STAT 801: Mathematical Statistics

Course notes

Chapter 1

Introduction

Statistics versus Probability

The standard view of scientific inference has a set of theories which make predictions about the outcomes of an experiment. In a very simple hypothetical case those predictions might be represented as in the following table:

Theory	Prediction
A	1
B	2
C	3

If we conduct the experiment and see outcome 2 we **infer** that Theory B is correct (or at least that A and C are wrong).

Now we add **Randomness**. Our table might look as follows:

Theory	Prediction
A	Usually 1 sometimes 2 never 3
B	Usually 2 sometimes 1 never 3
C	Usually 3 sometimes 1 never 2

Now if we actually see outcome 2 we infer that Theory B is probably correct, that Theory A is probably not correct and that Theory C is wrong.

Probability Theory is concerned with constructing the table just given: computing the likely outcomes of experiments.

Statistics is concerned with the inverse process of using the table to draw inferences from the outcome of the experiment. How should we do it and how wrong are our inferences likely to be?

Chapter 2

Probability

Probability Definitions

A **Probability Space** is an ordered triple $(\Omega, \mathcal{F}, P)$. The idea is that Ω is the set of possible outcomes of a random experiment, $\mathcal{F}$ is the set of those events, or subsets of Ω whose probability is defined and P is the rule for computing probabilities. Formally:

- Ω is a set.
- $\mathcal{F}$ is a family of subsets of Ω with the property that $\mathcal{F}$ is a σ-field (or Borel field or σ-algebra):
  1. The empty set $\emptyset$ and Ω are members of $\mathcal{F}$.
  2. The family $\mathcal{F}$ is closed under complementation. That is, if A is in $\mathcal{F}$ (meaning $P(A)$ is defined) then $A^c = \{\omega \in \Omega : \omega \notin A\}$ is in $\mathcal{F}$ (because we want to be able to say $P(A^c) = 1 - P(A)$).
  3. If $A_1, A_2, \cdots$ are all in $\mathcal{F}$ then so is $A = \cup_{i=1}^{\infty} A_i$. ($A$ is the event that at least one of the A_i happens and we want to be sure that if each of the A_i has a probability then so does this event A.)
- P is a function whose domain is $\mathcal{F}$ and whose range is a subset of $[0, 1]$ which satisfies the axioms for a probability:
  1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
  2. If $A_1, A_2, \cdots$ are **pairwise disjoint** (or **mutually exclusive**) (meaning for any $j \neq k$ $A_j \cap A_k = \emptyset$) then

$$P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$

     This property is called **countable additivity**.

These axioms guarantee that as we compute probabilities by the usual rules, including approximation of an event by a sequence of others we don't get caught in any logical contradictions. As we go through the course we may, from time to time, need various facts about σ-fields and probability spaces. The following theorems will give a useful reference.

Theorem 1 *Suppose $\mathcal{F}$ is a σ-field of subsets of a set Ω. Then*

1. *The family $\mathcal{F}$ is closed under countable intersections meaning that if $A_1, A_2, \ldots$ are all members of $\mathcal{F}$ then so is*
$$\cap_{i=1}^{\infty} A_i$$
2. *The family $\mathcal{F}$ is closed under finite unions and intersections. That is, if $A_1, \ldots, A_n$ are all members of $\mathcal{F}$ then so are*
$$\cap_{i=1}^{n} A_i \qquad \text{and} \cup_{i=1}^{n} A_i.$$

Proof: For part 1 apply de Morgan's laws. One of these is

$$\cap A_i = (\cup A_i^c)^c$$

The right hand side is in $\mathcal{F}$ by the properties of a σ-field applied one after another: closure under complementation, then closure under countable unions then closure under complementation again. •

Theorem 2 *Suppose that $(\Omega, \mathcal{F}, P)$ is a probability space. Then*

1. *If*

   A vector valued random variable is a function X whose domain is Ω and whose range is in some p dimensional Euclidean space, R^p with the property that the events whose probabilities we would like to calculate from their definition in terms of X are in $\mathcal{F}$. We will write $X = (X_1, \ldots, X_p)$. We will want to make sense of
   $$P(X_1 \leq x_1, \ldots, X_p \leq x_p)$$
   for any constants $(x_1, \ldots, x_p)$. In our formal framework the notation
   $$X_1 \leq x_1, \ldots, X_p \leq x_p$$
   is just shorthand for an event, that is a subset of Ω, defined as
   $$\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_p(\omega) \leq x_p\}$$
   Remember that X is a function on Ω so that X_1 is also a function on Ω. In almost all of probability and statistics the dependence of a random variable on a point in the probability space is hidden! You almost always see X not $X(\omega)$.

   Now for formal definitions:

Definition: *The* **Borel** σ*-field in* R^p *is the smallest* σ*-field in* R^p *containing every open ball. Every common set is a Borel set, that is, in the Borel* σ*-field.*

Definition: *An* R^p *valued* **random variable** *is a map* $X : \Omega \mapsto R^p$ *such that when* A *is Borel then* $\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$.

Fact: this is equivalent to

$$\{\omega \in \Omega : X_1(\omega) \leq x_1, \ldots, X_p(\omega) \leq x_p\} \in \mathcal{F}$$

for all $(x_1, \ldots, x_p) \in R^p$.

Jargon and notation: we write $P(X \in A)$ *for* $P(\{\omega \in \Omega : X(\omega) \in A)$ *and define the* **distribution** *of* X *to be the map*

$$A \mapsto P(X \in A)$$

which is a probability on the set R^p *with the Borel* σ*-field rather than the original* Ω *and* $\mathcal{F}$.

Definition: *The* **Cumulative Distribution Function** *(or CDF) of* X *is the function* F_X *on* R^p *defined by*

$$F_X(x_1, \ldots, x_p) = P(X_1 \leq x_1, \ldots, X_p \leq x_p)$$

Properties of F_X *(or just* F *when there's only one CDF under consideration):*

(a) $0 \leq F(x) \leq 1$.

(b) $x > y \Rightarrow F(x) \geq F(y)$ *(*F *is monotone non-decreasing). Note: for* $p > 1$ *the notation* $x > y$ *means* $x_i \geq y_i$ *for all* i *and* $x_i > y_i$ *for at least one* i.

(c) $\lim_{x \to -\infty} F(x) = 0$. *Note: for* $p > 1$ *this means* $\lim_{x_i \to -\infty} F(x) = 0$ *for each* i.

(d) $\lim_{x \to \infty} F(x) = 1$. *Note: for* $p > 1$ *this means* $\lim_{x_1, \ldots, x_p \to \infty} F(x) = 1$.

(e) $\lim_{x \searrow y} F(x) = F(y)$ *(*F *is right continuous).*

(f) $\lim_{x \nearrow y} F(x) \equiv F(y-)$ *exists.*

(g) For $p = 1$, $F(x) - F(x-) = P(X = x)$.

(h) $F_X(t) = F_Y(t)$ *for all* t *implies that* X *and* Y *have the same distribution, that is,* $P(X \in A) = P(Y \in A)$ *for any (Borel) set* A.

The distribution of a random variable X *is* **discrete** *(we also call the random variable discrete) if there is a countable set* $x_1, x_2, \cdots$ *such that*

$$P(X \in \{x_1, x_2 \cdots\}) = 1 = \sum_i P(X = x_i)$$

In this case the **discrete density** *or* **probability mass function** *of* X *is*

$$f_X(x) = P(X = x)$$

The distribution of a random variable X is **absolutely continuous** *if there is a function f such that*

$$P(X \in A) = \int_A f(x)dx$$

for any (Borel) set A. This is a p dimensional integral in general. This condition is equivalent (when $p = 1$) to

$$F(x) = \int_{-\infty}^{x} f(y)\, dy$$

or for general p to

$$F(x_1, \ldots, x_p) = \int_{-\infty}^{x-1} \cdots \int_{-\infty}^{x_p} f(y_1, \ldots, y_p) dy_1 \cdots dy_p$$

We call f the **density** *of X. For most values of x we then have F is differentiable at x and, for $p = 1$*

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

or in general

$$\frac{\partial^p}{\partial x_1 \cdots \partial x_p} F(x_1, \ldots, x_p) = f(x_1, \ldots, x_p)\,.$$

Example: X is Uniform[0,1].

$$F(x) = \begin{cases} 0 & x \leq 0 \\ x & 0 < x < 1 \\ 1 & x \geq 1 \end{cases}$$

$$f(x) = \begin{cases} 1 & 0 < x < 1 \\ \textit{undefined} & x \in \{0, 1\} \\ 0 & \textit{otherwise} \end{cases}$$

Example: X is exponential.

$$F(x) = \begin{cases} 1 - e^{-x} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

$$f(x) = \begin{cases} e^{-x} & x > 0 \\ \textit{undefined} & x = 0 \\ 0 & x < 0 \end{cases}$$

Chapter 3

Distribution Theory

Distribution Theory

General Problem: Start with assumptions about the density or CDF of a random vector $X = (X_1, \ldots, X_p)$. Define $Y = g(X_1, \ldots, X_p)$ to be some function of X (usually some statistic of interest). How can we compute the distribution or CDF or density of Y?

Univariate Techniques

Method 1: compute the CDF by integration and differentiate to find f_Y.

Example: $U \sim$ Uniform$[0,1]$ and $Y = -\log U$. Then

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(-\log U \le y) \\ &= P(\log U \ge -y) = P(U \ge e^{-y}) \\ &= \begin{cases} 1 - e^{-y} & y > 0 \\ 0 & y \le 0 \end{cases} \end{aligned}$$

so that Y has a standard exponential distribution.

Example: $Z \sim N(0,1)$, i.e.

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

and $Y = Z^2$. Then

$$F_Y(y) = P(Z^2 \le y) = \begin{cases} 0 & y < 0 \\ P(-\sqrt{y} \le Z \le \sqrt{y}) & y \ge 0 \end{cases}$$

Now

$$P(-\sqrt{y} \le Z \le \sqrt{y}) = F_Z(\sqrt{y}) - F_Z(-\sqrt{y})$$

can be differentiated to obtain

$$f_Y(y) = \begin{cases} 0 & y < 0 \\ \frac{d}{dy}\left[F_Z(\sqrt{y}) - F_Z(-\sqrt{y})\right] & y > 0 \\ \textit{undefined} & y = 0 \end{cases}$$

Then

$$\begin{aligned}\frac{d}{dy}F_Z(\sqrt{y}) &= f_Z(\sqrt{y})\frac{d}{dy}\sqrt{y} \\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\left(\sqrt{y}\right)^2/2\right)\frac{1}{2}y^{-1/2} \\ &= \frac{1}{2\sqrt{2\pi y}}e^{-y/2}\end{aligned}$$

with a similar formula for the other derivative. Thus

$$f_Y(y) = \begin{cases} \frac{1}{\sqrt{2\pi y}}e^{-y/2} & y > 0 \\ 0 & y < 0 \\ \textit{undefined} & y = 0 \end{cases}$$

We will find **indicator** *notation useful:*

$$1(y > 0) = \begin{cases} 1 & y > 0 \\ 0 & y \leq 0 \end{cases}$$

which we use to write

$$f_Y(y) = \frac{1}{\sqrt{2\pi y}}e^{-y/2}1(y > 0)$$

(changing the definition unimportantly at $y = 0$). One convenient convention here is to regard $0 \times$ undefined as meaning 0. That is, by convention the use of the indicator above makes $f_Y(0) = 0$.

Notice*: I never evaluated F_Y before differentiating it. In fact F_Y and F_Z are integrals I can't do but I can differentiate then anyway. You should remember the fundamental theorem of calculus:*

$$\frac{d}{dx}\int_a^x f(y)\,dy = f(x)$$

at any x where f is continuous.

So far: for $Y = g(X)$ with X and Y each real valued

$$P(Y \leq y) = P(g(X) \leq y) = P(X \in g^{-1}((-\infty, y]))$$

Take the derivative with respect to y to compute the density

$$f_Y(y) = \frac{d}{dy}\int_{\{x:g(x)\leq y\}} f(x)\,dx$$

Often we can differentiate this integral without doing the integral.

Method 2*: Change of variables.*

Now **assume g is one to one**. *I will do the case where g is increasing and I will be assuming that g is differentiable. The density has the following interpretation (mathematically what follows is just the expression of the fact that the density is the derivative of the CDF):*

$$f_Y(y) = \lim_{\delta y \to 0} \frac{P(y \leq Y \leq y + \delta y)}{\delta y} = \lim_{\delta y \to 0} \frac{F_Y(y + \delta y) - F_Y(y)}{\delta y}$$

and

$$f_X(x) = \lim_{\delta x \to 0} \frac{P(x \leq X \leq x + \delta x)}{\delta x}$$

Now assume that $y = g(x)$. Then

$$P(y \leq Y \leq g(x + \delta x)) = P(x \leq X \leq x + \delta x)$$

Each of these probabilities is the integral of a density. The first is the integral of the density of Y over the small interval from $y = g(x)$ to $y = g(x + \delta x)$. Since the interval is narrow the function f_Y is nearly constant over this interval and we get

$$P(y \leq Y \leq g(x + \delta x)) \approx f_Y(y)(g(x + \delta x) - g(x))$$

Since g has a derivative the difference

$$g(x + \delta x) - g(x) \approx \delta x g'(x)$$

and we get

$$P(y \leq Y \leq g(x + \delta x)) \approx f_Y(y) g'(x) \delta x$$

On the other hand the same idea applied to the probability expressed in terms of X gives

$$P(x \leq X \leq x + \delta x) \approx f_X(x) \delta x$$

which gives

$$f_Y(y) g'(x) \delta x \approx f_X(x) \delta x$$

or, cancelling the δx in the limit

$$f_Y(y) g'(x) = f_X(x)$$

If you remember $y = g(x)$ then you get

$$f_X(x) = f_Y(g(x)) g'(x)$$

or if you solve the equation $y = g(x)$ to get x in terms of y, that is, $x = g^{-1}(y)$ then you get the usual formula

$$f_Y(y) = f_X(g^{-1}(y)) / g'(g^{-1}(y))$$

I find it easier to remember the first of these formulas. **This is just the change of variables formula for doing integrals.**

Remark: *If g had been decreasing the derivative g' would have been negative but in the argument above the interval $(g(x), g(x+\delta x))$ would have to have been written in the other order. This would have meant that our formula had $g(x) - g(x+\delta x) \approx -g'(x)\delta x$. In both cases this amounts to the formula*

$$f_X(x) = f_Y(g(x))|g'(x)|\,.$$

Example: *$X \sim$ Weibull(shape α, scale β) or*

$$f_X(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left\{-(x/\beta)^\alpha\right\} 1(x>0)$$

Let $Y = \log X$ so that $g(x) = \log(x)$. Setting $y = \log x$ and solving gives $x = \exp(y)$ so that $g^{-1}(y) = e^y$. Then $g'(x) = 1/x$ and $1/g'(g^{-1}(y)) = 1/(1/e^y) = e^y$. Hence

$$f_Y(y) = \frac{\alpha}{\beta}\left(\frac{e^y}{\beta}\right)^{\alpha-1} \exp\left\{-(e^y/\beta)^\alpha\right\} 1(e^y>0)e^y\,.$$

The indicator is always equal to 1 since e^y is always positive. Simplifying we get

$$f_Y(y) = \frac{\alpha}{\beta^\alpha} \exp\left\{\alpha y - e^{\alpha y}/\beta^\alpha\right\}\,.$$

If we define $\phi = \log\beta$ and $\theta = 1/\alpha$ then the density can be written as

$$f_Y(y) = \frac{1}{\theta}\exp\left\{\frac{y-\phi}{\theta} - \exp\left\{\frac{y-\phi}{\theta}\right\}\right\}$$

which is called an **Extreme Value** *density with* **location** *parameter ϕ and* **scale** *parameter θ. (Note: there are several distributions going under the name Extreme Value.)*

Marginalization

Now we turn to multivariate problems. The simplest version has $X = (X_1, \ldots, X_p)$ and $Y = X_1$ (or in general any X_j).

