\right)$, while $F_X(x) = 1(x \geq 0)$. 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:

$X_n \Rightarrow X$ if

(1) $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ for all values of x where $F_X(.)$ is continuous.

(2) $\mathbf{E}(g(X_n)) \to \mathbf{E}(g(X))$ for any continuous bounded function g.

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

As it turns out, convergence in distribution obtains when $a'X_n \xrightarrow{d} a'X$ 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.

Relationships between the modes of convergence and some useful results

(1) If $X_n \xrightarrow{ms} X$ then $X_n \xrightarrow{p} X$.

Proof: From Chebychev's inequality $P(|X_n - X| \geq \varepsilon) \leq E(X_n - X)^2/\varepsilon^2$

(2) If $X_n \xrightarrow{as} X$ 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 $P(\omega|\ |X_n(\omega) - X(\omega)| > \varepsilon) < \delta$ 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 $P(\omega|\ |X_n(\omega) - X(\omega)| < \varepsilon) > 1 - \delta$, 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 $P(\omega|\ |X_n(\omega) - X(\omega)| > \varepsilon) < \delta$ for all $n > N$ as required.

Note: $X_n \xrightarrow{p} X$ does not imply that $X_n \xrightarrow{as} X$.

(3) If $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{d} X$.

Note: $X_n \xrightarrow{d} X$. does not imply $X_n \xrightarrow{p} X$.

Slutsky's theorem (Rao page 122)

(1) $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} 0$ implies $X_n Y_n \xrightarrow{p} 0$

(2) Let c be a constant and suppose $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then

(a) $X_n + Y_n \xrightarrow{d} X + c$

(b) $X_n Y_n \xrightarrow{d} Xc$

(c) $X_n/Y_n \xrightarrow{d} X/c$ if $c \neq 0$

(3) $(Y_n - X_n) \xrightarrow{p} 0$ and $X_n \xrightarrow{d} X$ then $Y_n \xrightarrow{d} X$

Continuous Mapping Theorem (Rao page 124)

Let $g(\cdot)$ be a continuous function, then

$X_n \xrightarrow{d} X$ implies $g(X_n) \xrightarrow{d} g(X)$

$X_n \xrightarrow{p} X$ implies $g(X_n) \xrightarrow{p} g(X)$

$(X_n - Y_n) \xrightarrow{p} 0$ and $Y_n \xrightarrow{d} Y$ then $g(X_n) - g(Y_n) \xrightarrow{p} 0$

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

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

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

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

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:

(1) 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 $s > 0$

$X_n + Y_n = o_p(\max\{f_n, g_n\})$

(2) 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 $s > 0$

$X_n + Y_n = O_p(\max\{f_n, g_n\})$

(3) If $X_n = o_p(f_n)$ and $Y_n = O_p(g_n)$, then

$X_n Y_n = o_p(f_n g_n)$

Laws of Large Numbers

A Weak Law of Large Numbers:

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

Proof:

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

where the first inequality follows from Chebyshev's inequality.

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

A Strong Law of Large Number

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

Central Limit Theorems

A CLT based on the moment generating function:

Lemma (Continuity Theorem): Let X_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 $M_0(t)$ be the moment generating function of the random variable X, which also exists on $(-h, h)$. If $\lim_{n\to\infty} M_n(t) = M_0(t)$ for all $t \in (-h, h)$, then $X_n \xrightarrow{d} X$.

A Central Limit Theorem: Let $Y_1, Y_2, \ldots$ denote a sequence of *i.i.d.* random variables with $E(Y_i) = 0$ and $\text{var}(Y_i) = 1$ and MGF $M_Y(t)$ that exists for $t \in (-h, h)$ for some $h > 0$. Then $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \Rightarrow Z \sim N(0,1)$.

Proof:

Using a mean value expansion:

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

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

Letting $Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$, then

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

where τ_n is between 0 and $t/n^{1/2}$.

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

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

Corollary: Let $X_1, X_2, ..$ 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 $\sqrt{n}(\bar{X} - \mu) \Rightarrow \sigma Z \sim N(0,\sigma^2)$.

Proof: Let $Y_i = \frac{X_i - \mu}{\sigma}$, and note that Y_i satisfies the assumption of the CLT. Note

$\sqrt{n}(\bar{X} - \mu) = \sigma \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$ and the result follows directly.

We did not go over the following in class, but here is a more general treatment of the CLT based on the characteristic function.

Characteristic Functions

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

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

where $i = \sqrt{-1}$. Thus, $C(t) = M(it)$. This change is useful because $e^{iz} = \cos(z) + i\sin(z)$ so that $|e^{iz}| = 1$ for all z. This means that $C(t)$ will always exists, while $M(t)$ exists only for certain distributions.

Some useful results:

1. Let $\alpha_r = \mathbf{E}(X^r)$, which is assumed to exist. Then $\frac{d^r C(t)}{dt^r} = i^r \int x^r e^{itx} dF(x)$ 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 $\kappa_0 = 0$, $\kappa_1 = \alpha_1 = \mu$, and $\kappa_2 = \alpha_2 - \alpha_1^2 = \sigma^2$.

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 $X_n \xrightarrow{d} X$, 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 $X_n \xrightarrow{d} X$.

A Central Limit Theorem

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

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n-\mu)\xrightarrow{d}N(0,1)$$

Proof:

Let $Z_n=\frac{\sqrt{n}}{\sigma}(\bar{X}_n-\mu)=\sum_{i=1}^{n}(\frac{X_i-\mu}{\sqrt{n}\sigma})$. Since $\mathbf{E}(\frac{X_i-\mu}{\sigma})=0$ and $var(\frac{X_i-\mu}{\sigma})=1$,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

$$\phi_{Z_n}(t)=\sum_{i=1}^{n}\left[-(\frac{1}{2})(\frac{t}{\sqrt{n}})^2+O[(\frac{t}{\sqrt{n}})^3\right]$$

$$=-\frac{1}{2}t^2+n\times O(\frac{t^3}{n^{3/2}})\to-\frac{1}{2}t^2$$

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

Example: Suppose that X_i are *i.i.d.* Bernoulli random variables with parameter p .

The CLT says that

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

which implies that for large n

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

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

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

Suppose $p=.25$ and n =100, and you are interested in the probability that $\overline{X}\leq 0.20$. Noting that $\overline{X}\leq 0.20$ is the same as $Y\leq 20,\ Y=\sum_{i=1}^{100}X_i$, the probability can be computed from the binomial distribution. A direct calculation yields: $P(Y\leq 20)=.14$. The normal approximation gives

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

Multivariate CLT:

Suppose $X_i \sim$ iid (μ, Σ). Then $\sqrt{n}(\bar{X}-\mu)\xrightarrow{d}N(0,\Sigma)$.

Sketch of proof: (We'll use the "Cramer-Wold device"). Let $Y_n = \sqrt{n}\left(\bar{X}-\mu\right)$, 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 N(0,Σ) random variable. Note that $M_{Y_n}(t) = E(e^{t'Y_n}) = E(\mathrm{e}^{n^{-1/2}\sum_{i=1}^{n}(t'X_i - t'\mu)})$. But $(t'X_i - t'\mu) \sim$ iid(0,$t'\Sigma t$), so the proof to the univariate CLT shows $M_{Y_n}(t) = M_{t'Y_n}(1) \to e^{t'\Sigma t/2}$

The Delta Method

Let Y_n denote a sequence of random variables, and let $X_n = n^{1/2}(Y_n - a)$, where a is a constant. Let $g(\cdot)$ be a continuously differentiable function.