Theorem 3 *If X has (joint) density $f(x_1, \ldots, x_p)$ then $Y = (X_1, \ldots, X_q)$ (with $q < p$) has a density f_Y given by*

$$f_{X_1,\ldots,X_q}(x_1, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, x_2, \ldots, x_p)\, dx_{q+1} \ldots dx_p$$

We call $f_{X_1,\ldots,X_q}$ the **marginal** *density of $X_1,\ldots,X_q$ and use the expression* **joint** *density for f_X but $f_{X_1,\ldots,X_q}$ is exactly the usual density of $(X_1,\ldots,X_q)$. The adjective "marginal" is just there to distinguish the object from the joint density of X.*

Example *The function*

$$f(x_1,x_2) = Kx_1x_2 1(x_1>0)1(x_2>0)1(x_1+x_2<1)$$

is a density for a suitable choice of K, namely the value of K making

$$P(X\in R^2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,x_2)\,dx_1\,dx_2 = 1\,.$$

The integral is

$$\begin{aligned} K\int_0^1\int_0^{1-x_1} x_1x_2\,dx_1\,dx_2 &= K\int_0^1 x_1(1-x_1)^2\,dx_1/2 \\ &= K(1/2-2/3+1/4)/2 \\ &= K/24 \end{aligned}$$

so that $K=24$. The marginal density of x_1 is

$$f_{X_1}(x_1) = \int_{-\infty}^{\infty} 24x_1x_2 1(x_1>0)1(x_2>0)1(x_1+x_2<1)\,dx_2$$

which is the same as

$$f_{X_1}(x_1) = 24\int_0^{1-x_1} x_1x_2 1(x_1>0)1(x_1<1)\,dx_2 = 12x_1(1-x_1)^2 1(0<x_1<1)$$

This is a Beta$(2,3)$ density.

The general multivariate problem has

$$Y = (Y_1,\ldots,Y_q) = (g_1(X_1,\ldots,X_p),\ldots,g_q(X_1,\ldots,X_p))$$

Case 1*: If $q>p$ then Y will* **not** *have a density for "smooth" g. Y will have a* **singular** *or discrete distribution. This sort of problem is rarely of real interest. (However, variables of interest* **often** *have a singular distribution – this is almost always true of the set of residuals in a regression problem.)*

Case 2 *If $q=p$ then we will be able to use a change of variables formula which generalizes the one derived above for the case $p=q=1$. (See below.)*

Case 3*: If $q<p$ we will try a two step process. In the first step we pad out Y by adding on $p-q$ more variables (carefully chosen) and calling them $Y_{q+1},\ldots,Y_p$. Formally we find functions $g_{q+1},\ldots,g_p$ and define*

$$Z = (Y_1,\ldots,Y_q, g_{q+1}(X_1,\ldots,X_p),\ldots,g_p(X_1,\ldots,X_p))$$

If we have chosen the functions carefully we will find that $g = (g_1, \ldots, g_p)$ satisfies the conditions for applying the change of variables formula from the previous case. Then we apply that case to compute f_Z. Finally we marginalize the density of Z to find that of Y:

$$f_Y(y_1, \ldots, y_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_Z(y_1, \ldots, y_q, z_{q+1}, \ldots, z_p)\, dz_{q+1} \ldots dz_p$$

Change of Variables

Suppose $Y = g(X) \in R^p$ with $X \in R^p$ having density f_X. **Assume the g is a one to one ("injective") map,** *that is, $g(x_1) = g(x_2)$ if and only if $x_1 = x_2$. Then we find f_Y as follows:*

Step 1: Solve for x in terms of y: $x = g^{-1}(y)$.

Step 2: Remember the following basic equation

$$f_Y(y)dy = f_X(x)dx$$

and rewrite it in the form

$$f_Y(y) = f_X(g^{-1}(y))\frac{dx}{dy}$$

It is now a matter of interpreting this derivative $\frac{dx}{dy}$ when $p > 1$. The interpretation is simply

$$\frac{dx}{dy} = \left| det\left(\frac{\partial x_i}{\partial y_j}\right)\right|$$

which is the so called **Jacobian** *of the transform. An equivalent formula inverts the matrix and writes*

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

This notation means

$$\left|\frac{dy}{dx}\right| = \left| det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_p} \\ & \vdots & & \\ \frac{\partial y_p}{\partial x_1} & \frac{\partial y_p}{\partial x_2} & \cdots & \frac{\partial y_p}{\partial x_p} \end{bmatrix} \right|$$

but *with x replaced by the corresponding value of y, that is, replace x by $g^{-1}(y)$.*

Example*: The density*

$$f_X(x_1, x_2) = \frac{1}{2\pi} \exp\left\{-\frac{x_1^2 + x_2^2}{2}\right\}$$

is called the **standard bivariate normal density**. *Let $Y = (Y_1, Y_2)$ where $Y_1 = \sqrt{X_1^2 + X_2^2}$ and Y_2 is the angle (between 0 and 2π) in the plane from the positive x*

axis to the ray from the origin to the point (X_1, X_2). *In other words,* Y *is* X *in polar co-ordinates.*

The first step is to solve for x *in terms of* y *which gives*

$$\begin{aligned} X_1 &= Y_1 \cos(Y_2) \\ X_2 &= Y_1 \sin(Y_2) \end{aligned}$$

so that in formulas

$$\begin{aligned} g(x_1, x_2) &= (g_1(x_1, x_2), g_2(x_1, x_2)) \\ &= (\sqrt{x_1^2 + x_2^2}, \mathit{argument}(x_1, x_2)) \\ g^{-1}(y_1, y_2) &= (g_1^{-1}(y_1, y_2), g_2^{-1}(y_1, y_2)) \\ &= (y_1 \cos(y_2), y_1 \sin(y_2)) \\ \left|\frac{dx}{dy}\right| &= \left| det \begin{pmatrix} \cos(y_2) & -y_1 \sin(y_2) \\ \sin(y_2) & y_1 \cos(y_2) \end{pmatrix} \right| \\ &= y_1 \end{aligned}$$

It follows that

$$f_Y(y_1, y_2) = \frac{1}{2\pi} \exp\left\{-\frac{y_1^2}{2}\right\} y_1 1(0 \le y_1 < \infty) 1(0 \le y_2 < 2\pi)$$

Next problem: what are the marginal densities of Y_1 *and* Y_2*? Note that* f_Y *can be factored into* $f_Y(y_1, y_2) = h_1(y_1)h_2(y_2)$ *where*

$$h_1(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)$$

and

$$h_2(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)$$

It is then easy to see that

$$f_{Y_1}(y_1) = \int_{-\infty}^{\infty} h_1(y_1)h_2(y_2)\, dy_2 = h_1(y_1) \int_{-\infty}^{\infty} h_2(y_2)\, dy_2$$

which says that the marginal density of Y_1 *must be a multiple of* h_1. *The multiplier needed will make the density integrate to 1 but in this case we can easily get*

$$\int_{-\infty}^{\infty} h_2(y_2)\, dy_2 = \int_0^{2\pi} (2\pi)^{-1} dy_2 = 1$$

so that

$$f_{Y_1}(y_1) = y_1 e^{-y_1^2/2} 1(0 \le y_1 < \infty)$$

which is special Weibull density also called a Rayleigh distribution. Similarly

$$f_{Y_2}(y_2) = 1(0 \le y_2 < 2\pi)/(2\pi)$$

which is the **Uniform***$(0, 2\pi)$ density. You should be able to check that $W = Y_1^2/2$ has a standard exponential distribution. You should also know that by definition $U = Y_1^2$ has a χ^2 distribution on 2 degrees of freedom and be able to find the χ^2_2 density.*

Note*: When a joint density factors into a product you will always see the phenomenon above — the factor involving the variable not be integrated out will come out of the integral and so the marginal density will be a multiple of the factor in question. We will see shortly that this happens when and only when the two parts of random vector are independent.*

Several-to-one Transformations

The map $g(z) = z^2$ does not have a functional inverse. Rather each $y > 0$ has two square roots $\sqrt{y}$ and $-\sqrt{y}$. In general if we have $Y = g(X)$ with both X and Y in R^p we may find that the **support** *of X (technically the smallest closed set K such that $P(X \in K) = 1$) can be split up into say $m + 1$ pieces say $S_1, \ldots, S_{m+1}$ in such a way that:*

(a) On each S_i for $i = 1, \ldots, m$ the function g is 1 to 1 with a derivative matrix

$$\frac{\partial g(x)}{\partial x}$$

which is not singular anywhere on S_i.

(b) The remaining set S_{m+1} has measure 0; that is $P(X \in S_m) = 0$

Then for each i there is a function h_i defined on $g(S_i)$ such that

$$h_i\{g(x)\} = x$$

for all $x \in S_i$ and we have the following formula for density of Y:

$$f_Y(y) = \sum_{j \le m : y \in g(S_j)} f_X\{h_j(y)\} \left| \det\left(\frac{\partial h_j(y)}{\partial y} \right) \right|$$

As an example consider the problem I did first: $X \sim N(0, 1)$ and $Y = X^2$. Now the function $g(x) = x^2$. We take $S_1 = (-\infty, 0)$, $S_2 = (0, \infty)$ and $S_3 = 0$. The sets $g(S_1)$ and $g(S_2)$ are just $(0, \infty)$ and the functions h_i are given by $h_1(y) = -\sqrt{y}$ and $h_2(y) = \sqrt{y}$. If $y > 0$ then the sum in the formula above for f_Y is over $j = 1, 2$ while for $y \le 0$ the sum is empty – meaning $f_Y(y) = 0$ for $y \le 0$. For $y > 0$ the two Jacobians are just $1/(2\sqrt{y})$ and you should now be able to reproduce the formula I gave for the χ^2_1 density.

Independence, conditional distributions

In the examples so far the density for X has been specified explicitly. In many situations, however, the process of modelling the data leads to a specification in terms of marginal and conditional distributions.

Definition*: Events A and B are independent if*

$$P(AB) = P(A)P(B)\,.$$

(Note the notation: AB is the event that both A and B happen. It is also written $A \cap B$.)

Definition*: Events A_i, $i = 1, \ldots, p$ are* **independent** *if*

$$P(A_{i_1} \cdots A_{i_r}) = \prod_{j=1}^{r} P(A_{i_j})$$

for any set of distinct indices $i_1, \ldots, i_r$ between 1 and p.

Example: $p = 3$

$$\begin{aligned} P(A_1A_2A_3) &= P(A_1)P(A_2)P(A_3) \\ P(A_1A_2) &= P(A_1)P(A_2) \\ P(A_1A_3) &= P(A_1)P(A_3) \\ P(A_2A_3) &= P(A_2)P(A_3) \end{aligned}$$

You need all these equations to be true for independence!

Example*: : Toss a coin twice. If A_1 is the event that the first toss is a Head, A_2 is the event that the second toss is a Head and A_3 is the event that the first toss and the second toss are different. then $P(A_i) = 1/2$ for each i and for $i \neq j$*

$$P(A_i \cap A_j) = \frac{1}{4}$$

but

$$P(A_1 \cap A_2 \cap A_3) = 0 \neq P(A_1)P(A_2)P(A_3)$$

Definition*: Random variables X and Y are* **independent** *if*

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

for all A and B.

Definition*: Random variables $X_1, \ldots, X_p$ are* **independent** *if*

$$P(X_1 \in A_1, \cdots, X_p \in A_p) = \prod P(X_i \in A_i)$$

for any choice of $A_1, \ldots, A_p$.

Theorem 4 *(a) If X and Y are independent then*

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for all x, y

(b) If X and Y are independent and have joint density $f_{X,Y}(x,y)$ then X and Y have densities, say f_X and f_Y, and

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

(c) If X and Y are independent and have marginal densities f_X and f_Y then (X,Y) has joint density $f_{X,Y}(x,y)$ given by

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

(d) If

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for **all** *x, y then X and Y are independent.*

(e) If (X,Y) has density $f(x,y)$ and there are functions $g(x)$ and $h(y)$ such that

$$f(x,y) = g(x)h(y)$$

for **all** *(well technically almost all) (x,y) then X and Y are independent and they each have a density given by*

$$f_X(x) = g(x)/\int_{-\infty}^{\infty} g(u)du$$

and

$$f_Y(y) = h(y)/\int_{-\infty}^{\infty} h(u)du\,.$$

Proof*:*

(a) Since X and Y are independent so are the events $X \le x$ and $Y \le y$; hence

$$P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$$

(b) For clarity suppose X and Y are real valued. In assignment 2 I have asked you to prove that the existence of $f_{X,Y}$ implies that f_X and f_Y exist (and are given by the marginal density formula). Then for any sets A and B

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B f_{X,Y}(x,y)dydx \\ P(X \in A)P(Y \in B) &= \int_A f_X(x)dx \int_B f_Y(y)dy \\ &= \int_A \int_B f_X(x)f_Y(y)dydx \end{aligned}$$

Since $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ *we see that for any sets* A *and* B

$$\int_A \int_B [f_{X,Y}(x,y) - f_X(x)f_Y(y)]dydx = 0$$

It follows (via measure theory) that the quantity in [] is 0 (for almost every pair (x,y)*).*

(c) For any A *and* B *we have*

$$\begin{aligned} P(X \in A, Y \in B) &= P(X \in A)P(Y \in B) \\ &= \int_A f_X(x)dx \int_B f_Y(y)dy \\ &= \int_A \int_B f_X(x)f_Y(y)]dydx \end{aligned}$$

If we **define** $g(x,y) = f_X(x)f_Y(y)$ *then we have proved that for* $C = A \times B$

$$P((X,Y) \in C) = \int_C g(x,y)dydx$$

To prove that g *is the joint density of* (X,Y) *we need only prove that this integral formula is valid for an arbitrary Borel set* C*, not just a rectangle* $A \times B$*. This is proved via a* monotone class *argument. You prove that the collection of sets* C *for which the identity holds has closure properties which guarantee that this collection includes the Borel sets.*

(d) This is proved via another monotone class argument.

(e)

$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B g(x)h(y)dydx \\ &= \int_A g(x)dx \int_B h(y)dy \end{aligned}$$

Take $B = R^1$ *to see that*

$$P(X \in A) = c_1 \int_A g(x)dx$$

where $c_1 = \int h(y)dy$*. From the definition of density we see that* $c_1 g$ *is the density of* X*. Since* $\int \int f_{X,Y}(xy)dxdy = 1$ *we see that* $\int g(x)dx \int h(y)dy = 1$ *so that* $c_1 = 1/\int g(x)dx$*. A similar argument works for* Y*.*

Theorem 5 *If* $X_1, \ldots, X_p$ *are independent and* $Y_i = g_i(X_i)$ *then* $Y_1, \ldots, Y_p$ *are independent. Moreover,* $(X_1, \ldots, X_q)$ *and* $(X_{q+1}, \ldots, X_p)$ *are independent.*

Conditional probability

Definition: *$P(A|B) = P(AB)/P(B)$ provided $P(B) \neq 0$.*

Definition: *For discrete random variables X and Y the conditional probability mass function of Y given X is*

$$\begin{aligned} f_{Y|X}(y|x) &= P(Y = y|X = x) \\ &= f_{X,Y}(x, y)/f_X(x) \\ &= f_{X,Y}(x, y)/\sum_t f_{X,Y}(x, t) \end{aligned}$$

For absolutely continuous X the problem is that $P(X = x) = 0$ for all x so how can we define $P(A|X = x)$ or $f_{Y|X}(y|x)$? The solution is to take a limit

$$P(A|X = x) = \lim_{\delta x \to 0} P(A|x \leq X \leq x + \delta x)$$

If, for instance, X, Y have joint density $f_{X,Y}$ then with $A = \{Y \leq y\}$ we have

$$P(A|x \leq X \leq x + \delta x) \& \frac{P(A \cap \{x \leq X \leq x + \delta x\})}{P(x \leq X \leq x + \delta x)} = \frac{\int_{-\infty}^{y} \int_{x}^{x+\delta x} f_{X,Y}(u, v) du dv}{\int_{x}^{x+\delta x} f_X(u) du}$$

Divide the top and bottom by δx and let δx tend to 0. The denominator converges to $f_X(x)$ while the numerator converges to

$$\int_{-\infty}^{y} f_{X,Y}(x, v) dv$$

So we define the conditional CDF of Y given $X = x$ to be

$$P(Y \leq y|X = x) = \frac{\int_{-\infty}^{y} f_{X,Y}(x, v) dv}{f_X(x)}$$

Differentiate with respect to y to get the definition of the conditional density of Y given $X = x$ namely

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

or in words "conditional = joint/marginal".

The Multivariate Normal Distribution

Definition: *$Z \in R^1 \sim N(0, 1)$ iff*

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

Definition: *$Z \in R^p \sim MVN(0, I)$ iff $Z = (Z_1, \ldots, Z_p)^t$ (a column vector for later use) with the Z_i independent and each $Z_i \sim N(0, 1)$.*

In this case according to our theorem

$$\begin{aligned} f_Z(z_1, \ldots, z_p) &= \prod \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} \\ &= (2\pi)^{-p/2} \exp\{-z^t z/2\} \end{aligned}$$

where the superscript t denotes matrix transpose.

Definition: *$X \in R^p$ has a multivariate normal distribution if it has the same distribution as $AZ+\mu$ for some $\mu \in R^p$, some $p \times p$ matrix of constants A and $Z \sim MVN(0, I)$.*

If the matrix A is singular then X will not have a density. If A is invertible then we can derive the multivariate normal density by the change of variables formula:

$$X = AZ + \mu \Leftrightarrow Z = A^{-1}(X - \mu)$$

$$\frac{\partial X}{\partial Z} = A \qquad \frac{\partial Z}{\partial X} = A^{-1}$$

So

$$\begin{aligned} f_X(x) &= f_Z(A^{-1}(x - \mu))|\det(A^{-1})| \\ &= \frac{(2\pi)^{-p/2} \exp\{-(x - \mu)^t (A^{-1})^t A^{-1}(x - \mu)/2\}}{|\det A|} \end{aligned}$$

Now define $\Sigma = AA^t$ and notice that

$$\Sigma^{-1} = (A^t)^{-1} A^{-1} = (A^{-1})^t A^{-1}$$

and

$$\det \Sigma = \det A \det A^t = (\det A)^2$$

Thus

$$f_X(x) = (2\pi)^{-p/2} (\det \Sigma)^{-1/2} \exp\{-(x - \mu)^t \sigma^{-1}(x - \mu)/2\}$$

which is called the $MVN(\mu, \Sigma)$ density. Notice that the density is the same for all A such that $AA^t = \Sigma$. This justifies the notation $MVN(\mu, \Sigma)$.

For which vectors μ and matrices Σ is this a density? Any μ will work but if $x \in R^p$ then

$$\begin{aligned} x^t \Sigma x &= x^t A A^t x \\ &= (A^t x)^t (A^t x) \\ &= \sum_1^p y_i^2 \\ &\geq 0 \end{aligned}$$

where $y = A^t x$. The inequality is strict except for $y = 0$ which is equivalent to $x = 0$. Thus Σ is a positive definite symmetric matrix. Conversely, if Σ is a positive definite symmetric matrix then there is a square invertible matrix A such that $AA^t = \Sigma$ so that there is a $MVN(\mu, \Sigma)$ distribution. (The matrix A can be found via the Cholesky decomposition, for example.)

More generally we say that X has a multivariate normal distribution if it has the same distribution as $AZ + \mu$ where now we remove the restriction that A be non-singular. When A is singular X will not have a density because there will exist an a such that $P(a^t X = a^t \mu) = 1$, that is, so that X is confined to a hyperplane. It is still true, however, that the distribution of X depends only on the matrix $\Sigma = AA^t$ in the sense that if $AA^t = BB^t$ then $AZ + \mu$ and $BZ + \mu$ have the same distribution.

Properties of the MVN distribution

1*: All margins are multivariate normal: if*

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and

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

then $X \sim MVN(\mu, \Sigma)$ implies that $X_1 \sim MVN(\mu_1, \Sigma_{11})$.

2*: All conditionals are normal: the conditional distribution of X_1 given $X_2 = x_2$ is $MVN(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$*

Remark*: If Σ_{22} is singular then it is still possible to make sense of these formulas:*

3*: $MX + \nu \sim MVN(M\mu + \nu, M\Sigma M^t)$. That is, an affine transformation of a Multivariate Normal is normal.*

Normal samples: Distribution Theory

Theorem *6* *Suppose $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ random variables. (That is each satisfies my definition above in 1 dimension.) Then*

(a) The sample mean $\bar{X}$ and the sample variance s^2 are independent.

(b) $n^{1/2}(\bar{X} - \mu)/\sigma \sim N(0, 1)$

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

(d) $n^{1/2}(\bar{X} - \mu)/s \sim t_{n-1}$

Proof: *Let* $Z_i = (X_i - \mu)/\sigma$. *Then* $Z_1, \ldots, Z_p$ *are independent* $N(0,1)$. *Thus* $Z = (Z_1, \ldots, Z_p)^t$ *is multivariate standard normal. Note that* $\bar{X} = \sigma\bar{Z} + \mu$ *and* $s^2 = \sum(X_i - \bar{X})^2/(n-1) = \sigma^2 \sum(Z_i - \bar{Z})^2/(n-1)$ *Thus*

$$\frac{n^{1/2}(\bar{X} - \mu)}{\sigma} = n^{1/2}\bar{Z}$$

$$\frac{(n-1)s^2}{\sigma^2} = \sum(Z_i - \bar{Z})^2$$

and

$$T = \frac{n^{1/2}(\bar{X} - \mu)}{s} = \frac{n^{1/2}\bar{Z}}{s_Z}$$

where $(n-1)s_Z^2 = \sum(Z_i - \bar{Z})^2$.

Thus we need only give a proof in the special case $\mu = 0$ *and* $\sigma = 1$.

Step 1: *Define*

$$Y = (\sqrt{n}\bar{Z}, Z_1 - \bar{Z}, \ldots, Z_{n-1} - \bar{Z})^t .$$

(This choice means Y *has the same dimension as* Z*.) Now*

$$Y = \begin{bmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ -\frac{1}{n} & 1 - \frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{bmatrix}$$

or letting M *denote the matrix*

$$Y = MZ$$

It follows that $Y \sim MVN(0, MM^t)$ *so we need to compute* MM^t*:*

$$MM^t = \left[\begin{array}{c|cccc} 1 & 0 & 0 & \cdots & 0 \\ \hline 0 & 1 - \frac{1}{n} & -\frac{1}{n} & \cdots & -\frac{1}{n} \\ \vdots & -\frac{1}{n} & \ddots & \cdots & -\frac{1}{n} \\ 0 & \vdots & \cdots & & 1 - \frac{1}{n} \end{array}\right]$$

$$= \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q \end{array}\right]$$

Now it is easy to solve for Z *from* Y *because* $Z_i = n^{-1/2}Y_1 + Y_{i+1}$ *for* $1 \le i \le n-1$. *To get* Z_n *we use the identity*

$$\sum_{i=1}^{n}(Z_i - \bar{Z}) = 0$$

to see that $Z_n = -\sum_{i=2}^{n} Y_i + n^{-1/2}Y_1$. *This proves that* M *is invertible and we find*

$$\Sigma^{-1} \equiv (MM^t)^{-1} = \left[\begin{array}{c|c} 1 & 0 \\ \hline 0 & Q^{-1} \end{array}\right]$$

Now we use the change of variables formula to compute the density of Y. It is helpful to let $\mathbf{y}_2$ denote the $n-1$ vector whose entries are $y_2, \ldots, y_n$. Note that

$$y^t \Sigma^{-1} y = y_1^2 + \mathbf{y}_2^t Q^{-1} \mathbf{y_2}$$

Then

$$\begin{aligned} f_Y(y) =& (2\pi)^{-n/2} \exp[-y^t \Sigma^{-1} y/2]/|\det M| \\ =& \frac{1}{\sqrt{2\pi}} e^{-y_1^2/2} \times \frac{(2\pi)^{-(n-1)/2} \exp[-\mathbf{y}_2^t Q^{-1} \mathbf{y_2}/2]}{|\det M|} \end{aligned}$$

Notice that this is a factorization into a function of y_1 times a function of $y_2, \ldots, y_n$. Thus $\sqrt{n}\bar{Z}$ is independent of $Z_1 - \bar{Z}, \ldots, Z_{n-1} - \bar{Z}$. Since s_Z^2 is a function of $Z_1 - \bar{Z}, \ldots, Z_{n-1} - \bar{Z}$ we see that $\sqrt{n}\bar{Z}$ and s_Z^2 are independent.

Furthermore the density of Y_1 is a multiple of the function of y_1 in the factorization above. But the factor in question is the standard normal density so $\sqrt{n}\bar{Z} N(0,1)$.

We have now done the first 2 parts of the theorem. The third part is a homework exercise but I will outline the derivation of the χ^2 density.

Suppose that $Z_1, \ldots, Z_n$ are independent $N(0,1)$. We define the χ_n^2 distribution to be that of $U = Z_1^2 + \cdots + Z_n^2$. Define angles $\theta_1, \ldots, \theta_{n-1}$ by

$$\begin{aligned} Z_1 &= U^{1/2} \cos\theta_1 \\ Z_2 &= U^{1/2} \sin\theta_1 \cos\theta_2 \\ \vdots &= \vdots \\ Z_{n-1} &= U^{1/2} \sin\theta_1 \cdots \sin\theta_{n-2} \cos\theta_{n-1} \\ Z_n &= U^{1/2} \sin\theta_1 \cdots \sin\theta_{n-1} \end{aligned}$$

(These are spherical co-ordinates in n dimensions. The θ values run from 0 to π except for the last θ whose values run from 0 to 2π.) Then note the following derivative formulas

$$\frac{\partial Z_i}{\partial U} = \frac{1}{2U} Z_i$$

and

$$\frac{\partial Z_i}{\partial \theta_j} = \begin{cases} 0 & j > i \\ -Z_i \tan\theta_i & j = i \\ Z_i \cot\theta_j & j < i \end{cases}$$

I now fix the case $n = 3$ to clarify the formulas. The matrix of partial derivatives is

$$\begin{bmatrix} U^{-1/2}\cos\theta_1/2 & -U^{1/2}\sin\theta_1 & 0 \\ U^{-1/2}\sin\theta_1\cos\theta_2/2 & U^{1/2}\cos\theta_1\cos\theta_2 & -U^{1/2}\sin\theta_1\sin\theta_2 \\ U^{-1/2}\sin\theta_1\sin\theta_2/2 & U^{1/2}cos\theta_1\sin\theta_2 & U^{1/2}\sin\theta_1\cos\theta_2 \end{bmatrix}$$

The determinant of this matrix may be found by adding $2U^{1/2}\cos\theta_j/\sin\theta_j$ times column 1 to column $j+1$ (which doesn't change the determinant). The resulting matrix is

lower triangular with diagonal entries (after a small amount of algebra) $U^{-1/2}\cos\theta_1/2$, $U^{1/2}\cos\theta_2/\cos\theta_1$ *and* $U^{1/2}\sin\theta_1/\cos\theta_2$. *We multiply these together to get*

$$U^{1/2}\sin(\theta_1)/2$$

which is non-negative for all U *and* θ_1. *For general* n *we see that every term in the first column contains a factor* $U^{-1/2}/2$ *while every other entry has a factor* $U^{1/2}$. *Multiplying a column in a matrix by* c *multiplies the determinant by* c *so the Jacobian of the transformation is* $u^{(n-1)/2}u^{-1/2}/2$ *times some function, say* h, *which depends only on the angles. Thus the joint density of* $U, \theta_1, \ldots \theta_{n-1}$ *is*

$$(2\pi)^{-n/2}\exp(-u/2)u^{(n-2)/2}h(\theta_1, \cdots, \theta_{n-1})/2$$

To compute the density of U *we must do an* $n-1$ *dimensional multiple integral* $d\theta_{n-1}\cdots d\theta_1$. *We see that the answer has the form*

$$cu^{(n-2)/2}\exp(-u/2)$$

for some c *which we can evaluate by making*

$$\int_{-\infty}^{\infty} f_U(u)du = c\int_0^{\infty} u^{(n-2)/2}\exp(-u/2)du = 1$$

Substitute $y = u/2$, $du = 2dy$ *to see that*

$$c2^{(n-2)/2}2\int_0^{\infty} y^{(n-2)/2}e^{-y}dy = c2^{(n-1)/2}\Gamma(n/2) = 1$$

so that the χ^2 *density is*

$$\frac{1}{2\Gamma(n/2)}\left(\frac{u}{2}\right)^{(n-2)/2}e^{-u/2}$$

Finally the fourth part of the theorem is a consequence of the first 3 parts of the theorem and the definition of the t_ν *distribution, namely, that* $T \sim t_\nu$ *if it has the same distribution as*

$$Z/\sqrt{U/\nu}$$

where $Z \sim N(0,1)$, $U \sim \chi^2_\nu$ *and* Z *and* U *are independent.*

However, I now derive the density of T *in this definition:*

$$\begin{aligned} P(T \le t) &= P(Z \le t\sqrt{U/\nu}) \\ &= \int_0^{\infty}\int_{-\infty}^{t\sqrt{u/\nu}} f_Z(z)f_U(u)dzdu \end{aligned}$$

I can differentiate this with respect to t *by simply differentiating the inner integral:*

$$\frac{\partial}{\partial t}\int_{at}^{bt} f(x)dx = bf(bt) - af(at)$$

by the fundamental theorem of calculus. Hence

$$\frac{d}{dt}P(T \le t) = \int_0^\infty f_U(u)\sqrt{u/\nu}\frac{\exp[-t^2u/(2\nu)]}{\sqrt{2\pi}}du\,.$$

Now I plug in

$$f_U(u) = \frac{1}{2\Gamma(\nu/2)}(u/2)^{(\nu-2)/2}e^{-u/2}$$

to get

$$f_T(t) = \int_0^\infty \frac{1}{2\sqrt{\pi\nu}\Gamma(\nu/2)}(u/2)^{(\nu-1)/2}\exp[-u(1+t^2/\nu)/2]\,du\,.$$

Make the substitution $y = u(1+t^2/\nu)/2$, $dy = (1+t^2/\nu)du/2$ $(u/2)^{(\nu-1)/2} = [y/(1+t^2/\nu)]^{(\nu-1)/2}$ *to get*

$$f_T(t) = \frac{1}{\sqrt{\pi\nu}\Gamma(\nu/2)}(1+t^2/\nu)^{-(\nu+1)/2}\int_0^\infty y^{(\nu-1)/2}e^{-y}dy$$

or

$$f_T(t) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\pi\nu}\Gamma(\nu/2)}\frac{1}{(1+t^2/\nu)^{(\nu+1)/2}}$$

Expectation, moments

In elementary courses we give two definitions of expected values:

Definition *If X has density f then*

$$E(g(X)) = \int g(x)f(x)\,dx\,.$$

Definition*: If X has discrete density f then*

$$E(g(X)) = \sum_x g(x)f(x)\,.$$

Now if $Y = g(X)$ for smooth g then

$$E(Y) = \int yf_Y(y) = \int g(x)f_Y(g(x))g'(x)\,dy = E(g(X))$$

by the change of variables formula for integration. This is good because otherwise we might have two different values for $E(e^X)$.