Suppose $X_n \Rightarrow X \sim N(0, \sigma^2)$.

Then $n^{1/2}[g(Y_n) - g(a)] \Rightarrow X\,\partial g(a)/\partial a \sim N(0, [\partial g(a)/\partial a]^2\sigma^2)$.

Proof: By the mean value theorem

$$g(Y_n) = g(a) + (Y_n - a)\,\partial g(\tilde{Y}_n)/\partial \tilde{Y}_n$$

where $\tilde{Y}_n$ is between a and Y_n. Since $X_n \Rightarrow X$, $n^{-1/2}X_n \xrightarrow{p} 0$, so that $Y_n \xrightarrow{p} a$.

Since g is continuously differentiable $\partial g(\tilde{Y}_n)/\partial \tilde{Y}_n \xrightarrow{p} \partial g(a)/\partial a$. Thus

$$n^{1/2}[g(Y_n) - g(a)] = n^{1/2}[(Y_n - a)]\,\partial g(\tilde{Y}_n)/\partial \tilde{Y}_n \Rightarrow X\partial g(a)/\partial a \sim N(0, [\partial g(a)/\partial a]^2\sigma^2)$$

by Slutsky's theorem.

A similar result can also be proved for vectors. With Y_n, X, a, etc., denoting vectors: Suppose $X_n \Rightarrow X \sim N(0, \Sigma)$, then

$$n^{1/2}[g(Y_n) - g(a)] = [\partial g(\tilde{Y}_n)/\partial \tilde{Y}_n']\, n^{1/2}[(Y_n - a)] \Rightarrow GX \sim N(0, G\Sigma G)$$

where $G = \partial g(a)/\partial a'$.

Example: Suppose X_i is distributed *i.i.d.* $N(5,4)$ for $i = 1, \ldots, n$. Then we know

$$\bar{X} \sim N(5, 4/n)$$

or

$$\sqrt{n}\ (\bar{X} - 5) \sim N(0, 4)$$

Let $Y = \bar{X}^2$. The delta-method implies

$$\sqrt{n}\,(Y - 5^2) \overset{a}{\sim} N(0, 10 \times 4 \times 10)$$

Suppose $n = 100$ and we want to know $P(Y \leq 23.5)$.

An exact calculation yields $P(Y \leq 23.5). = 0.223$.

The delta-method yields:

$$P(Y \leq 23.5) = P\left(\frac{Y - 25}{(400/100)^{1/2}} \leq \frac{23.5 - 25}{(400/100)^{1/2}} \right) \approx P(Z \leq -0.75) = 0.227$$

Estimators

Let Y denote an $n \times 1$ vector of observations with CDF $F(y,\theta)$. 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, $i = 1, \ldots n$ be scalar *i.i.d.*$N(\mu,\sigma^2)$ random variables. Then $\mathbf{E}(Y_i) = \mu$ and $\mathbf{E}[(Y_i - \mu)^2] = \sigma^2$. Method-of-moment estimators are therefore

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

which use sample moments as estimators of corresponding population moments.

Properties of Estimators

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 the true value of the parameter is θ. (Think of Loss as the negative of utility).

For any $\hat{\theta} = g(Y)$, the expected value of the loss is

$$R(\hat{\theta}, \theta) = \mathbf{E}[L(\hat{\theta}, \theta)] = \mathbf{E}[L(g(Y)), \theta)].$$

$R(\hat{\theta}, \theta)$ is called the ***Risk Function***. (With Loss interpreted as the negative of utility, then risk is the negative of expected utility.)

A good estimator is an estimator that yields small risk (high expected utility). The best estimator has the smallest risk (highest expected utility).

In most cases, the risk of an estimator will depend on the value of θ (hence the notation $R(\hat{\theta}, \theta)$) and thus the "best" estimator will depend the value of θ. Since θ is unknown we must find an estimator that works well for a range of values of θ.

Examples:

(1) If we know that $\theta \in \Theta$, then we might try to find an estimator that solves

$$\min_{\hat{\theta}} \max_{\theta \in \Theta} R(\hat{\theta}, \theta)$$

This produces a mini-max estimator.

(2) We might want to find an estimator that minimizes the weighted average risk using a weight function $w(\theta)$. Thus we could consider

$$r(\hat{\theta}) = \int R(\hat{\theta}, \theta) w(\theta) d\theta$$

which is called the average risk of $\hat{\theta}$. The best estimator is the function $\hat{\theta}$ that minimizes $r(\hat{\theta})$.

Jargon: $R(\hat{\theta}, \theta)$ is sometimes called ***Classical*** or ***Frequentist*** risk.
$r(\hat{\theta})$ is often called ***Bayes*** risk.

Note that the average risk $r(\hat{\theta})$ depends on the weighting function w, so that different weighting functions give rise to difference measures of average (or Bayes) risk and different optimal Bayes estimators.

Admissability*:** $\hat{\theta}$ is said to be ***inadmissable is 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 from some θ. Thus, inadmissable estimators are dominated.

***Quadratic Loss*:** A useful loss function is

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

which is quadratic loss. The associated risk is called ***mean squared error*** (*m.s.e.*). Since

$$\hat{\theta}-\theta = [\hat{\theta}-E(\hat{\theta})] + [E(\hat{\theta})-\theta]$$

then

$$E[(\hat{\theta}-\theta)^2] = E[(\hat{\theta}-E(\hat{\theta}))^2] + [E(\hat{\theta})-\theta]^2 + E[(\hat{\theta}-E(\hat{\theta}))(E(\hat{\theta})-\theta)],$$

and the last term is equal to zero, so that

$$m.s.e, = \text{var}(\hat{\theta}) + [\text{Bias}(\hat{\theta})]^2$$

where the **Bias** is defined by

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

An estimator is ***unbiased*** if $\text{Bias}(\hat{\theta}) = 0$, or equivalently, $E(\hat{\theta}) = \theta$.

Exercise: X_i is *i.i.d.* $(\mu, 1)$. Consider the estimator $\hat{\mu}_1 = \overline{X}$ and $\hat{\mu}_2 = \frac{1}{2}\overline{X}$. Loss is quadratic. Derive the risk functions, $R(\hat{\mu}_1, \mu)$ and $R(\hat{\mu}_2, \mu)$.

Often it is difficult to deduce the exact distribution of an estimator, and so various approximations based on large-sample theory are used. The relevant jargon is

$\hat{\theta}$ is ***consistent*** if $\hat{\theta} \xrightarrow{p} \theta$. (When $\hat{\theta} \xrightarrow{as} \theta$, some say $\hat{\theta}$ is "strongly" consistent.)

If $\hat{\theta} - \theta$ is $O_p(1/a_n)$, then $\hat{\theta}$ is said to be a_n-consistent. Typically $a_n = n^{1/2}$ as will see.

Suppose some scaled and centered version of an estimator satisfies a CLT, e.g.,

$$a_n(\hat{\theta} - \gamma) \xrightarrow{d} N(0,1)$$

where a_n is sequence of real numbers (usually $a_n = \sqrt{n}/\sigma$) and γ is a constant. We then say that $\hat{\theta}$ is ***asymptotically normal***.