In general, there are random variables which are neither absolutely continuous nor discrete. Here's how probabilists define E in general.

Definition*: A random variable X is simple if we can write*

$$X(\omega) = \sum_1^n a_i 1(\omega \in A_i)$$

for some constants $a_1, \ldots, a_n$ and events A_i.

Def'n*: For a simple rv X we define*

$$E(X) = \sum a_i P(A_i)$$

For positive random variables which are not simple we extend our definition by approximation:

Def'n*: If $X \geq 0$ then*

$$E(X) = \sup\{E(Y) : 0 \leq Y \leq X, Y \text{ simple}\}$$

Def'n*: We call X* **integrable** *if*

$$E(|X|) < \infty\,.$$

In this case we define

$$E(X) = E(\max(X, 0)) - E(\max(-X, 0))$$

Facts: E is a linear, monotone, positive operator:

(a) **Linear***: $E(aX + bY) = aE(X) + bE(Y)$ provided X and Y are integrable.*

(b) **Positive***: $P(X \geq 0) = 1$ implies $E(X) \geq 0$.*

(c) **Monotone***: $P(X \geq Y) = 1$ and X, Y integrable implies $E(X) \geq E(Y)$.*

Major technical theorems:

Monotone Convergence*: If $0 \leq X_1 \leq X_2 \leq \cdots$ and $X = \lim X_n$ (which has to exist) then*

$$E(X) = \lim_{n\to\infty} E(X_n)$$

Dominated Convergence*: If $|X_n| \leq Y_n$ and there is a random variable X such that $X_n \to X$ (technical details of this convergence later in the course) and a random variable Y such that $Y_n \to Y$ with $E(Y_n) \to E(Y) < \infty$ then*

$$E(X_n) \to E(X)$$

This is often used with all Y_n the same random variable Y.

Fatou's Lemma*: If $X_n \geq 0$ then*

$$E(\limsup X_n) \leq \limsup E(X_n)$$

Theorem*: With this definition of E if X has density $f(x)$ (even in R^p say) and $Y = g(X)$ then*

$$E(Y) = \int g(x) f(x) dx\,.$$

(This could be a multiple integral.) If X has pmf f then

$$E(Y) = \sum_x g(x) f(x) \, .$$

This works for instance even if X has a density but Y doesn't.

Def'n*: The r^{th} moment (about the origin) of a real random variable X is $\mu'_r = E(X^r)$ (provided it exists). We generally use μ for $E(X)$. The r^{th} central moment is*

$$\mu_r = E[(X - \mu)^r]$$

We call $\sigma^2 = \mu_2$ the variance.

Def'n*: For an R^p valued random vector X we define $\mu_X = E(X)$ to be the vector whose i^{th} entry is $E(X_i)$ (provided all entries exist).*

Def'n*: The $(p \times p)$ variance covariance matrix of X is*

$$Var(X) = E\left[(X - \mu)(X - \mu)^t\right]$$

which exists provided each component X_i has a finite second moment. More generally if $X \in R^p$ and $Y \in R^q$ both have all components with finite second moments then

$$Cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)^T\right]$$

We have

$$Cov(AX + b, CY + d) = ACov(X, Y)B^T$$

for general (conforming) matrices A, C and vectors b and d.

Moments and probabilities of rare events are closely connected as will be seen in a number of important probability theorems. Here is one version of Markov's inequality (one case is Chebyshev's inequality):

$$\begin{aligned} P(|X - \mu| \geq t) &= E[1(|X - \mu| \geq t)] \\ &\leq E\left[\frac{|X - \mu|^r}{t^r} 1(|X - \mu| \geq t)\right] \\ &\leq \frac{E[|X - \mu|^r]}{t^r} \end{aligned}$$

The intuition is that if moments are small then large deviations from average are unlikely.

Example moments*: If Z is standard normal then*

$$\begin{aligned} E(Z) &= \int_{-\infty}^{\infty} z e^{-z^2/2} dz / \sqrt{2\pi} \\ &= \left. \frac{-e^{-z^2/2}}{\sqrt{2\pi}} \right|_{-\infty}^{\infty} \\ &= 0 \end{aligned}$$

and (integrating by parts)

$$\begin{aligned} E(Z^r) &= \int_{-\infty}^{\infty} z^r e^{-z^2/2} dz/\sqrt{2\pi} \\ &= \left. \frac{-z^{r-1}e^{-z^2/2}}{\sqrt{2\pi}} \right|_{-\infty}^{\infty} + (r-1)\int_{-\infty}^{\infty} z^{r-2} e^{-z^2/2} dz/\sqrt{2\pi} \end{aligned}$$

so that

$$\mu_r = (r-1)\mu_{r-2}$$

for $r \geq 2$. *Remembering that* $\mu_1 = 0$ *and*

$$\mu_0 = \int_{-\infty}^{\infty} z^0 e^{-z^2/2} dz/\sqrt{2\pi} = 1$$

we find that

$$\mu_r = \begin{cases} 0 & r \text{ odd} \\ (r-1)(r-3)\cdots 1 & r \text{ even} \end{cases}$$

If now $X \sim N(\mu, \sigma^2)$, *that is,* $X \sim \sigma Z + \mu$, *then* $E(X) = \sigma E(Z) + \mu = \mu$ *and*

$$\mu_r(X) = E[(X-\mu)^r] = \sigma^r E(Z^r)$$

In particular, we see that our choice of notation $N(\mu, \sigma^2)$ *for the distribution of* $\sigma Z + \mu$ *is justified;* σ *is indeed the variance.*

Moments and independence

Theorem*: If* $X_1, \ldots, X_p$ *are independent and each* X_i *is integrable then* $X = X_1 \cdots X_p$ *is integrable and*

$$E(X_1 \cdots X_p) = E(X_1) \cdots E(X_p)$$

Proof*: Suppose each* X_i *is simple:* $X_i = \sum_j x_{ij} 1(X_i = x_{ij})$ *where the* x_{ij} *are the possible values of* X_i. *Then*

$$\begin{aligned} E(X_1 \cdots X_p) &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} E(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p}) \\ &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\ &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\ &= \left[\sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1}) \right] \cdots \left[\sum_{j_p} x_{pj_p} P(X_p = x_{pj_p}) \right] \\ &= \prod E(X_i) \end{aligned}$$

For general $X_i > 0$ we create a sequence of simple approximations by rounding X_i down to the nearest multiple of 2^{-n} (to a maximum of n) and applying the case just done and the monotone convergence theorem. The general case uses the fact that we can write each X_i as the difference of its positive and negative parts:

$$X_i = \max(X_i, 0) - \max(-X_i, 0)$$

Moment Generating Functions

Def'n: *The moment generating function of a real valued X is*

$$M_X(t) = E(e^{tX})$$

defined for those real t for which the expected value is finite.

Def'n: *The moment generating function of $X \in R^p$ is*

$$M_X(u) = E[\exp u^t X]$$

defined for those vectors u for which the expected value is finite.

The moment generating function has the following formal connection to moments:

$$M_X(t) = \sum_{k=0}^{\infty} E[(tX)^k]/k!$$

$$= \sum_{k=0}^{\infty} \mu'_k t^k/k!$$

It is thus sometimes possible to find the power series expansion of M_X and read off the moments of X from the coefficients of the powers $t^k/k!$. (An analogous multivariate version is available.)

Theorem: *If M is finite for all $t \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$ then*

(a) Every moment of X is finite.

(b) M is C^∞ (in fact M is analytic).

(c) $\mu'_k = \frac{d^k}{dt^k} M_X(0)$.

The proof, and many other facts about moment generating functions, rely on techniques of complex variables.

Moment Generating Functions and Sums

If $X_1, \ldots, X_p$ are independent and $Y = \sum X_i$ then the moment generating function of Y is the product of those of the individual X_i:

$$E(e^{tY}) = \prod_i E(e^{tX_i})$$

or $M_Y = \prod M_{X_i}$.

However this formula makes the power series expansion of M_Y *not a particularly nice function of the expansions of the individual* M_{X_i}. *In fact this is related to the following observation. The first 3 moments (meaning* μ, σ^2 *and* μ_3*) of* Y *are just the sums of those of the* X_i *but this doesn't work for the fourth or higher moment.*

$$E(Y) = \sum E(X_i)$$
$$\mathrm{Var}(Y) = \sum \mathrm{Var}(X_i)$$
$$E[(Y - E(Y))^3] = \sum E[(X_i - E(X_i))^3]$$

but

$$E[(Y - E(Y))^4] = \sum \{E[(X_i - E(X_i))^4] - E^2[(X_i - E(X_i))^2]\} + \left\{\sum E[(X_i - E(X_i))^2]\right\}^2$$

It is possible, however, to replace the moments by other objects called **cumulants** *which do add up properly. The way to define them relies on the observation that the log of the moment generating function of* Y *is the sum of the logs of the moment generating functions of the* X_i. *We define the cumulant generating function of a variable* X *by*

$$K_X(t) = \log(M_X(t))$$

Then

$$K_Y(t) = \sum K_{X_i}(t)$$

The moment generating functions are all positive so that the cumulative generating functions are defined wherever the moment generating functions are. This means we can give a power series expansion of K_Y*:*

$$K_Y(t) = \sum_{r=1}^{\infty} \kappa_r t^r / r!$$

We call the κ_r *the cumulants of* Y *and observe*

$$\kappa_r(Y) = \sum \kappa_r(X_i)$$

To see the relation between cumulants and moments proceed as follows: the cumulant generating function is

$$\begin{aligned} K(t) &= \log(M(t)) \\ &= \log(1 + [\mu_1 t + \mu_2' t^2/2 + \mu_3' t^3/3! + \cdots]) \end{aligned}$$

To compute the power series expansion we think of the quantity in $[\ldots]$ *as* x *and expand*

$$\log(1 + x) = x - x^2/2 + x^3/3 - x^4/4 \cdots .$$

When you stick in the power series

$$x = \mu t + \mu_2' t^2/2 + \mu_3' t^3/3! + \cdots$$

you have to expand out the powers of x and collect together like terms. For instance,

$$\begin{aligned} x^2 &= \mu^2 t^2 + \mu\mu_2' t^3 + [2\mu_3'\mu/3! + (\mu_2')^2/4]t^4 + \cdots \\ x^3 &= \mu^3 t^3 + 3\mu_2'\mu^2 t^4/2 + \cdots \\ x^4 &= \mu^4 t^4 + \cdots \end{aligned}$$

Now gather up the terms. The power t^1 occurs only in x with coefficient μ. The power t^2 occurs in x and in x^2 and so on. Putting these together gives

$$\begin{aligned} K(t) = \mu t &\\ &+ [\mu_2' - \mu^2]t^2/2 \\ &+ [\mu_3' - 3\mu\mu_2' + 2\mu^3]t^3/3! \\ &+ [\mu_4' - 4\mu_3'\mu - 3(\mu_2')^2 + 12\mu_2'\mu^2 - 6\mu^4]t^4/4! + \cdots \end{aligned}$$

Comparing coefficients of $t^r/r!$ we see that

$$\begin{aligned} \kappa_1 &= \mu \\ \kappa_2 &= \mu_2' - \mu^2 = \sigma^2 \\ \kappa_3 &= \mu_3' - 3\mu\mu_2' + 2\mu^3 = E[(X-\mu)^3] \\ \kappa_4 &= \mu_4' - 4\mu_3'\mu - 3(\mu_2')^2 + 12\mu_2'\mu^2 - 6\mu^4 \\ &= E[(X-\mu)^4] - 3\sigma^4 \end{aligned}$$

Check the book by Kendall and Stuart (or the new version called Kendall's Theory of Statistics *by Stuart and Ord) for formulas for larger orders r.*

Example*: If $X_1, \ldots, X_p$ are independent and X_i has a $N(\mu_i, \sigma_i^2)$ distribution then*

$$\begin{aligned} M_{X_i}(t) &= \int_{-\infty}^{\infty} e^{tx} e^{-(x-\mu_i)/\sigma_i^2} dx/(\sqrt{2\pi}\sigma_i) \\ &= \int_{-\infty}^{\infty} e^{t(\sigma_i z + \mu_i)} e^{-z^2/2} dz/\sqrt{2\pi} \\ &= e^{t\mu_i} \int_{-\infty}^{\infty} e^{-(z-t\sigma_i)^2/2 + t^2\sigma_i^2/2} dz/\sqrt{2\pi} \\ &= e^{\sigma_i^2 t^2/2 + t\mu_i} \end{aligned}$$

This makes the cumulant generating function

$$K_{X_i}(t) = \log(M_{X_i}(t)) = \sigma_i^2 t^2/2 + \mu_i t$$

and the cumulants are $\kappa_1 = \mu_i$, $\kappa_2 = \sigma_i^2$ and every other cumulant is 0. The cumulant generating function for $Y = \sum X_i$ is

$$K_Y(t) = \sum \sigma_i^2 t^2/2 + t \sum \mu_i$$

which is the cumulant generating function of $N(\sum \mu_i, \sum \sigma_i^2)$.

Example*: I am having you derive the moment and cumulant generating function and all the moments of a Gamma rv. Suppose that* $Z_1, \ldots, Z_\nu$ *are independent* $N(0,1)$ *rvs. Then we have defined* $S_\nu = \sum_1^\nu Z_i^2$ *to have a* χ^2 *distribution. It is easy to check* $S_1 = Z_1^2$ *has density*

$$(u/2)^{-1/2} e^{-u/2}/(2\sqrt{\pi})$$

and then the mgf of S_1 *is*

$$(1-2t)^{-1/2}$$

It follows that

$$M_{S_\nu}(t) = (1-2t)^{-\nu/2};$$

you will show in homework that this is the mgf of a Gamma$(\nu/2, 2)$ *rv. This shows that the* χ^2_ν *distribution has the Gamma*$(\nu/2, 2)$ *density which is*

$$(u/2)^{(\nu-2)/2} e^{-u/2}/(2\Gamma(\nu/2))\,.$$

Example*: The Cauchy density is*

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

and the corresponding moment generating function is

$$M(t) = \int_{-\infty}^{\infty} \frac{e^{tx}}{\pi(1+x^2)} dx$$

which is $+\infty$ *except for* $t = 0$ *where we get 1. This mgf is exactly the mgf of* every t *distribution so it is not much use for distinguishing such distributions. The problem is that these distributions do not have infinitely many finite moments.*

This observation has led to the development of a substitute for the mgf which is defined for every distribution, namely, the characteristic function.

Characteristic Functions

Definition*: The characteristic function of a real rv* X *is*

$$\phi_X(t) = E(e^{itX})$$

where $i = \sqrt{-1}$ *is the imaginary unit.*

Aside on complex arithmetic.

The complex numbers are the things you get if you add $i = \sqrt{-1}$ *to the real numbers and require that all the usual rules of algebra work. In particular if* i *and any real numbers* a *and* b *are to be complex numbers then so must be* $a + bi$*. If we multiply a*

complex number $a+bi$ with a and b real by another such number, say $c+di$ then the usual rules of arithmetic (associative, commutative and distributive laws) require

$$\begin{aligned}(a+bi)(c+di) =& ac+adi+bci+bdi^2\\ =& ac+bd(-1)+(ad+bc)i\\ =& (ac-bd)+(ad+bc)i\end{aligned}$$

so this is precisely how we define multiplication. Addition is simply (again by following the usual rules)

$$(a+bi)+(c+di)=(a+b)+(c+d)i$$

Notice that the usual rules of arithmetic then don't require any more numbers than things of the form

$$x+yi$$

where x and y are real. We can identify a single such number $x+yi$ with the corresponding point (x,y) in the plane. It often helps to picture the complex numbers as forming a plane.

Now look at transcendental functions. For real x we know $e^x=\sum x^k/k!$ so our insistence on the usual rules working means

$$e^{x+iy}=e^x e^{iy}$$

and we need to know how to compute e^{iy}. Remember in what follows that $i^2=-1$ so $i^3=-i$, $i^4=1$ $i^5=i^1=i$ and so on. Then

$$\begin{aligned}e^{iy} =& \sum_0^\infty \frac{(iy)^k}{k!}\\ =& 1+iy+(iy)^2/2+(iy)^3/6+\cdots\\ =& 1-y^2/2+y^4/4!-y^6+\cdots\\ & +iy-iy^3/3!+iy^5/5!+\cdots\\ =& \cos(y)+i\sin(y)\end{aligned}$$

We can thus write

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

Now every point in the plane can be written in polar co-ordinates as $(r\cos\theta, r\sin\theta)$ and comparing this with our formula for the exponential we see we can write

$$x+iy=\sqrt{x^2+y^2}e^{i\theta}$$

for an angle $\theta\in[0,2\pi)$.

We will need from time to time a couple of other definitions:

Definition*: The* **modulus** *of the complex number $x+iy$ is*

$$|x+iy|=\sqrt{x^2+y^2}$$

Definition: *The* **complex conjugate** *of* $x + iy$ *is* $\overline{x + iy} = x - iy$.

Notes on calculus with complex variables. Essentially the usual rules apply so, for example,

$$\frac{d}{dt}e^{it} = ie^{it}$$

We will (mostly) be doing only integrals over the real line; the theory of integrals along paths in the complex plane is a very important part of mathematics, however.

End of Aside

Since

$$e^{itX} = \cos(tX) + i\sin(tX)$$

we find that

$$\phi_X(t) = E(\cos(tX)) + iE(\sin(tX))$$

Since the trigonometric functions are bounded by 1 the expected values must be finite for all t *and this is precisely the reason for using characteristic rather than moment generating functions in probability theory courses.*

Theorem 7 *For any two real rvs* X *and* Y *the following are equivalent:*

(a) X *and* Y *have the same distribution, that is, for any (Borel) set* A *we have*

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

(b) $F_X(t) = F_Y(t)$ *for all* t.

(c) $\phi_X = E(e^{itX}) = E(e^{itY}) = \phi_Y(t)$ *for all real* t.

Moreover, all of these are implied if there is a positive ϵ *such that for all* $|t| \leq \epsilon$

$$M_X(t) = M_Y(t) < \infty\,.$$

Inversion

The previous theorem is a non-constructive characterization. It does not show us how to get from ϕ_X *to* F_X *or* f_X. *For CDFs or densities with reasonable properties, however, there are effective ways to compute* F *or* f *from* ϕ. *In homework I am asking you to prove the following basic* **inversion** *formula:*

If X *is a random variable taking only integer values then for each integer* k

$$\begin{aligned} P(X = k) &= \frac{1}{2\pi}\int_0^{2\pi} \phi_X(t)e^{-itk}dt \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_X(t)e^{-itk}dt\,. \end{aligned}$$

The proof proceeds from the formula

$$\phi_X(t) = \sum_k e^{ikt} P(X = k)\,.$$

Now suppose that X has a continuous bounded density f. Define

$$X_n = [nX]/n$$

where $[a]$ denotes the integer part (rounding down to the next smallest integer). We have

$$\begin{aligned} P(k/n \le X < (k+1)/n) &= P([nX] = k) \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \phi_{[nX]}(t) e^{-itk} dt\,. \end{aligned}$$

Make the substitution $t = u/n$, and get

$$nP(k/n \le X < (k+1)/n) = \frac{1}{2\pi}\int_{-n\pi}^{n\pi} \phi_{[nX]}(u/n) e^{iuk/n} du$$

Now, as $n \to \infty$ we have

$$\phi_{[nX]}(u/n) = E(e^{iu[nX]/n}) \to E(e^{iuX})$$

(by the dominated convergence theorem – the dominating random variable is just the constant 1). The range of integration converges to the whole real line and if $k/n \to x$ we see that the left hand side converges to the density $f(x)$ while the right hand side converges to

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_X(u) e^{-iux} du$$

which gives the inversion formula

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_X(u) e^{-iux} du$$

Many other such formulas are available to compute things like $F(b) - F(a)$ and so on.

All such formulas are sometimes referred to as Fourier inversion formulas; the characteristic function itself is sometimes called the Fourier transform of the distribution or CDF or density of X.

Inversion of the Moment Generating Function

The moment generating function and the characteristic function are related formally by

$$M_X(it) = \phi_X(t)$$

When M_X exists this relationship is not merely formal; the methods of complex variables mean there is a "nice" (analytic) function which is $E(e^{zX})$ for any complex $z = x + iy$ for which $M_X(x)$ is finite. All this means that there is an inversion formula for M_X. This formula requires a complex contour integral. *In general if z_1 and z_2 are two points in the complex plane and C a path between these two points we can define the path integral*

$$\int_C f(z)dz$$

by the methods of line integration. When it comes to doing algebra with such integrals the usual theorems of calculus still work. The Fourier inversion formula was

$$2\pi f(x) = \int_{-\infty}^{\infty} \phi(t)e^{-itx}dt$$

so replacing ϕ by M we get

$$2\pi f(x) = \int_{-\infty}^{\infty} M(it)e^{-itx}dt$$

If we just substitute $z = it$ then we find

$$2\pi i f(x) = \int_C M(z)e^{-zx}dz$$

where the path C is the imaginary axis. This formula becomes of use by the methods of complex integration which permit us to replace the path C by any other path which starts and ends at the same place. It is possible, in some cases, to choose this path to make it easy to do the integral approximately; this is what saddlepoint approximations are. This inversion formula is called the inverse Laplace transform; the mgf is also called the Laplace transform of the distribution or of the CDF or of the density.

Applications of Inversion

1)*: Numerical calculations*

Example: Many statistics have a distribution which is approximately that of

$$T = \sum \lambda_j Z_j^2$$

where the Z_j are iid $N(0,1)$. In this case

$$\begin{aligned} E(e^{itT}) &= \prod E(e^{it\lambda_j Z_j^2}) \\ &= \prod (1 - 2it\lambda_j)^{-1/2}\,. \end{aligned}$$

*Imhof (*Biometrika*, 1961) gives a simplification of the Fourier inversion formula for*

$$F_T(x) - F_T(0)$$

which can be evaluated numerically.

2)*: The central limit theorem (in some versions) can be deduced from the Fourier inversion formula: if $X_1, \ldots, X_n$ are iid with mean 0 and variance 1 and $T = n^{1/2}\bar{X}$ then with ϕ denoting the characteristic function of a single X we have*

$$\begin{aligned} E(e^{itT} &= E(e^{in^{-1/2}t\sum X_j}) \\ &= \left[\phi(n^{-1/2}t)\right]^n \\ &\approx \left[\phi(0) + n^{-1/2}t\phi'(0) + n^{-1}t^2\phi''(0)/2 + o(n^{-1}\right]^n \end{aligned}$$

But now $\phi(0) = 1$ and

$$\phi'(t) = \frac{d}{dt}E(e^{itX_1}) = iE(X_1 e^{itX_1})$$

So $\phi'(0) = E(X_1) = 0$. Similarly

$$\phi''(t) = i^2 E(X_1^2 e^{itX_1})$$

so that

$$\phi''(0) = -E(X_1^2) = -1$$

It now follows that

$$\begin{aligned} E(e^{itT}) &\approx [1 - t^2/(2n) + o(1/n)]^n \\ &\to e^{-t^2/2} \end{aligned}$$

With care we can then apply the Fourier inversion formula and get

$$\begin{aligned} f_T(x) &= \frac{1}{2\pi i}\int_{-\infty}^{\infty} e^{-itx}[\phi(tn^{-1/2})]^n dt \\ &\to \frac{1}{2\pi i}\int_{-\infty}^{\infty} e^{-itx}e^{-t^2/2}dt \\ &= \frac{1}{\sqrt{2\pi}}\phi_Z(-x) \end{aligned}$$

where ϕ_Z is the characteristic function of a standard normal variable Z. Doing the integral we find

$$\phi_Z(x) = \phi_Z(-x) = e^{-x^2/2}$$

so that

$$f_T(x) \to \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

which is a standard normal random variable.

This proof of the central limit theorem is not terribly general since it requires T to have a bounded continuous density. The central limit theorem itself is a statement about CDFs not densities and is

$$P(T \le t) \to P(Z \le t)$$

Last time derived the Fourier inversion formula

$$f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \phi_X(u)e^{-iux}du$$

and the moment generating function inversion formula

$$2\pi i f(x) = \int_{-i\infty}^{i\infty} M(z)e^{-zx}dz$$

(where the limits of integration indicate a contour integral up the imaginary axis.) The methods of complex variables permit this path to be replaced by any contour running up a line like $Re(z) = c$. ($Re(Z)$ denotes the real part of z, that is, x when $z = x + iy$ with x and y real.) The value of c has to be one for which $M(c) < \infty$. Rewrite the inversion formula using the cumulant generating function $K(t) = \log(M(t))$ to get

$$2\pi i f(x) = \int_{c-i\infty}^{c+i\infty} \exp(K(z) - zx)dz\,.$$

Along the contour in question we have $z = c + iy$ so we can think of the integral as being

$$i\int_{-\infty}^{\infty} \exp(K(c+iy) - (c+iy)x)dy$$

Now do a Taylor expansion of the exponent:

$$K(c+iy) - (c+iy)x = K(c) - cx + iy(K'(c) - x) - y^2K''(c)/2 + \cdots$$

Ignore the higher order terms and select a c so that the first derivative

$$K'(c) - x$$

vanishes. Such a c is a saddlepoint. We get the formula

$$2\pi f(x) \approx \exp(K(c) - cx)\int_{-\infty}^{\infty} \exp(-y^2K''(c)/2)dy$$

The integral is just a normal density calculation and gives $\sqrt{2\pi/K''(c)}$. The saddlepoint approximation is

$$f(x) = \frac{\exp(K(c) - cx)}{\sqrt{2\pi K''(c)}}$$

Essentially the same idea lies at the heart of the proof of Sterling's approximation to the factorial function:

$$n! = \int_0^{\infty} \exp(n\log(x) - x)dx$$

The exponent is maximized when $x = n$. For n large we approximate $f(x) = n\log(x) - x$ by

$$f(x) \approx f(x_0) + (x - x_0)f'(x_0) + (x - x_0)^2 f''(x_0)/2$$

and choose $x_0 = n$ to make $f'(x_0) = 0$. Then

$$n! \approx \int_0^\infty \exp[n\log(n) - n - (x-n)^2/(2n)]dx$$

Substitute $y = (x-n)/\sqrt{n}$ to get the approximation

$$n! \approx n^{1/2} n^n e^{-n} \int_{-\infty}^{\infty} e^{-y^2/2} dy$$

or

$$n! \approx \sqrt{2\pi} n^{n+1/2} e^{-n}$$

This tactic is called Laplace's method. Notice that I am very sloppy about the limits of integration. To make the foregoing rigorous you must show that the contribution to the integral from x not sufficiently close to n is negligible.

Applications of Inversion

1)*: Numerical calculations*

Example: Many statistics have a distribution which is approximately that of

$$T = \sum \lambda_j Z_j^2$$

where the Z_j are iid $N(0,1)$. In this case

$$\begin{aligned} E(e^{itT}) &= \prod E(e^{it\lambda_j Z_j^2}) \\ &= \prod (1 - 2it\lambda_j)^{-1/2} \,. \end{aligned}$$

*Imhof (*Biometrika*, 1961) gives a simplification of the Fourier inversion formula for*

$$F_T(x) - F_T(0)$$

which can be evaluated numerically.

Here is how it works:

$$\begin{aligned} F_T(x) - F_T(0) &= \int_0^x f_T(y)dy \\ &= \int_0^x \frac{1}{2\pi} \int_{-\infty}^{\infty} \prod (1 - 2it\lambda_j)^{-1/2} e^{-ity} dt dy \end{aligned}$$

Multiply

$$\phi(t) = \left[\frac{1}{\prod (1 - 2it\lambda_j)} \right]^{1/2}$$

top and bottom by the complex conjugate of the denominator:

$$\phi(t) = \left[\frac{\prod(1+2it\lambda_j)}{\prod(1+4t^2\lambda_j^2)}\right]^{1/2}$$

The complex number $1+2it\lambda_j$ *is* $r_j e^{i\theta_j}$ *where* $r_j = \sqrt{1+4t^4\lambda_j^2}$ *and* $\tan(\theta_j) = 2t\lambda_j$ *This allows us to rewrite*

$$\phi(t) = \left[\frac{(\prod r_j)\, e^{i\sum\theta_j}}{\prod r_j^2}\right]^{1/2}$$

or

$$\phi(t) = \frac{e^{i\sum \tan^{-1}(2t\lambda_j)/2}}{\prod(1+4t^2\lambda_j^2)^{1/4}}$$

Assemble this to give

$$F_T(x) - F_T(0) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{e^{i\theta(t)}}{\rho(t)}\int_0^x e^{-iyt}dydt$$

where $\theta(t) = \sum \tan^{-1}(2t\lambda_j)/2$ *and* $\rho(t) = \prod(1+4t^2\lambda_j^2)^{1/4}$*. But*

$$\int_0^x e^{-iyt}dy = \frac{e^{-ixt}-1}{-it}$$

We can now collect up the real part of the resulting integral to derive the formula given by Imhof. I don't produce the details here.