When $\hat{\theta}$ is asymptotically normal, probabilities involving $\hat{\theta}$ can be computed using the normal distribution: If

$$a_n(\hat{\theta} - \gamma) \xrightarrow{d} N(0,1)$$

then (at least for n large)

$$a_n(\hat{\theta} - \gamma) \overset{a}{\sim} N(0,1)$$

where I use the symbol $\overset{a}{\sim}$ to denote "approximately distributed as." Thus,

$$\hat{\theta} \overset{a}{\sim} N(\gamma, 1/a_n^2)$$

***Example*:** if $Y_i \sim i.i.d.\ (\mu, \sigma^2)$ then

$$(\sqrt{n}/\sigma)(\hat{\mu} - \mu) \xrightarrow{d} N(0,1)$$

where $\hat{\mu} = n^{-1}\sum_{i=1}^{n} Y_i = \overline{Y}$. This suggests using the approximation

$$\hat{\mu} \overset{a}{\sim} N(\mu, \sigma^2/n).$$

Bayes Estimators

Consider a probability model in which θ is random with pdf $w(\cdot)$. The thought experiment is that the random variable Y is generated in a two-step process. First, θ is drawn from w, and then Y is drawn from $f_{Y|\theta}$. Here we can think of the estimation problem as estimating the value of θ drawn in the first step after seeing the value of Y drawn in the second step.

In this setting, both θ and Y are random variables. The risk $R(\hat{\theta}, \theta)$ is the expected value of the loss, conditional on the value of θ: $R(\hat{\theta}, \theta) = E_{Y|\theta}[L(\hat{\theta}, \theta)]$. The overall risk (over both Y and θ) is $E_{Y,\theta}[L(\hat{\theta}, \theta)] = E_{\theta}\left[E_{Y|\theta}[L(\hat{\theta}, \theta)]\right] = E_{\theta}\left[R(\hat{\theta}, \theta)\right]$. Thus, the risk is

$$r(\hat{\theta}) = \int R(\hat{\theta}, \theta) w(\theta) d\theta,$$

which is the "average" or "Bayes" risk that was introduced above, where w is the marginal pdf of θ.

From what we know about marginal, joint and conditional distributions.

- The joint density of θ and Y is the product of the marginal density of θ, $w(\theta)$, and the conditional density of Y given θ, $f_{Y|\theta}$. Thus the joint density of Y and θ evaluated at $Y = y$ and $\theta = \tilde{\theta}$ is given by
$$f_{Y,\theta}(y, \tilde{\theta}) = f_{Y|\theta}(y \mid \tilde{\theta}) w(\tilde{\theta})$$
- The marginal density of Y is the joint density, integrated with respect to θ. That is
$$f_Y(y) = \int f_{Y,\theta}(y, \theta) d\theta = \int f_{Y|\theta}(y|\theta) w(\theta) d\theta.$$
- The conditional density of θ given $Y=y$ is given by

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

Jargon: $w(\theta)$ is called the ***prior density*** for θ. (It is the density of θ prior to seeing the value of Y.)

$f_{Y|\theta}(y \mid \theta = \tilde{\theta})$ is called the ***Likelihood*** function. With y fixed, it is a function of the value $\tilde{\theta}$.

$f_{\theta|Y}(\tilde{\theta} \mid Y = y)$ is called the ***posterior*** density of θ. (It is the density of θ after seeing the value of Y.)

$f_Y(y)$ is the called the ***marginal likelihood*** of y.

Using this notation:

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

and noting that the Marginal likelihood does not depend on θ:

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

The ***Posterior Risk***, or ***Posterior Expected Loss*** is $E_{\theta|Y}[L(\hat{\theta},\theta) \mid Y = y]$.

Bayes estimators are constructed to minimize posterior risk. When $L(\hat{\theta},\theta) = (\hat{\theta} - \theta)^2$, the Bayes estimator is $\hat{\theta}^{Bayes} = E_{\theta|Y}(\theta \mid Y = y)$, the mean of the posterior distribution. (This is the result that the mean minimizes quadratic loss that was proved earlier in these notes.)

Here's an important result: by minimizing posterior risk, Bayes estimators also minimize Average (or Bayes) risk, $r(\hat{\theta})$. Here's an identity for $r(\hat{\theta})$ that shows why:

$$
\begin{aligned}
r(\hat{\theta}) &= \int R(\hat{\theta},\theta) w(\theta) d\theta \\
&= \int \left[E_{Y|\theta} L(\hat{\theta},\theta) \right] w(\theta) d\theta \\
&= \int \left[\int L(\hat{\theta},\theta) f_{Y|\theta}(y \mid \theta) dy \right] w(\theta) d\theta \\
&= \int \int L(\hat{\theta},\theta) f_{\theta|Y}(\theta \mid y) f_Y(y) d\theta\, dy \\
&= \int E_{\theta|Y}[L(\hat{\theta},\theta) \mid Y = y] f_Y(y) dy
\end{aligned}
$$

Thus, the Bayes risk is the posterior risk averaged over all values of Y.

By minimizing $E_{\theta|Y}[L(\hat{\theta},\theta) \mid Y = y]$ for *each* value of y, the Bayes estimator minimizes the average value, $r(\hat{\theta})$.

Example 1:

$Y|\mu \sim N(\mu,1)$ 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 posterior. Thus, the posterior will only have mass at μ =2 and μ = 4.

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

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

Plugging in the numbers we find:

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

so

P($\mu = 4 \mid Y$=2.7) ≈ 1 –0.65 = 0.35.

Example 2: Now suppose the prior is $\mu \sim N(1,4)$. We carry out the same calculations

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

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 $f_{\mu}(m) \sim N(1,4)$, then

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

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

Example 3: Suppose $Y_i|\mu$ are *i.i.d.* $N(\mu, 1)$ and $\mu \sim N(\tau, \omega^2)$, where τ and ω are known constants. Let $Y = (Y_1, \ldots, Y_n)$ and $Y|\mu=m \sim N(ml, I_n)$, where l is an $n\times 1$ vector of 1's. $w(\theta)$ is therefore the $N(\tau, \omega^2)$ density and $f_{Y|\mu=m}$ is the $N(ml, I_n)$ density. The normal-normal densities imply that $(Y' \mu)'$ have a joint normal density, and you can verify that

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

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

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:

$\mu \mid Y = y \sim N(\lambda\tau + (1-\lambda)\bar{y},\ (1-\lambda)^2/n)$, where $\lambda = 1/(1+n\omega^2)$

This is the posterior for μ. If loss is quadratic, the Bayes estimator is therefore

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

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***.

Some Properties of Bayes Estimators:

Bayes Estimators are admissible: Consider a model in which θ can take on only k values, say $\theta_1, \ldots, \theta_k$. Let $\hat{\theta}^w$ denote a Bayes estimator using a prior $w(\theta) = w_1, \ldots, w_k$, where $w_i = P(\theta=\theta_i)$. Assume that w_i is positive for all θ_i.

Suppose that $\hat{\theta}^w$ is inadmissible. Then there exist an estimator $\hat{\theta}$ such that $R(\hat{\theta},\theta_i) \leq R(\hat{\theta}^w,\theta_i)$ for all i, and $R(\hat{\theta},\theta_i) < R(\hat{\theta}^w,\theta_i)$ for some i. But this means that $r(\hat{\theta}) = \sum_{i=1}^{k} w_i R(\hat{\theta},\theta_i) < \sum_{i=1}^{k} w_i R(\hat{\theta}^w,\theta_i) = r(\hat{\theta}^w)$. This is a contradiction because $\hat{\theta}^w$ is a Bayes estimator, and Bayes estimators minimize Bayes risk.