2)*: The central limit theorem (the version called "local") can be deduced from the Fourier inversion formula: if* $X_1, \ldots, X_n$ *are iid with mean 0 and variance 1 and* $T = n^{1/2}\bar{X}$ *then with* ϕ *denoting the characteristic function of a single* X *we have*

$$\begin{aligned} E(e^{itT}) &= E(e^{in^{-1/2}t\sum X_j}) \\ &= \left[\phi(n^{-1/2}t)\right]^n \\ &\approx \left[\phi(0) + n^{-1/2}t\phi'(0) + n^{-1}t^2\phi''(0)/2 + o(n^{-1})\right]^n \end{aligned}$$

But now $\phi(0) = 1$ *and*

$$\phi'(t) = \frac{d}{dt}E(e^{itX_1}) = iE(X_1e^{itX_1})$$

So $\phi'(0) = E(X_1) = 0$*. Similarly*

$$\phi''(t) = i^2E(X_1^2e^{itX_1})$$

so that

$$\phi''(0) = -E(X_1^2) = -1$$

It now follows that

$$\begin{aligned} E(e^{itT}) &\approx [1 - t^2/(2n) + o(1/n)]^n \\ &\to e^{-t^2/2} \end{aligned}$$

With care we can then apply the Fourier inversion formula and get

$$\begin{aligned} f_T(x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}[\phi(tn^{-1/2})]^n dt \\ &\to \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}e^{-t^2/2}dt \\ &= \frac{1}{\sqrt{2\pi}}\phi_Z(-x) \end{aligned}$$

where ϕ_Z is the characteristic function of a standard normal variable Z. Doing the integral we find

$$\phi_Z(x) = \phi_Z(-x) = e^{-x^2/2}$$

so that

$$f_T(x) \to \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$$

which is a standard normal random variable.

This proof of the central limit theorem is not terribly general since it requires T to have a bounded continuous density. The usual central limit theorem is a statement about CDFs not densities and is

$$P(T \le t) \to P(Z \le t)$$

Convergence in Distribution

In undergraduate courses we often teach the central limit theorem: if $X_1, \ldots, X_n$ are iid from a population with mean μ and standard deviation σ then $n^{1/2}(\bar{X} - \mu)/\sigma$ has approximately a normal distribution. We also say that a Binomial(n, p) random variable has approximately a $N(np, np(1-p))$ distribution.

To make precise sense of these assertions we need to assign a meaning to statements like "X and Y have approximately the same distribution". The meaning we want to give is that X and Y have nearly the same CDF but even here we need some care. If n is a large number is the $N(0, 1/n)$ distribution close to the distribution of $X \equiv 0$? Is it close to the $N(1/n, 1/n)$ distribution? Is it close to the $N(1/\sqrt{n}, 1/n)$ distribution? If $X_n \equiv 2^{-n}$ is the distribution of X_n close to that of $X \equiv 0$?

The answer to these questions depends in part on how close close needs to be so it's a matter of definition. In practice the usual sort of approximation we want to make is to say that some random variable X, say, has nearly some continuous distribution, like $N(0, 1)$. In this case we must want to calculate probabilities like $P(X > x)$ and know that this is nearly $P(N(0, 1) > x)$. The real difficulty arises in the case of discrete

random variables; in this course we will not actually need to approximate a distribution by a discrete distribution.

When mathematicians say two things are close together they either can provide an upper bound on the distance between the two things or they are talking about taking a limit. In this course we do the latter.

Definition*: A sequence of random variables X_n converges in distribution to a random variable X if*

$$E(g(X_n)) \to E(g(X))$$

for every bounded continuous function g.

Theorem*: The following are equivalent:*

(a) X_n converges in distribution to X.

(b) $P(X_n \leq x) \to P(X \leq x)$ for each x such that $P(X = x) = 0$

(c) The characteristic functions of X_n converge to that of X:

$$E(e^{itX_n}) \to E(e^{itX})$$

for every real x.

These are all implied by

$$M_{X_n}(t) \to M_X(t) < \infty$$

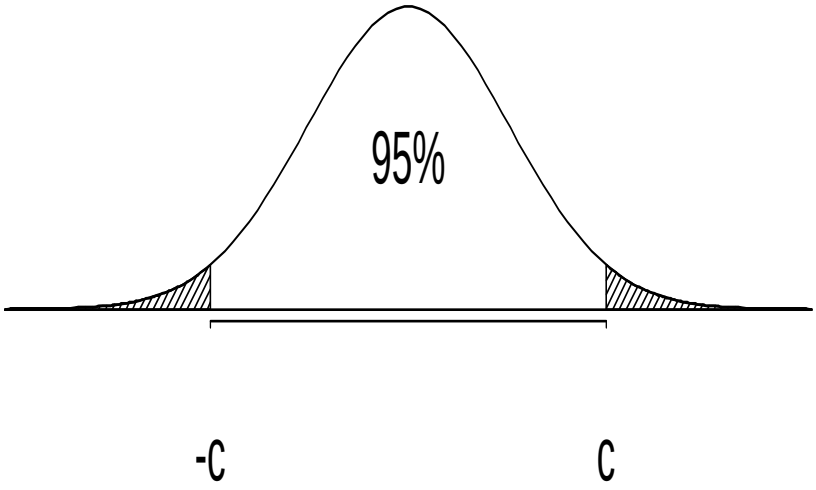
for all $|t| \leq \epsilon$ for some positive ϵ.

Now let's go back to the questions I asked:

- *$X_n \sim N(0, 1/n)$ and $X = 0$. Then*

$$P(X_n \leq x) \to \begin{cases} 1 & x > 0 \\ 0 & x < 0 \\ 1/2 & x = 0 \end{cases}$$

Now the limit is the CDF of $X = 0$ except for $x = 0$ and the CDF of X is not continuous at $x = 0$ so yes, X_n converges to X in distribution.



- *I asked if* $X_n \sim N(1/n, 1/n)$ *had a distribution close to that of* $Y_n \sim N(0, 1/n)$. *The definition I gave really requires me to answer by finding a limit* X *and proving that both* X_n *and* Y_n *converge to* X *in distribution. Take* $X = 0$. *Then*

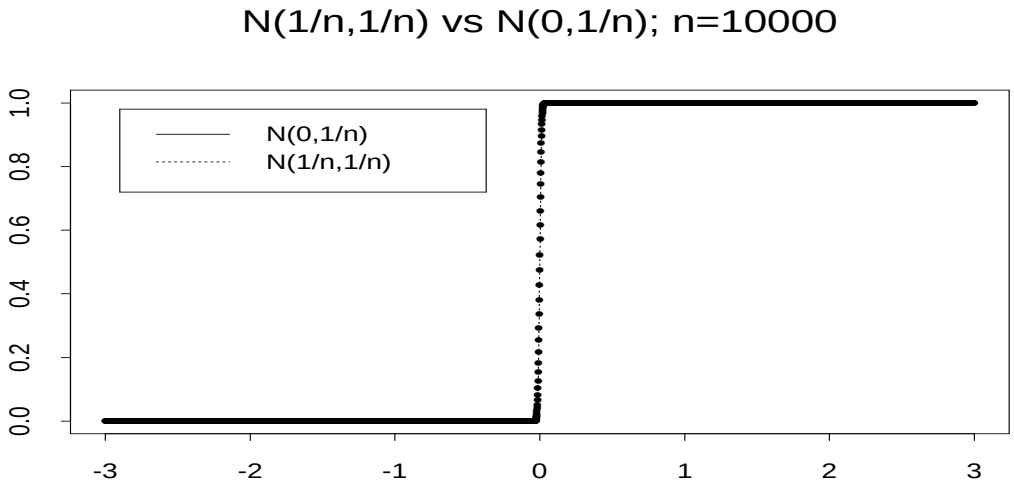
$$E(e^{tX_n}) = e^{t/n+t^2/(2n)} \to 1 = E(e^{tX})$$

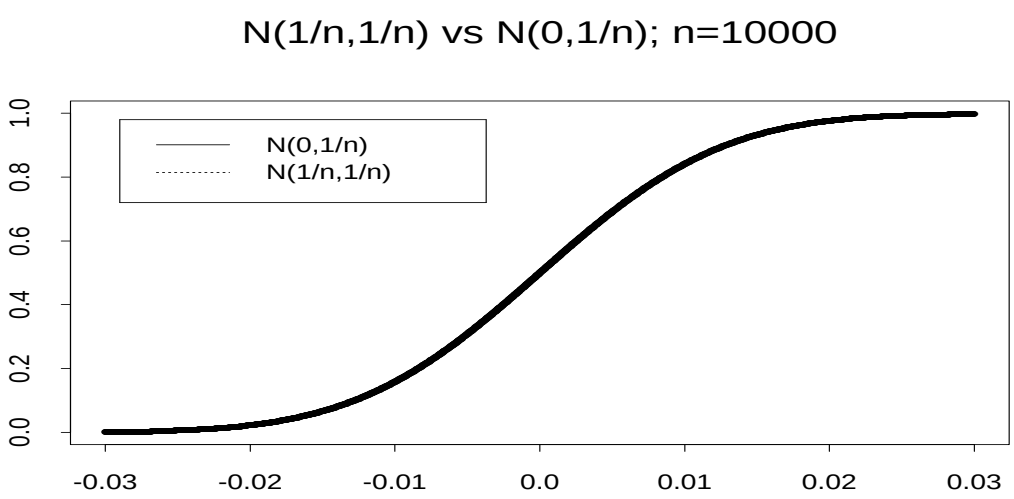
and

$$E(e^{tY_n}) = e^{t^2/(2n)} \to 1$$

so that both X_n *and* Y_n *have the same limit in distribution.*

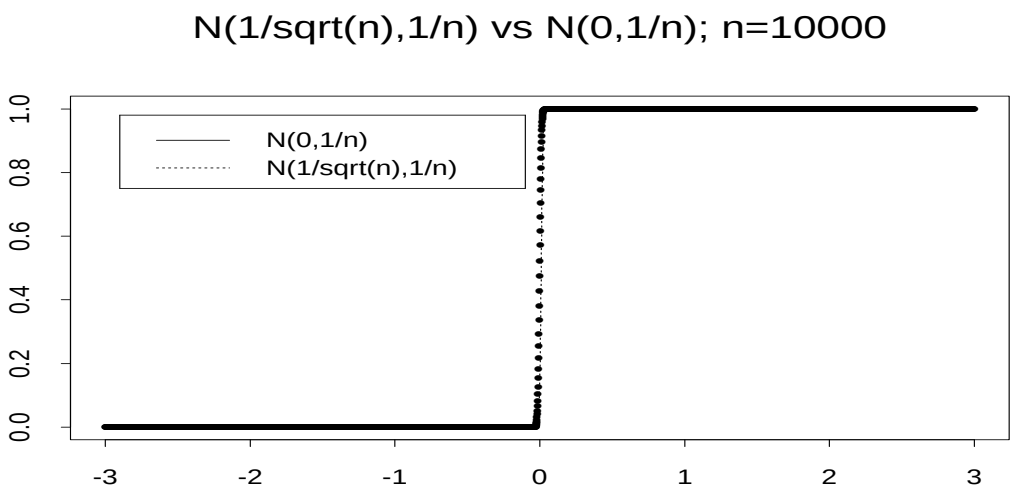
- *Now multiply both* X_n *and* Y_n *by* $n^{1/2}$ *and let* $X \sim N(0, 1)$. *Then* $\sqrt{n}X_n \sim N(n^{-1/2}, 1)$ *and* $\sqrt{n}Y_n \sim N(0, 1)$. *You can use characteristic functions to prove that both* $\sqrt{n}X_n$ *and* $\sqrt{n}Y_n$ *converge to* $N(0, 1)$ *in distribution.*

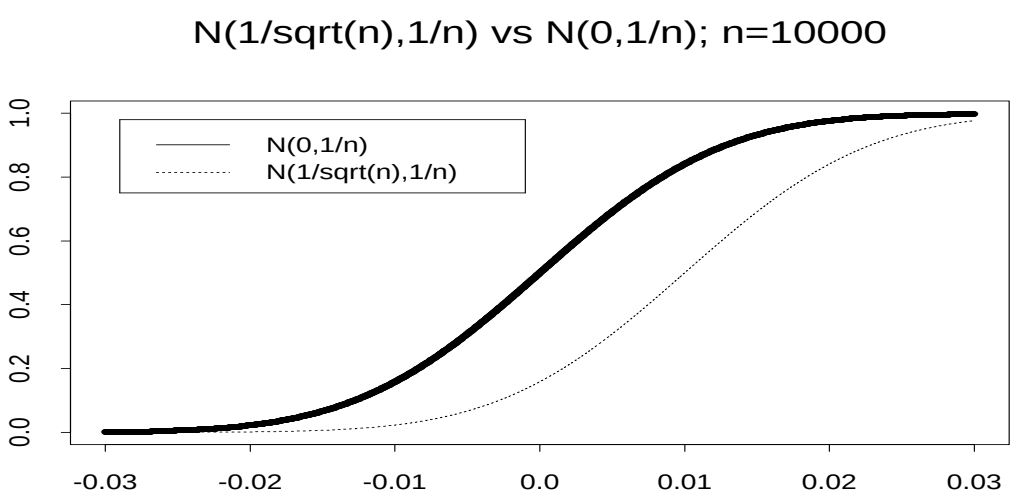




- *If you now let* $X_n \sim N(n^{-1/2}, 1/n)$ *and* $Y_n \sim N(0, 1/n)$ *then again both* X_n *and* Y_n *converge to 0 in distribution.*

- *If you multiply these* X_n *and* Y_n *by* $n^{1/2}$ *then* $n^{1/2}X_n \sim N(1, 1)$ *and* $n^{1/2}Y_n \sim N(0, 1)$ *so that* $n^{1/2}X_n$ *and* $n^{1/2}Y_n$ *are* **not** *close together in distribution.*





- *You can check that* $2^{-n} \to 0$ *in distribution.*

Here is the message you are supposed to take away from this discussion. You do distributional approximations by showing that a sequence of random variables X_n *converges to some* X*. The limit distribution should be non-trivial, like say* $N(0,1)$*. We don't say* X_n *is approximately* $N(1/n, 1/n)$ *but that* $n^{1/2}X_n$ *converges to* $N(0,1)$ *in distribution.*

The Central Limit Theorem

If $X_1, X_2, \cdots$ *are iid with mean 0 and variance 1 then* $n^{1/2}\bar{X}$ *converges in distribution to* $N(0,1)$*. That is,*

$$P(n^{1/2}\bar{X} \le x) \to \frac{1}{2\pi}\int_{-\infty}^{x} e^{-y^2/2}dy\,.$$

Proof: *As before*

$$E(e^{itn^{1/2}\bar{X}}) \to e^{-t^2/2}$$

This is the characteristic function of a $N(0,1)$ *random variable so we are done by our theorem.*

Edgeworth expansions

In fact if $\gamma = E(X^3)$ *then*

$$\phi(t) \approx 1 - t^2/2 - i\gamma t^3/6 + \cdots$$

keeping one more term. Then

$$\log(\phi(t)) = \log(1+u)$$

where

$$u = -t^2/2 - i\gamma t^3/6 + \cdots$$

Use $\log(1+u) = u - u^2/2 + \cdots$ *to get*

$$\log(\phi(t)) \approx -[t^2/2 - i\gamma t^3/6 + cdots] - [\ldots]^2/2 + \cdots$$

which rearranged is

$$\log(\phi(t)) \approx -t^2/2 + i\gamma t^3/6 + \cdots$$

Now apply this calculation to

$$\log(\phi_T(t)) \approx -t^2/2 + iE(T^3)t^3/6 + \cdots$$

Remember $E(T^3) = \gamma/\sqrt{n}$ *and exponentiate to get*

$$\phi_T(t) \approx e^{-t^2/2}\exp\{i\gamma t^3/(6\sqrt{n}) + \cdots\}$$

You can do a Taylor expansion of the second exponential around 0 because of the square root of n *and get*

$$\phi_T(t) \approx e^{-t^2/2}(1 - i\gamma t^3/(6\sqrt{n}))$$

neglecting higher order terms. This approximation to the characteristic function of T *can be inverted to get an* **Edgeworth** *approximation to the density (or distribution) of* T *which looks like*

$$f_T(x) \approx \frac{1}{\sqrt{2\pi}} e^{-x^2/2}[1 - \gamma(x^3 - 3x)/(6\sqrt{n}) + \cdots]$$

Remarks:

(a) *The error using the central limit theorem to approximate a density or a probability is proportional to* $n^{-1/2}$

(b) *This is improved to* n^{-1} *for symmetric densities for which* $\gamma = 0$.

(c) *The expansions are* **asymptotic**. *This means that the series indicated by* $\cdots$ *usually does* **not** *converge. When* $n = 25$ *it may help to take the second term but get worse if you include the third or fourth or more.*

(d) *You can integrate the expansion above for the density to get an approximation for the CDF.*

Multivariate convergence in distribution

Definition: *$X_n \in R^p$ converges in distribution to $X \in R^p$ if*

$$E(g(X_n)) \to E(g(X))$$

for each bounded continuous real valued function g on R^p.

This is equivalent to either of

Cramér Wold Device: *$a^t X_n$ converges in distribution to $a^t X$ for each $a \in R^p$*

or

Convergence of Characteristic Functions:

$$E(e^{ia^t X_n}) \to E(e^{ia^t X})$$

for each $a \in R^p$.

Extensions of the CLT

(a) *If $Y_1, Y_2, \cdots$ are iid in R^p with mean μ and variance covariance Σ then $n^{1/2}(\bar{Y} - \mu)$ converges in distribution to $MVN(0, \Sigma)$.*

(b) *If for each n we have a set of independent mean 0 random variables $X_{n1}, \ldots, X_{nn}$ and $E(X_{ni}) = 0$ and $Var(\sum_i X_{ni}) = 1$ and*

$$\sum E(|X_{ni}|^3) \to 0$$

then $\sum_i X_{ni}$ converges in distribution to $N(0, 1)$. This is the Lyapunov central limit theorem.

(c) *As in the Lyapunov central CLT but replace the third moment condition with*

$$\sum E(X_{ni}^2 1(|X_{ni}| > \epsilon)) \to 0$$

for each $\epsilon > 0$ then again $\sum_i X_{ni}$ converges in distribution to $N(0, 1)$. This is the Lindeberg central limit theorem. (Lyapunov's condition implies Lindeberg's.)

(d) *There are extensions to random variables which are not independent. Examples include the m-dependent central limit theorem, the martingale central limit theorem, the central limit theorem for mixing processes.*

(e) *Many important random variables are not sums of independent random variables. We handle these with Slutsky's theorem and the δ method.*

Slutsky's Theorem: *If X_n converges in distribution to X and Y_n converges in distribution (or in probability) to c, a constant, then $X_n + Y_n$ converges in distribution to $X + c$.*

Warning: the hypothesis that the limit of Y_n be constant is essential.

The delta method: *Suppose a sequence Y_n of random variables converges to some y a constant and that if we define $X_n = a_n(Y_n - y)$ then X_n converges in distribution to some random variable X. Suppose that f is a differentiable function on the range of Y_n. Then $a_n(f(Y_n) - f(y))$ converges in distribution to $f'(y)X$. If X_n is in R^p and f maps R^p to R^q then f' is the $q \times p$ matrix of first derivatives of components of f.*

Example: *Suppose $X_1, \ldots, X_n$ are a sample from a population with mean μ, variance σ^2, and third and fourth central moments μ_3 and μ_4. Then*

$$n^{1/2}(s^2 - \sigma^2) \Rightarrow N(0, \mu_4 - \sigma^4)$$

where $\Rightarrow$ is notation for convergence in distribution. For simplicity I define $s^2 = \overline{X^2} - \bar{X}^2$.

We take Y_n to be the vector with components $(\overline{X^2}, \bar{X})$. Then Y_n converges to $y = (\mu^2 + \sigma^2, \mu)$. Take $a_n = n^{1/2}$. Then

$$n^{1/2}(Y_n - y)$$

converges in distribution to $MVN(0, \Sigma)$ with

$$\Sigma = \begin{bmatrix} \mu_4 - \sigma^4 & \mu_3 - \mu(\mu^2 + \sigma^2) \\ \mu_3 - \mu(\mu^2 + \sigma^2) & \sigma^2 \end{bmatrix}$$

Define $f(x_1, x_2) = x_1 - x_2^2$. Then $s^2 = f(Y_n)$ and the gradient of f has components $(1, -2x_2)$. This leads to

$$n^{1/2}(s^2 - \sigma^2) \approx n^{1/2}(1, -2\mu) \begin{bmatrix} \overline{X^2} - (\mu^2 + \sigma^2) \\ \bar{X} - \mu \end{bmatrix}$$

which converges in distribution to the law of $(1, -2\mu)Y$ which is $N(0, a^t \Sigma a)$ where $a = (1, -2\mu)^t$. This boils down to $N(0, \mu_4 - \sigma^2)$.

Remark: In this sort of problem it is best to learn to recognize that the sample variance is unaffected by subtracting μ from each X. Thus there is no loss in assuming $\mu = 0$ which simplifies Σ and a.

Special case: if the observations are $N(\mu, \sigma^2)$ then $\mu_3 = 0$ and $\mu_4 = 3\sigma^4$. Our calculation has

$$n^{1/2}(s^2 - \sigma^2) \Rightarrow N(0, 2\sigma^4)$$

You can divide through by σ^2 and get

$$n^{1/2}(\frac{s^2}{\sigma^2} - 1) \Rightarrow N(0, 2)$$

In fact $(n-1)s^2/\sigma^2$ *has a* χ^2_{n-1} *distribution and so the usual central limit theorem shows that*

$$(n-1)^{1/2}[(n-1)s^2/\sigma^2 - (n-1)] \Rightarrow N(0,2)$$

(using mean of χ^2_1 *is 1 and variance is 2). Factoring out* $n-1$ *gives the assertion that*

$$(n-1)^{1/2}(s^2/\sigma^2 - 1) \Rightarrow N(0,2)$$

which is our δ *method calculation except for using* $n-1$ *instead of* n*. This difference is unimportant as can be checked using Slutsky's theorem.*

Monte Carlo

The last method of distribution theory that I will review is Monte Carlo simulation. Suppose you have some random variables $X_1, \ldots, X_n$ *whose joint distribution is specified and a statistic* $T(X_1, \ldots, X_n)$ *whose distribution you want to know. To compute something like* $P(T > t)$ *for some specific value of* t *we appeal to the limiting relative frequency interpretation of probability:* $P(T > t)$ *is the limit of the proportion of trials in a long sequence of trials in which* T *occurs. We use a (pseudo) random number generator to generate a sample* $X_1, \ldots, X_n$ *and then calculate the statistic getting* T_1*. Then we generate a new sample (independently of our first, say) and calculate* T_2*. We repeat this a large number of times say* N *and just count up how many of the* T_k *are larger than* t*. If there are* M *such* T_k *we estimate that* $P(T > t) = M/N$*.*

The quantity M *has a Binomial*$(N, p = P(T > t))$ *distribution. The standard error of* M/N *is then* $p(1-p)/N$ *which is estimated by* $M(N-M)/N^3$*. This permits us to guess the accuracy of our study.*

Notice that the standard deviation of M/N *is* $\sqrt{p(1-p)}/\sqrt{N}$ *so that to improve the accuracy by a factor of 2 requires 4 times as many samples. This makes Monte Carlo a relatively time consuming method of calculation. There are a number of tricks to make the method more accurate (though they only change the constant of proportionality – the SE is still inversely proportional to the square root of the sample size).*

Generating the Sample

Most computer languages have a facility for generating pseudo uniform random numbers, that is, variables U *which have (approximately of course) a Uniform*$[0,1]$ *distribution. Other distributions are generated by transformation:*

Exponential*:* $X = -\log U$ *has an exponential distribution:*

$$P(X > x) = P(-\log(U) > x) = P(U \le e^{-x}) = e^{-x}$$

Random uniforms generated on the computer sometimes have only 6 or 7 digits or so of detail. This can make the tail of your distribution grainy. If U *were actually a multiple of* 10^{-6} *for instance then the largest possible value of* X *is* $6\log(10)$*. This problem can be ameliorated by the following algorithm:*

- *Generate U a Uniform[0,1] variable.*
- *Pick a small ϵ like 10^{-3} say. If $U > \epsilon$ take $Y = -\log(U)$.*
- *If $U \leq \epsilon$ remember that the conditional distribution of $Y - y$ given $Y > y$ is exponential. You use this by generating a new U' and computing $Y' = -\log(U')$. Then take $Y = Y' - \log(\epsilon)$. The resulting Y has an exponential distribution. You should check this by computing $P(Y > y)$.*

Normal: *In general if F is a continuous CDF and U is Uniform[0,1] then $Y = F^{-1}(U)$ has CDF F because*

$$P(Y \leq y) = P(F^{-1}(U) \leq y) = P(U \leq F(y)) = F(y)$$

This is almost the technique in the exponential distribution. For the normal distribution $F = \Phi$ (Φ is a common notation for the standard normal CDF) there is no closed form for F^{-1}. You could use a numerical algorithm to compute F^{-1} or you could use the following Box Müller trick. Generate U_1, U_2 two independent Uniform[0,1] variables. Define $Y_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$ and $Y_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$. Then you can check using the change of variables formula that Y_1 and Y_2 are independent $N(0,1)$ variables.

Acceptance Rejection

If you can't easily calculate F^{-1} but you know f you can try the acceptance rejection method. Find a density g and a constant c such that $f(x) \leq cg(x)$ for each x and G^{-1} is computable or you otherwise know how to generate observations $W_1, W_2, \ldots$ independently from g. Generate W_1. Compute $p = f(W_1)/(cg(W_1)) \leq 1$. Generate a uniform[0,1] random variable U_1 independent of all the Ws and let $Y = W_1$ if $U_1 \leq p$. Otherwise get a new W and a new U and repeat until you find a $U_i \leq f(W_i)/(cg(W_i))$. You make Y be the last W you generated. This Y has density f.

Markov Chain Monte Carlo

In the last 10 years the following tactic has become popular, particularly for generating multivariate observations. If $W_1, W_2, \ldots$ is an (ergodic) Markov chain with stationary transitions and the stationary initial distribution of W has density f then you can get random variables which have the marginal density f by starting off the Markov chain and letting it run for a long time. The marginal distributions of the W_i converge to f. So you can estimate things like $\int_A f(x)dx$ by computing the fraction of the W_i which land in A.

There are now many versions of this technique. Examples include Gibbs Sampling and the Metropolis-Hastings algorithm. (The technique was invented in the 1950s by physicists: Metropolis et al. One of the authors of the paper was Edward Teller "father of the hydrogen bomb".)

Importance Sampling

If you want to compute

$$\theta \equiv E(T(X)) = \int T(x)f(x)dx$$

you can generate observations from a different density g and then compute

$$\hat{\theta} = n^{-1} \sum T(X_i) f(X_i)/g(X_i)$$

Then

$$\begin{aligned} E(\hat{\theta}) &= n^{-1} \sum E(T(X_i) f(X_i)/g(X_i) \\ &= \int [T(x) f(x)/g(x)] g(x) dx \\ &= \int T(x) f(x) dx \\ &= \theta \end{aligned}$$

Variance reduction

Consider the problem of estimating the distribution of the sample mean for a Cauchy random variable. The Cauchy density is

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

We generate $U_1, \ldots, U_n$ uniforms and then define $X_i = \tan^{-1}(\pi(U_i - 1/2))$. Then we compute $T = \bar{X}$. Now to estimate $p = P(T > t)$ we would use

$$\hat{p} = \sum_{i=1}^{N} 1(T_i > t)/N$$

after generating N samples of size n. This estimate is unbiased and has standard error $\sqrt{p(1-p)/N}$.

We can improve this estimate by remembering that $-X_i$ also has Cauchy distribution. Take $S_i = -T_i$. Remember that S_i has the same distribution as T_i. Then we try (for $t > 0$)

$$\tilde{p} = [\sum_{i=1}^{N} 1(T_i > t) + \sum_{i=1}^{N} 1(S_i > t)]/(2N)$$

which is the average of two estimates like $\hat{p}$. The variance of $\tilde{p}$ is

$$(4N)^{-1} Var(1(T_i > t) + 1(S_i > t)) = (4N)^{-1} Var(1(|T| > t))$$

which is

$$\frac{2p(1-2p)}{4N} = \frac{p(1-2p)}{2N}$$

Notice that the variance has an extra 2 in the denominator and that the numerator is also smaller – particularly for p near 1/2. So this method of variance reduction has resulted in a need for only half the sample size to get the same accuracy.

Regression estimates

Suppose we want to compute

$$\theta = E(|Z|)$$

where Z is standard normal. We generate N iid $N(0,1)$ variables $Z_1, \ldots, Z_N$ and compute $\hat{\theta} = \sum |Z_i|/N$. But we know that $E(Z_i^2) = 1$ and can see easily that $\hat{\theta}$ is positively correlated with $\sum Z_i^2/N$. So we consider using

$$\tilde{\theta} = \hat{\theta} - c(\sum Z_i^2/N - 1)$$

Notice that $E(\tilde{\theta}) = \theta$ and

$$Var(\tilde{\theta}) = Var(\hat{\theta}) - 2cCov(\hat{\theta}, \sum Z_i^2/n) + c^2 Var(\sum Z_i^2/N)$$

The value of c which minimizes this is

$$c = \frac{Cov(\hat{\theta}, \sum Z_i^2/n)}{Var(\sum Z_i^2/N)}$$

and this value can be estimated by regressing the $|Z_i|$ on the Z_i^2!

Chapter 4

Estimation

Statistical Inference

Definition: *A* **model** *is a family* $\{P_\theta; \theta \in \Theta\}$ *of possible distributions for some random variable* X. *(Our data set is* X*, so* X *will generally be a big vector or matrix or even more complicated object.)*

We will assume throughout this course that the true distribution P *of* X *is in fact some* P_{θ_0} *for some* $\theta_0 \in \Theta$. *We call* θ_0 *the true value of the parameter. Notice that this assumption will be wrong; we hope it is not wrong in an important way. If we are very worried that it is wrong we enlarge our model putting in more distributions and making* Θ *bigger.*

Our goal is to observe the value of X *and then guess* θ_0 *or some property of* θ_0. *We will consider the following classic mathematical versions of this:*

(a) Point estimation: we must compute an estimate $\hat{\theta} = \hat{\theta}(X)$ *which lies in* Θ *(or something close to* Θ*).*

(b) Point estimation of a function of θ*: we must compute an estimate* $\hat{\phi} = \hat{\phi}(X)$ *of* $\phi = g(\theta)$.

(c) Interval (or set) estimation. We must compute a set $C = C(X)$ *in* Θ *which we think will contain* θ_0.

(d) Hypothesis testing: We must decide whether or not $\theta_0 \in \Theta_0$ *where* $\Theta_0 \subset \Theta$.