(Note: If we have time later in the semester we'll work through a "complete class theorem" which says that any admissible estimator can be interpreted as a Bayes estimator.)

***In general, Bayes Estimators are biased*:** By "biased" I mean $E_{Y|\theta=\theta_0}(\hat{\theta}^{Bayes}) \neq \theta_0$. That is, if θ is fixed at θ_0 and multiple draws of Y are obtained, the average value of $\hat{\theta}^{Bayes}$ is not equal to θ_0.

- In example 3, $\hat{\mu}^{Bayes} = \lambda\tau + (1-\lambda)\bar{y}$, so that

$$\begin{aligned} E(\hat{\mu}^{Bayes} \mid \mu = \mu_0) &= \lambda\tau + (1-\lambda)E(\bar{Y} \mid \mu = \mu_0) \\ &= \lambda\tau + (1-\lambda)\mu_0 \\ &= \mu_0 + \lambda(\tau - \mu_0) \end{aligned}$$

***Bayes Estimators minimize r*($\hat{\theta}^{Bayes}$) but may have large values *R*($\hat{\theta}^{Bayes}$, *θ) f*or certain values of *θ*:**

Consider example 3: Conditional on $\mu = \mu_0$ the Risk is

$$R(\hat{\mu}^{Bayes}, \mu_o) = [\text{Bias}(\hat{\mu}^{Bayes})]^2 + \text{Variance}(\hat{\mu}^{Bayes})]$$
$$= \lambda^2(\tau - \mu_0)^2 + (1-\lambda)^2 \text{var}(\bar{Y})$$
$$= \lambda^2(\tau - \mu_0)^2 + n^{-1}(1-\lambda)^2.$$

Which is quite large when τ differs significantly from μ_0.

In contrast, the frequentist risk of the estimator $\bar{Y}$ is $R(\bar{Y}, \mu_0) = n^{-1}$.

Note that $R(\bar{Y}, \mu_0) > R(\hat{\mu}^{Bayes}, \mu_0)$ for values of μ_0 close to τ,

but $R(\bar{Y}, \mu_0) \ll R(\hat{\mu}^{Bayes}, \mu_0)$ when $(\tau - \mu_0)^2$ is large.

(Note: $r(\bar{Y}) = n^{-1}$, while (a calculation shows) $r(\hat{\mu}^{Bayes}) = n^{-1}(1-\lambda)$. Thus (of course) $r(\hat{\mu}^{Bayes}) < r(\bar{Y})$, so that "on average" $R(\hat{\mu}^{Bayes}, \mu_0) < R(\bar{Y}, \mu_0)$, where the "averaging" uses the prior/weight function *w* for the values of μ_0.)

Unbiased Estimators

We'll now discuss some a general result about unbiased estimators and then study the asymptotic properties of maximum likelihood and Bayes estimators. A useful result in this regard is the ***Cramer-Rao inequality***, which gives a lower bound on the variance of any unbiased estimator.

Cramer-Rao Inequality

Some preliminaries: Suppose $Y \sim F(y \mid \theta)$ with density $f(y \mid \theta)$. Then

$$1 = \int f(y \mid \theta) dy,$$

Differentiating both sides, and assuming the support of Y does not depend on θ

$$0 = \int \frac{\partial f(y \mid \theta)}{\partial \theta} dy$$

Let

$$S(\theta, y) = \frac{\partial \ln f(y \mid \theta)}{\partial \theta}$$

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

$$\frac{\partial f(y \mid \theta)}{\partial \theta} = S(\theta, y) \times f(y \mid \theta)$$

so that

$$0 = \int \frac{\partial f(y \mid \theta)}{\partial \theta} dy = \int S(\theta, y) f(y \mid \theta) dy = E[S(\theta, Y)].$$

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:

$$0 = \int \frac{\partial S(\theta, y)}{\partial \theta} f(y \mid \theta) dy + \int S(\theta, y)^2 f(y \mid \theta) dy$$

so that

$$-E\left[\frac{\partial S(\theta, Y)}{\partial \theta}\right] = E\left[S(\theta, Y)^2\right] = var\left[S(\theta, Y)\right]$$

Let

$$I(\theta) = -E\left[\frac{\partial S(\theta, Y)}{\partial \theta}\right] = -E\left(\frac{\partial^2 \ln[f(Y \mid \theta)]}{\partial \theta^2}\right) = E\left[S(\theta, Y)^2\right] = var\left[S(\theta, Y)\right]$$

which is called the ***Information***.

Now, let $\hat{\theta} = g(Y)$ denote an unbiased estimator of θ. Then

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

Differentiating both sides with respect to θ yields:

$$1 = \int g(y) S(\theta, y) f(y \mid \theta) dy$$

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

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

so that

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

and thus

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

which is the Cramer-Rao inequality.

The same set of results obtain when θ is a $k \times 1$ vector.

$S(\theta,Y)$ is a $k \times 1$ Score vector with $E[S(\theta,Y)] = 0$.

$Var(S(\theta,Y)) = E(S(\theta,Y)S(\theta,Y)') = -E\left(\partial S(\theta,Y / \partial \theta'\right) = I(\theta)$ a $k \times k$ Information matrix

If $\hat{\theta}$ is an unbiased estimator, then $E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)'] \geq I(\theta)^{-1}$.

Maximum Likelihood Estimators

Let Y denote a random vector with density $f(y \mid \theta)$. Then

$$Lik(\theta) = f(Y \mid \theta)$$

the density of Y evaluated at $Y = y$ and viewed as function of θ is referred to as the ***Likelihood Function***.

Let $Y_1, Y_2, \ldots, Y_n$ be *iid*, each with density $f(Y \mid \theta)$ Then

$$f(Y_{1:n} \mid \theta) = \prod_{i=1}^{n} f(Y_i \mid \theta)$$

is the likelihood function, where I have used the notation $Y_{1:n}$ to denote $Y_1, Y_2, \ldots, Y_n$
Let

$$L_n(\theta) = \ln(f(Y_{1:n} \mid \theta))$$

denote the log-likelihood function.
Suppose that θ is a $k \times 1$ vector. Let

$$S_i(\theta) = \partial L(Y_i \mid \theta) / \partial \theta$$

and

$$S_{1:n} = \sum_{i=1}^{n} S_i(\theta)$$

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

Let

$$I(\theta) = \mathrm{E}(S_i(\theta)S_i(\theta)') = -E[\partial S_i(\theta) / \partial \theta']$$

where there is no i subscript on I because Y_i are *iid*.

Let

$$I_{1:n}(\theta) = nI(\theta) = \mathrm{E}[S_{1:n}(\theta)S_{1:n}(\theta)']$$

denote the information in the sample.

Let $\hat{\theta}_{MLE}$ solve

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

$\hat{\theta}_{MLE}$ is the ***maximum likelihood estimator*** of θ.

Some asymptotic properties of MLEs (consistency, asymptotic normality, efficiency):

Given a set of "regularity" conditions:

$$\hat{\theta}_{MLE} \xrightarrow{p} \theta_o$$

and

$$n^{1/2}(\hat{\theta}_{MLE} - \theta_0) \Rightarrow X \sim N(0, I(\theta_0)^{-1})$$

so that

$$\hat{\theta}_{MLE} \overset{a}{\sim} \mathrm{N}(\theta_0, n^{-1} I(\theta_0)^{-1})$$

where θ_o is the true value of θ. Note that $nI(\theta_0) = I_{1:n}(\theta_0)$.

Sketch of consistency proof under iid sampling:

Let

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

where θ_o is the true value of θ and E_{θ_o} means taking the expected value using the density $f(y|\theta_o)$. Note that $C(\theta_0) = 0$.

Suppose $\theta \neq \theta_0$

$$C(\theta) = \mathbf{E}_{\theta_o} \ln\left[\frac{f(Y|\theta)}{f(Y,|\theta_o)}\right] \leq \ln \mathbf{E}_{\theta_o}\left[\frac{f(Y|\theta)}{f(Y|\theta_o)}\right] = \ln(1) = 0$$

where the first inequality follows from Jensen's inequality since the log function is concave, and the inequality is strict unless $\ln\left[\frac{f(Y|\theta)}{f(Y|\theta_0)}\right]$ has a degenerate distribution. When it does, we say that θ is **unidentified**, otherwise θ is **identified**. Thus, when θ is identified $C(\theta)$ is uniquely maximized at $\theta = 0$.

To complete the discussion of consistency, assume that

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

uniformly in θ over some set Θ. (This is *Uniform* LLN result —- see, for example, Gallant, A. R. (1997), *An Introduction to Econometric Theory*, Princeton University Press., page 135).

This means that the maximizer of $C_n(\theta)$ converges to the maximizer of $C(\theta)$, which we just showed was 0.

Thus the maximizer of $n^{-1}\sum\ln(f(Y_i,\theta) = n^{-1}L_n(\theta)$ converges to θ_0, so the MLE is consistent.

Sketch of Asymptotic Normality

First (**Asymptotic normality of score**):

$$\frac{1}{\sqrt{n}} S_{1:n}(\theta_0) \xrightarrow{d} N(0, I(\theta_0))$$

follows immediately from applying the CLT to $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_i(\theta_0)$.

Next (**mean-value expansion**)

$$S_n(\hat{\theta}_{MLE}) = S_n(\theta_0) + \frac{\partial S_n(\tilde{\theta})}{\partial \theta}(\hat{\theta}_{MLE} - \theta_0)$$

where $\tilde{\theta}$ is between θ_0 and $\hat{\theta}_{MLE}$.

(**First Order Conditions for Maximum**): Since $S_n(\hat{\theta}_{MLE}) = 0$,

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_0) = \left[\left\{ -\frac{1}{n} \frac{\partial S_n(\tilde{\theta}} {\partial \theta} \right\} \right]^{-1} \left[\frac{1}{\sqrt{n}} S_n(\theta_0) \right]$$

and (**Asymptotic behavior of Hessian**)

$$-\frac{1}{n} \frac{\partial S_n(\tilde{\theta})}{\partial \theta} \xrightarrow{p} \bar{I}(\theta_0)$$

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

Thus

$$\sqrt{n}(\hat{\theta}_{MLE} - \theta_o) \xrightarrow{d} N(0, \bar{I}(\theta_o)^{-1})$$

by Slutsky's Theorem.

These results also hold for vector $\hat{\theta}_{MLE}$ and vector values $S_{1:n}(\theta_0)$, etc.

Method of Moment Estimators

Suppose Y_i, $i = 1, \ldots, n$ is a sequence of *i.i.d.*(μ, Σ) random $l \times 1$ vectors.
The method of moments estimator of μ is

$$\hat{\mu}_{MM} = n^{-1} \sum Y_i.$$

From the LLN and CLT, we have

$$\hat{\mu}_{MM} \xrightarrow{as} \mu$$

and

$$\sqrt{n}(\hat{\mu}_{MM} - \mu) \xrightarrow{d} N(0, \Sigma).$$

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

Now suppose that $\mu = h(\theta_0)$ where μ is $l \times 1$, θ_0 is $k \times 1$ with $k \leq l$. The goal is to estimate θ_0. 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 $\hat{\theta}_{MM}$ denote the method of moments estimator. The properties of $\hat{\theta}_{MM}$ can be derived in a way that parallels the discussion of the maximum likelihood estimator.

Consistency follows by arguing that $J_n(\theta) \to J(\theta)$ where $J(\theta)$ is minimized at $\theta = \theta_0$.

Asymptotic normality is proved using the following steps

(*Asymptotic normality of gradient*): The gradient is

$$g_n(\theta) = \frac{\partial J_n(\theta)}{\partial \theta} = -2[\frac{\partial h(\theta)}{\partial \theta'}]'(\bar{Y} - h(\theta))$$

so that

$$\sqrt{n} g_n(\theta_o) = -2[\frac{\partial h(\theta_o)}{\partial \theta'}]'[\sqrt{n}(\bar{Y} - h(\theta_o))] \xrightarrow{d} N(0, 4[\frac{\partial h(\theta_o)}{\partial \theta'}]' \Sigma [\frac{\partial h(\theta_o)}{\partial \theta'}]$$

(*Mean Value Expansion*): Linearize $g_n(\hat{\theta}_{MM})$ around $g_n(\theta_o)$ and solve for $\hat{\theta}_{MM}$.

$$g_n(\hat{\theta}_{MM}) = g_n(\theta_0) + \frac{\partial g_n(\tilde{\theta})}{\partial \theta'}(\hat{\theta}_{MM} - \theta_0)$$

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

(*Asymptotic Behavior of Hessian*): Show $\frac{\partial g_n(\tilde{\theta})}{\partial \theta'} \xrightarrow{p} 2H$ where

$H = \left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]' \left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]$ is a a constant, non-singular matrix. To do this, write

$$\frac{\partial g_n(\theta)}{\partial \theta'} = 2\left[\frac{\partial h(\theta)}{\partial \theta'}\right]' \left[\frac{\partial h(\theta)}{\partial \theta'}\right] + m_n(\theta)(\bar{Y} - h(\theta))$$

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 2*H*. To finish argument, expand $g_n(\hat{\theta}_{MM})$ around $g_n(\theta_o)$ to yield

$$\sqrt{n}(\hat{\theta}_{MM} - \theta_o) = \left[\frac{\partial g_n(\tilde{\theta})}{\partial \theta'}\right]^{-1} \left[\sqrt{n} g_n(\theta_0)\right] \xrightarrow{d} N\left(0, H^{-1}\left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]' \Sigma \left[\frac{\partial h(\theta_o)}{\partial \theta'}\right] H^{-1}\right)$$

so that $\hat{\theta}_{MM} \overset{a}{\sim} N(\theta_o, V_n)$, where $V_n = \frac{1}{n} H^{-1}\left[\frac{\partial h(\theta_o)}{\partial \theta'}\right]' \Sigma \left[\frac{\partial h(\theta_o)}{\partial \theta'}\right] H^{-1}$.

Sufficiency

A key task is 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 affecting 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:

(1) Suppose Y_1 and Y_2 are *iid* Bernoulli random variables with $P(Y_i = 1) = \theta$. 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 $P(\{0,0\}|S = 0) = 1$; similarly if $S = 2$, then $Y_1 = Y_2 = 1$, so $P(\{1,1\}|S = 2) = 1$. 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 $P(\{0,1\}|S = 1) = P(\{1,0\}|S = 1) = 0.5$. In all of these case $P(y|S=s)$ does not depend on the value of θ, so S is a sufficient statistic.

(2) Suppose $Y_i \sim i.i.d.\ N(\mu,1)$, for $i = 1, \ldots, n$. Equivalently $Y \sim N(\mu \times l, I_n)$, where l is an $n \times 1$ vector of 1's. Let $S(Y) = \bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$ denote the sample mean. Using the conditional normal formula, the pdf of $Y \mid S$ is normal with mean vector $\mu \times l + l \times (n^{-1}/\, n^{-1})(\bar{Y} - \mu) = \bar{Y} \times l$, and covariance matrix $I_n - n^{-1}ll'$. Because this conditional distribution does not depend on μ, $\bar{Y}$ is sufficient for μ.

2 useful results for Sufficient Statistics:

(1) Factorization Theorem: Let $f_Y(y;\theta)$ denote the density of Y. Then S is a sufficient statistic 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:

(i) as a mechanical matter it provides another way to check that a candidate S is sufficient. For example, in the $Y_i \sim iidN(\mu,1)$ example, the pdf $f_Y(y;\theta)$ 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.

(ii) 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 the MLE of θ is

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

and the Bayes posterior is

$$f(\theta \mid Y=y) = \frac{f_{Y|\theta}(y \mid \theta)w(\theta)}{\int f_{Y|\theta}(y \mid \theta)w(\theta)\,d\theta} = \frac{h(y)f_{S|\theta}(s \mid \theta)w(\theta)}{\int h(y)f_{S|\theta}(s \mid \theta)w(\theta)\,d\theta}$$
$$= \frac{f_{S|\theta}(s \mid \theta)w(\theta)}{\int f_{S|\theta}(s \mid \theta)w(\theta)\,d\theta} = f(\theta \mid S=s)$$

(2) Rao-Blackwell Theorem:

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

$\sigma_Y^2 = var(\mu(X)) + var(Y - \mu(X))$. This implies that $var(\mu(X)) \leq \sigma_Y^2$.

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

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

Note that while $\tilde{\theta}(S) = \mathbf{E}[\hat{\theta}(Y)\,|S]$ 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, $\mathbf{E}(\hat{\theta}(Y)) = \mathbf{E}(\mathbf{E}(\hat{\theta}(Y) \mid S)) = \mathbf{E}(\tilde{\theta}(S))$ from the law of iterations, so $\tilde{\theta}(S)$ is unbiased, and from our result above, it has a variance that is weakly smaller than $\hat{\theta}(Y)$. 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, $\tilde{\theta}(S)$.

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

Hypothesis Testing

We will first cover hypothesis testing in a specific (important) example. We'll then move on to more general discussion of the hypothesis testing problem. Suppose we are interested in a $k\times1$ vector of parameters, say θ, that characterize 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_0 is called the **null hypothesis** and H_a is called the **alternative hypothesis**. How might we decide between H_0 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 $||\hat{\theta} - \theta_0||$ is likely to be small.

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

We are helped with "likely" and "small" and "large" because we know that $\hat{\theta} \sim N(\theta, \Omega)$. 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 the decision rule of the form:

(1) Choose H_0 if $\xi \leq c$ and (2) Choose H_a if $\xi > c$.

where the number c 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 c? That's easy. c

solves

$$P(\xi > c \mid \theta = \theta_0) = \alpha$$

so that c is the 1-α percentile of the χ^2_k distribution.

The **power** of the test is defined as $P(\xi > c \mid H_a$ is true). But because H_a includes many value of θ (the hypothesis is said to be **composite**), the power will be different for the different value 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 $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V)$. 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 $P(\xi > c) \to 1$ for any fixed value of c. The test therefore has power = 1 for any (fixed) value of θ under the alternative.

When power → 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.

Hypotheses involving linear functions of θ:

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

H_0: $R\theta = r_0$ where r_0 is known value and H_a: $R\theta \neq r_0$.

Note that $R\hat{\theta} \sim N(R\theta, R\Omega R')$, 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.)

Hypotheses involving nonlinear functions of θ:

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

H_0: $R(\theta) = r_0$ where r_0 is known value and H_a: $R(\theta) \neq r_0$.

The delta-method implies

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

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

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

With this important specific case out of the way, let's discuss the hypothesis testing problem more generally.

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: $Power = 1 - P(type\ 2\ error)$.

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

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

$$LR(Y) = \frac{Lik_a(Y)}{Lik_o(Y)}$$

where $Lik_a(Y)$ and $Lik_o(Y)$ are the likelihoods under the alternative and null, respectively. The critical region for a test with size α is

$$W_\alpha = \{y \mid LR(y) > c_\alpha\}$$

where c_α is chosen so that

$$P[LR(Y) > c_\alpha \mid Y \sim F_o\,] = \alpha$$

The proof of this result is given on the next page.

Suppose the random variables have a continuous distribution with density f_a and f_o under the alternative and null. Then $Lik_o = f_o$ and $Lik_a = f_a$.

Let W_α denote the NP critical region. Let X_α denote any other critical region with size α. Note

$$W_\alpha = (W_\alpha \cap X_\alpha) \cup (W_\alpha \cap X_\alpha^c)$$

and

$$X_\alpha = (X_\alpha \cap W_\alpha) \cup (X_\alpha \cap W_\alpha^c)$$

Now (because tests have size α):

$$\alpha = \int_{W_\alpha} f_o(y)dy = \int_{X_\alpha} f_o(y)dy$$

so that

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

which implies

$$\int_{W_\alpha \cap X_\alpha^c} f_o(y)dy = \int_{X_\alpha \cap W_\alpha^c} f_o(y)dy$$

But, for any $Y \in W_\alpha$ (and hence for any $Y \in (W_\alpha \cap X_\alpha^c)$), $f_a(Y) > c\, f_o(Y)$, and for any $Y \in W_\alpha^c$ (and hence for any $Y \in (X_\alpha \cap W_\alpha^c)$) $f_a(Y) \leq c\, f_o(Y)$). Thus

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

so that

$$\int_{W_\alpha} f_a(y) = \int_{W_\alpha \cap X_\alpha} f_a(y) + \int_{W_\alpha \cap X_\alpha^c} f_a(y)dy \geq \int_{X_\alpha \cap W_\alpha} f_a(y)dy + \int_{X_\alpha \cap W_\alpha^c} f_a(y)dy = \int_{X_\alpha} f_a(y)dy$$

or

$$P(Y \in W_\alpha \mid Y \sim F_a) \geq P(Y \in X_\alpha \mid Y \sim F_a).$$

Parametric Restrictions

Write the density of Y as $f(y,\theta)$, where θ is a $k\times 1$ vector of parameters. Suppose $\theta\in\Theta$, where

$$H_o\colon \theta\in\Theta_o$$

$$H_a\colon \theta\in\Theta_a$$

where $\Theta=\Theta_o\cup\Theta_a$ and $\Theta_o\cap\Theta_a=\emptyset$.

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

$$H_o\colon \mu=\mu_0$$

$$H_a\colon \mu=\mu_a$$

with $\mu_o \neq \mu_a$. Note

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

and thus

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

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

This means that we can write the LR testing procedure as

Reject H_o when $\bar{Y}>c_a$ where c_a is chosen so that

$$P(\bar{Y}>c_a\mid \bar{Y}\sim N(\mu_o,\frac{1}{n}))=\alpha$$

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.

Maximizing Weighted Average Power:

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

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

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 $\int_W f(y \mid \theta)dy$, so that weighted average power is

$WAP = \int_{\Theta_a}\left[\int_W f(y \mid \theta)dy\right]w(\theta)d\theta$. This can be written as

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

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

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

Example 1: $Y_i \sim iid$ N(μ,1), where Y is a scalar. We are interested in H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$. Without loss of generality, set $\mu_0 = 0$. Suppose we put weight of ½ 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(μ,1/n). A calculation shows that the WAP test rejects for large values of $\zeta = e^{n\overline{Y}} + e^{-n\overline{Y}} = e^{n|\overline{Y}|} + e^{-n|\overline{Y}|} = \zeta(|\overline{Y}|)$, where the function is increasing in the value of $|\overline{Y}|$. Thus, the test rejects for large values of $|\overline{Y}|$.

Example 2: $Y_i \sim iid$ $N_k(\mu,\Sigma)$ where Y is a $k \times 1$ vector. We are interested in H_0: $\mu = \mu_0$ versus H_a: $\mu \neq \mu_0$. Suppose we use a weight function with $\mu \sim N(0,\omega^2\Sigma)$. In this case we can see that the distribution of Y under this weighted average alternative is $H_{a,\text{weightd}}$: $Y \sim N(0,(1+\omega^2)\Sigma)$. Using sufficient statistics, the null and weighted-average alternative are:

$$H_0: \overline{Y} \sim N(\mu_0, n^{-1}\Sigma) \text{ versus } H_{a,\text{weights}}: \overline{Y} \sim N(0, (1+n\omega^2)n^{-1}\Sigma)$$

The WAP test is then the LR, which is

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

so the test rejects for large values of the exponent. When $\omega^2 \to \infty$, 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.

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 $\hat{\theta}$ 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 $\xi_{LR} > c$, where c is a critical value that satisfies $P_{H_o}(\xi_{LR} > c) = \alpha$, where α is the size of the test. To find c 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 $\frac{\partial L_n(\hat{\theta})}{\partial \theta} = 0$

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

From our earlier results

$$\sqrt{n}(\hat{\theta} - \theta_o)] \xrightarrow{d, H_o} N(0, \bar{I}(\theta_o)^{-1})$$

and

$$\left[-\frac{1}{n}\frac{\partial^2 L_n(\tilde{\theta})}{\partial\theta\,\partial\theta'}\right]=-\frac{1}{n}\sum\frac{\partial^2 \ln f(Y_i,\tilde{\theta})}{\partial\theta\,\partial\theta'}\xrightarrow{p,H_o}\bar{I}(\theta_o).$$

Thus

$$\xi_{LR}\xrightarrow{d,H_0}\xi\sim\chi_k^2$$

This final result follows from noting that ξ_{LR} is asymptotically a quadratic form of a $N(0,\bar{I})$ 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 last week, using the MLE of θ. They differ in only to the extent that they use different estimators of the covariance matrix.

Thus, as we discussed last week suppose $\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}N(0,V)$

We could then form a statistic:

$\xi = n(\hat{\theta}\text{-}\theta_0)'\hat{V}^{-1}(\hat{\theta}\text{-}\theta_0)$ where $\hat{V}$ is consistent for V. Then $\xi\xrightarrow{d,H_o}\chi_k^2$ random variable.

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

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

$$\hat{V}=\left[\frac{1}{n}\sum s_i(\theta_0)s_i(\theta_0)'\right]^{-1}$$

$$\hat{V}=\frac{1}{n}\sum s_i(\tilde{\theta})s_i(\tilde{\theta})'$$

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

Score tests: use $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i(\theta_0)$ … etc.

Confidence Sets

A $(1-\alpha)\times 100\%$ confidence set for θ is a (random) set of values of θ that contains θ_o, the true value of with probability $1-\alpha$. Let $C(Y)$ denote such a set. That is, suppose that $P[\theta_o \in C(Y)] = 1 - \alpha$. (Note that the randomness comes from Y conditional on $\theta = \theta_0$).

An easy way to construct such a set is to use hypothesis tests. 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 using a test with size α. Let $C(Y)$ denote the set of values for which the test does not reject the null.

Note that, since $\theta_o \in \Theta$, the null $\theta = \theta_o$ was one of the tests constructed. This test rejected the null with probability α, hence did not reject with probability $1-\alpha$. Hence $P[\theta_o \in C(Y)] = 1 - \alpha$.

When the hypothesis test is carried out using a Wald-statistic with a limiting χ^2 distribution, the confidence set is particularly easy to form. Note that H_o: $\theta = \theta_0$ is not rejected using a test of size α if

$$\xi \leq \chi^2_{\kappa,1-\alpha}$$

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

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

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

$$C(Y) = \{\theta \,|\, \hat{\theta} - Z_{1-\frac{\alpha}{2}}[\frac{1}{n}\widehat{V}]^{\frac{1}{2}} \leq \theta \leq \hat{\theta} + Z_{1-\frac{\alpha}{2}}[\frac{1}{n}\widehat{V}]^{\frac{1}{2}}\}$$

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

Note that confidence sets are "frequentist" constructs. They treat θ as fixed and the randomness in the set comes from Y in $C(Y)$. Bayes methods can be used to construct analogous sets called ***credible sets***, although their interpretations differ from confidence sets. Credible sets are discussed below.

Efficient Confidence Sets and Pratt's Insights

We have discussed constructed 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 insights.

Let $C(Y)$ denote a confidence set for a parameter θ. The "volume" of $C(Y)$ is

$$V_C(Y) = \int 1[\theta \in C(Y)]d\theta$$

where 1[ ] is the indicator function. ($1[x] = 1$ if x is 'true' and $1[x] = 0$ 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:

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

which serves as a criteria for evaluating confidence sets: $C_1(Y)$ is preferred to $C_2(Y)$ if $R_{C1} < R_{C2}$.

Now, consider the testing problem: H_o: $Y \sim f_\theta$ versus H_a: $Y \sim f$. If we have a 1-α 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 II error (i.e., 1-Power) is therefore

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

Now, rewrite the expression for expected volume:

$$\begin{aligned} R_C &= E[V_c(Y)] = \int V_c(y) f(y)\, dy \\ &= \int \left[\int 1\big(\theta \in C(y)\big)\, d\theta \right] f(y)\, dy \\ &= \int \left[\int 1\big(\theta \in C(y)\big)\, f(y)\, dy \right] d\theta \end{aligned}$$

So that R_C can be minimized by choosing a test, $1(\theta \in C(y)$, that minimizes the probability of type II error .. i.e., that maximizes power. This is achieved using a Neyman-Pearson test.

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

Example:

Above we considered the testing problem with $Y_i|\mu \sim$ i.i.d. $N_k(\mu, \Sigma)$ and

H_0: $\mu = \mu_0$ versus H_a: $\mu \sim N(0,\omega^2\Sigma)$.

We showed that, when ω^2 was large, the optimal test was the Wald test

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

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

More on Bayes Procedures

Recall the "normal-normal" example we studied a few days ago with

$Y_i|\mu \sim i.i.d.N(\mu, \sigma^2)$ and $\mu \sim N(m, \omega^2)$. Because $\overline{Y}$ is sufficient for μ, we can summarize the sample information with the conditional distribution $\overline{Y}|\mu \sim N(\mu, \sigma^2/n)$. The properties of the normal tell us that

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

The posterior for μ follows directly: $\mu|\overline{Y} \sim N(\lambda_n \overline{Y} + (1-\lambda_n)m, \omega^2(1-\lambda_n))$ with

$$\lambda_n = \frac{\omega^2}{\omega^2 + \sigma^2/n} .$$

If loss is quadratic, risk is MSE and the Bayes estimator is the posterior mean: $\hat{\mu}_{Bayes} = \lambda_n \overline{Y} + (1-\lambda_n)m$.

Some properties of:

(a.0) $E(\mu|\overline{Y}) = \hat{\mu}_{Bayes} = \lambda_n \overline{Y} + (1-\lambda_n)m$

(b.0) $var(\mu|\overline{Y}) = \omega^2(1-\lambda_n)$

(a.1) $E(\hat{\mu}_{Bayes}) = m$

(b.1) $var(\hat{\mu}_{Bayes}) = \lambda_n^2 (\sigma^2/n + \omega^2)$

(c.1) $E(\hat{\mu}_{Bayes} - \mu)^2 = \lambda_n^2 \sigma^2/n + (1-\lambda_n)^2\omega^2$

(a.2) $E(\hat{\mu}_{Bayes}|\mu) = \lambda_n\mu + (1-\lambda_n)(m-\mu)$

(b.2) $var(\hat{\mu}_{Bayes}|\mu) = \lambda_n^2 \sigma^2/n$

(c.2) $E[(\hat{\mu}_{Bayes} - \mu)^2 | \mu] = \lambda_n^2 \sigma^2/n + (1-\lambda_n)^2(m-\mu)^2$

(d.0.1) $\left.\dfrac{\mu-\lambda_n\overline{Y}-(1-\lambda_n)m}{\sqrt{\omega^2(1-\lambda_n)}}\right|\overline{Y}\sim \mathrm{N}(0,1)$

(d.0.2)
$P\left(\left(\lambda_n\overline{Y}+(1-\lambda_n)m-1.96\sqrt{\omega^2(1-\lambda_n)}\right)\le\mu\le\left(\lambda_n\overline{Y}+(1-\lambda_n)m+1.96\sqrt{\omega^2(1-\lambda_n)}\right)|\,\overline{Y}\right)$
$= 0.95$

(Note: (d.0) hold unconditionally as well .. over the joint normal distribution given above)

(d.2.1) $\left.\dfrac{\overline{Y}-\mu}{\sqrt{\sigma^2/n}}\right|\mu\sim \mathrm{N}(0,1)$

(d.2.2) $P\left(\left(\overline{Y}-1.96\sqrt{\sigma^2/n}\right)\le\mu\le\left(\overline{Y}+1.96\sqrt{\sigma^2/n}\right)|\,\mu\right)=0.95$

(Note: (d.2) hold unconditionally as well .. over the joint normal distribution given above.)

Which is the narrower .. the Bayes "credible set" or the Frequentist "confidence set"?

Large-sample results:

Suppose n is large and ω^2 is a fixed positive number. Then $\sqrt{n}\,(1-\lambda_n)\to 0$, so that $\sqrt{n}\,(\hat{\mu}_{Bayes}-\overline{Y})\xrightarrow{p}0$.

Also $\sqrt{n}\left(\mu-\hat{\mu}_{Bayes}\right)=\sqrt{n}\left(\mu-\lambda_n\overline{Y}-(1-\lambda_n)m\right)\xrightarrow{p}\sqrt{n}\left(\mu-\overline{Y}\right)$ (conditional on $\overline{Y}$).

But $\sqrt{n}\left(\mu-\overline{Y}\right)\mid\overline{Y}\sim N(0,\sigma^2)$.

Large-n Posterior: $\mu\mid\overline{Y}\sim N(\overline{Y},\sigma^2/n)$

Large-n Samplings distribution: $\overline{Y}\mid\mu\sim N(\mu,\sigma^2/n)$

and

$\overline{Y}-\mu\sim N(0,\sigma^2/n)$ (conditional on μ, conditional on $\overline{Y}$, and unconditionally)
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. The proof is sketched 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 $L_n(\theta) = \ln(f(Y_{1:n}|\theta))$. The prior is $w(\theta)$.

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

Let $\gamma = \sqrt{n}(\theta - \hat{\theta}^{MLE})$, so that $\theta = \hat{\theta}^{MLE} + \gamma/\sqrt{n}$. Then

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

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

$L_n(\hat{\theta}^{MLE} + \gamma/\sqrt{n}) - L_n(\hat{\theta}^{MLE}) = \dfrac{1}{2} \dfrac{\partial^2 L_n(\tilde{\theta})}{\partial \theta^2} \dfrac{\gamma^2}{n}$, where $\tilde{\theta}$ is between $\hat{\theta}^{MLE}$ and $\hat{\theta}^{MLE} + \gamma/\sqrt{n}$. Thus $L_n(\hat{\theta}^{MLE} + \gamma/\sqrt{n}) - L_n(\hat{\theta}^{MLE}) \xrightarrow{p} -\dfrac{1}{2} \bar{I}(\theta_0) \gamma^2$.

Using these approximations

$$f(\gamma \mid Y_{1:n}) = \frac{e^{-0.5\bar{I}(\theta_0)\gamma^2}}{\int e^{-0.5\bar{I}(\theta_0)\gamma^2}\, d\gamma} \frac{w(\theta_0)}{w(\theta_0)}$$
$$= \frac{1/\sqrt{2\pi\bar{I}(\theta_0)^{-1}}}{1/\sqrt{2\pi\bar{I}(\theta_0)^{-1}}} \frac{e^{-0.5\bar{I}(\theta_0)\gamma^2}}{\int e^{-0.5\bar{I}(\theta_0)\gamma^2}\, d\gamma}$$
$$= 1/\sqrt{2\pi\bar{I}(\theta_0)^{-1}}\, e^{-0.5\bar{I}(\theta_0)\gamma^2}$$

while the final equality follows by noting that the denominator of the preceding expression is 1.

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

A key assumption in this argument is that $\hat{\theta}^{MLE}$ is consistent for the true value of θ.

Here is an example in which θ is "set-identified" and the MLE is not unique, and thus not consistent: Suppose (X_i, Y_i) are *iid N* with means μ_X and μ_Y and identity covariance matrix. It is known that $\mu_X < \mu_Y$. The parameter of interest is θ with $\mu_X \leq \theta \leq \mu_Y$. The likelihood is maximized by any value of satisfying $\bar{X} \leq \hat{\theta}_{MLE} \leq \bar{Y}$. Let $w(\theta)$ denote the prior for θ. The posterior will converge to $P(\theta \mid X_{1:n}, Y_{1:n}) \rightarrow w(\theta) / \int_{\mu_X}^{\mu_Y} w(\theta)\, d\theta$ for $\mu_X \leq \theta \leq \mu_Y$ and zero elsewhere. The Bayes credible sets will have asymptotic Frequentist coverage of 0 or 1.