(e) Prediction: we must guess the value of an observable random variable Y *whose distribution depends on* θ_0. *Typically* Y *is the value of the variable* X *in a repetition of the experiment.*

There are several schools of statistical thinking with different views on how these problems should be done. The main schools of thought may be summarized roughly as follows:

- **Neyman Pearson**: *In this school of thought a statistical procedure should be evaluated by its long run frequency performance. You imagine repeating the data collection exercise many times, independently. The quality of a procedure is measured by its average performance when the true distribution of the X values is P_{θ_0}.*
- **Bayes**: *In this school of thought we treat θ as being random just like X. We compute the conditional distribution of what we don't know given what we do know. In particular we ask how a procedure will work on the data we actually got – no averaging over data we might have got.*
- **Likelihood**: *This school tries to combine the previous 2 by looking only at the data we actually got but trying to avoid treating θ as random.*

For the next several weeks we do only the Neyman Pearson approach, though we use that approach to evaluate the quality of likelihood methods.

Likelihood Methods of Inference

Suppose you toss a coin 6 times and get Heads twice. If p is the probability of getting H then the probability of getting 2 heads is

$$15p^2(1-p)^4$$

This probability, thought of as a function of p, is the **likelihood** *function for this particular data.*

Definition: *A* **model** *is a family $\{P_\theta; \theta \in \Theta\}$ of possible distributions for some random variable X. Typically the model is described by specifying $\{f_\theta(x); \theta \in \Theta\}$ the set of possible densities of X.*

Definition: *The likelihood function is the function L whose domain is Θ and whose values are given by*

$$L(\theta) = f_\theta(X)$$

The key point is to think about how the density depends on θ not about how it depends on X. Notice that X, the observed value of the data, has been plugged into the formula for the density. Notice also that the coin tossing example is like this but with f being the discrete density. We use the likelihood in most of our inference problems:

(a) Point estimation: we must compute an estimate $\hat{\theta} = \hat{\theta}(X)$ which lies in Θ. The **maximum likelihood estimate (MLE)** *of θ is the value $\hat{\theta}$ which maximizes $L(\theta)$ over $\theta \in \Theta$ if such a $\hat{\theta}$ exists.*

(b) Point estimation of a function of θ: we must compute an estimate $\hat{\phi} = \hat{\phi}(X)$ of $\phi = g(\theta)$. We use $\hat{\phi} = g(\hat{\theta})$ where $\hat{\theta}$ is the MLE of θ.

(c) *Interval (or set) estimation. We must compute a set* $C = C(X)$ *in* Θ *which we think will contain* θ_0*. We will use*

$$\{\theta \in \Theta : L(\theta) > c\}$$

for a suitable c*.*

(d) *Hypothesis testing: We must decide whether or not* $\theta_0 \in \Theta_0$ *where* $\Theta_0 \subset \Theta$*. We base our decision on the likelihood ratio*

$$\frac{\sup\{L(\theta); \theta \in \Theta_0\}}{\sup\{L(\theta); \theta \in \Theta \setminus \Theta_0\}}$$

Maximum Likelihood Estimation

To find an MLE we maximize L*. This is a typical function maximization problem which we approach by setting the gradient of* L *equal to 0 and then checking to see that the root is a maximum, not a minimum or saddle point.*

We begin by examining some likelihood plots in examples:

Cauchy Data

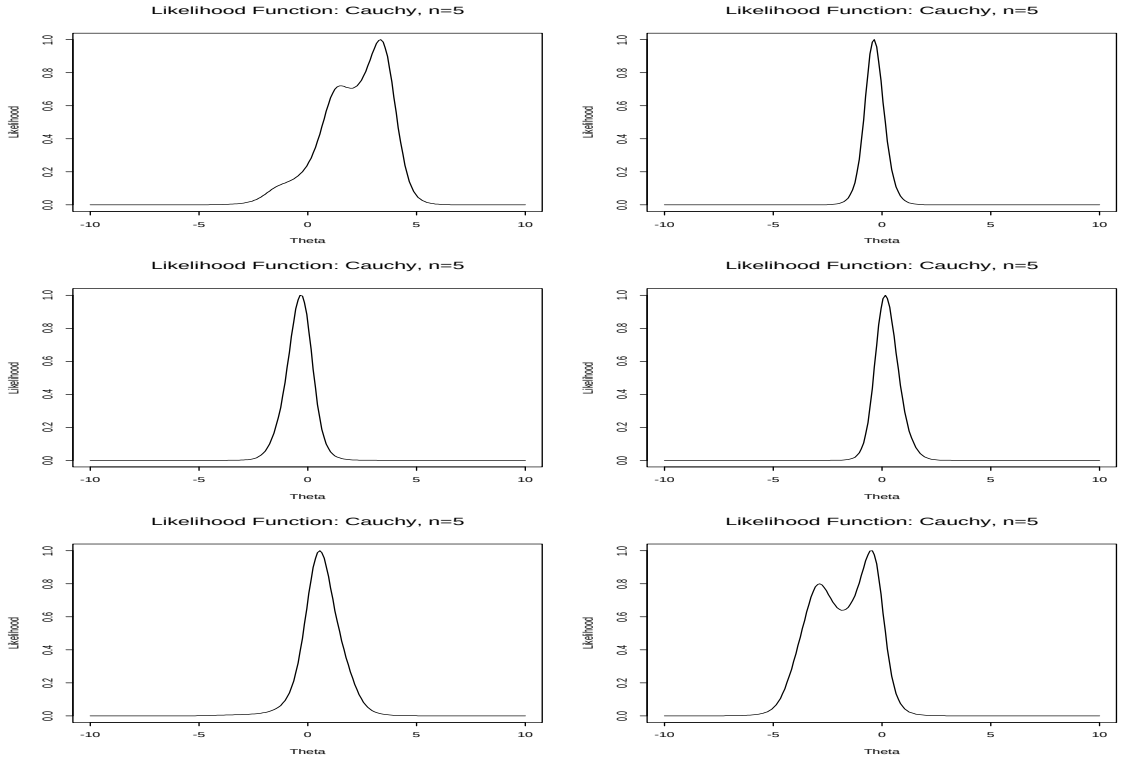
We have a sample $X_1, \ldots, X_n$ *from the Cauchy*(θ) *density*

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

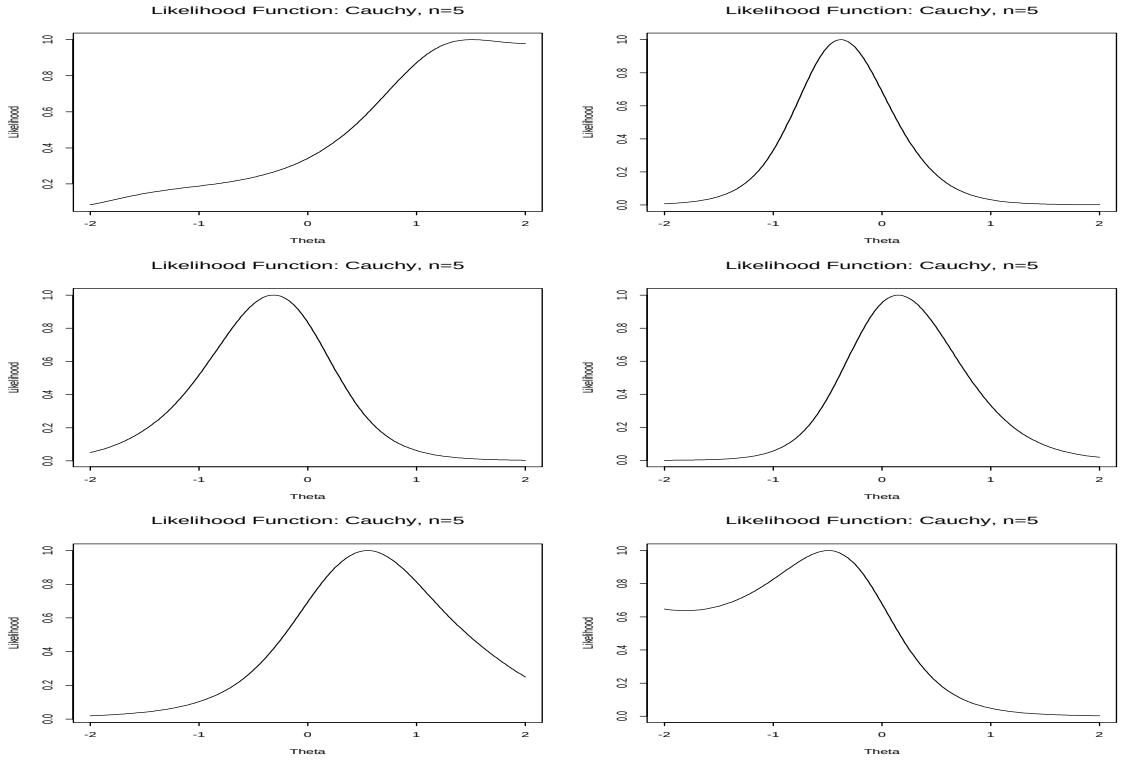
The likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\pi(1 + (X_i - \theta)^2)}$$

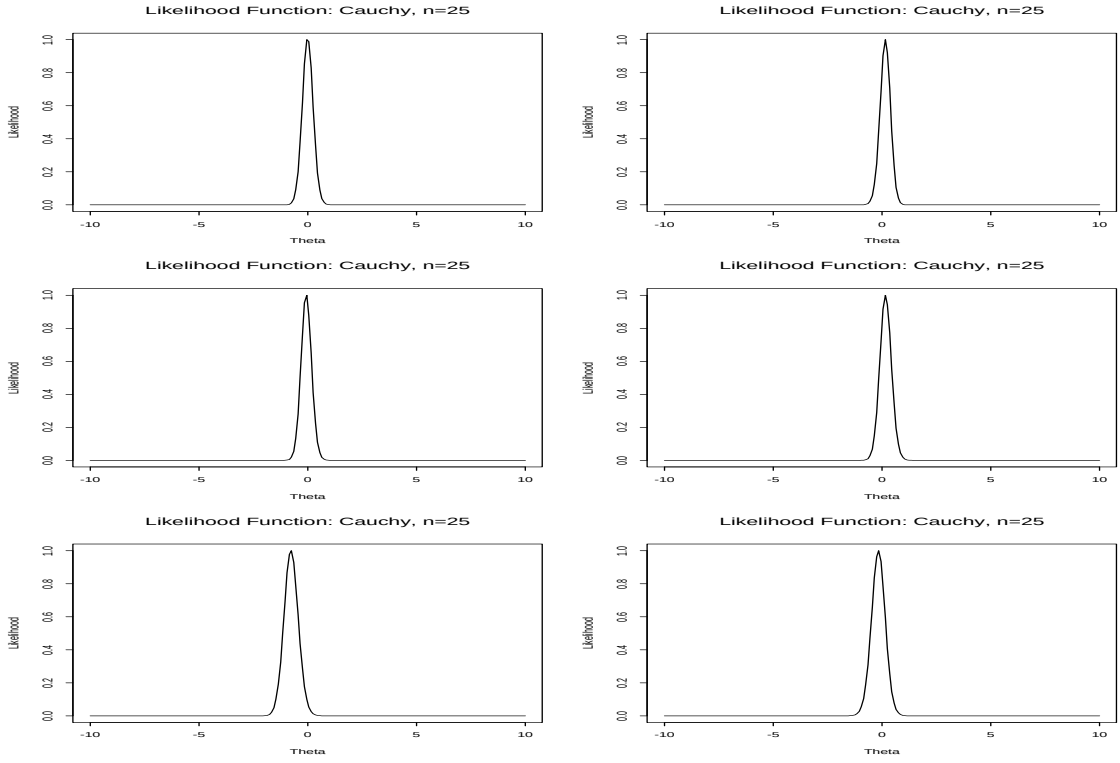
Here are some plots of this function for 6 samples of size 5.



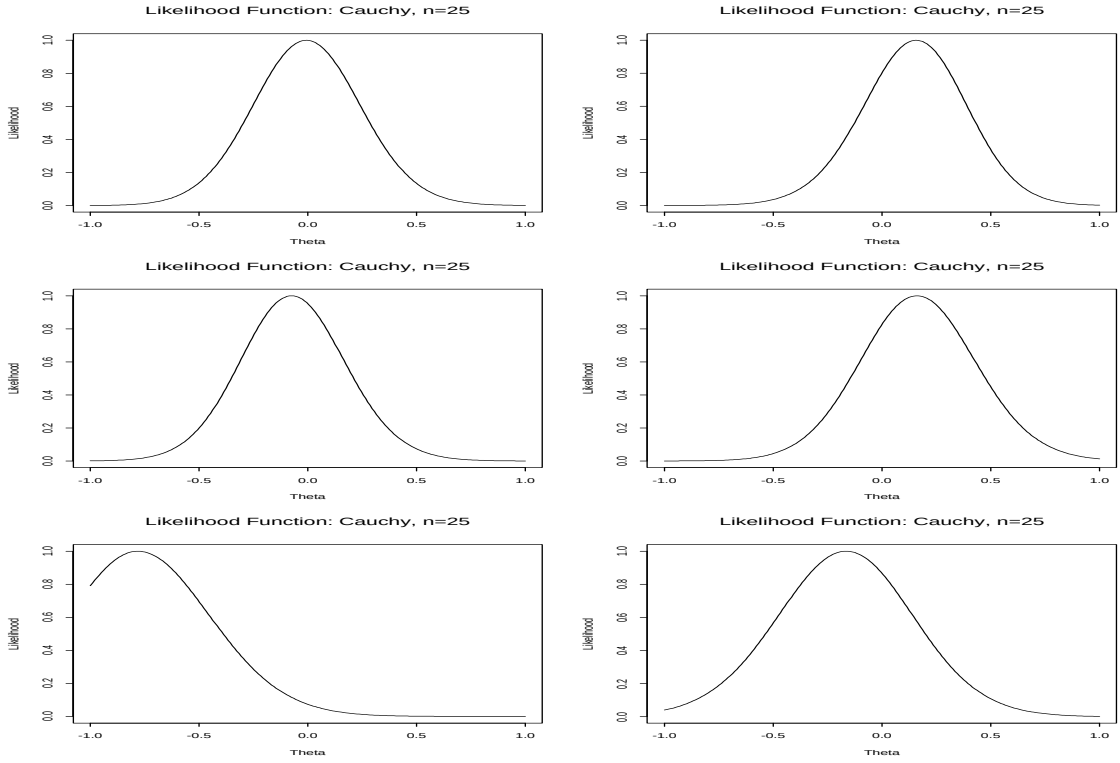
Here are close up views of these plots for θ between -2 and 2.



Now for sample size 25.



Here are close up views of these plots for θ between -2 and 2.



I want you to notice the following points:

- *The likelihood functions have peaks near the true value of θ (which is 0 for the*

data sets I generated).

- *The peaks are narrower for the larger sample size.*
- *The peaks have a more regular shape for the larger value of n.*
- *I actually plotted $L(\theta)/L(\hat{\theta})$ which has exactly the same shape as L but runs from 0 to 1 on the vertical scale.*

To maximize this likelihood we would have to differentiate L and set the result equal to 0. Notice that L is a product of n terms and the derivative will then be

$$\sum_{i=1}^{n} \prod_{j \neq i} \frac{1}{\pi(1+(X_j-\theta)^2)} \frac{2(X_i-\theta)}{\pi(1+(X_i-\theta)^2)^2}$$

which is quite unpleasant. It is much easier to work with the logarithm of L since the log of a product is a sum and the logarithm is monotone increasing.

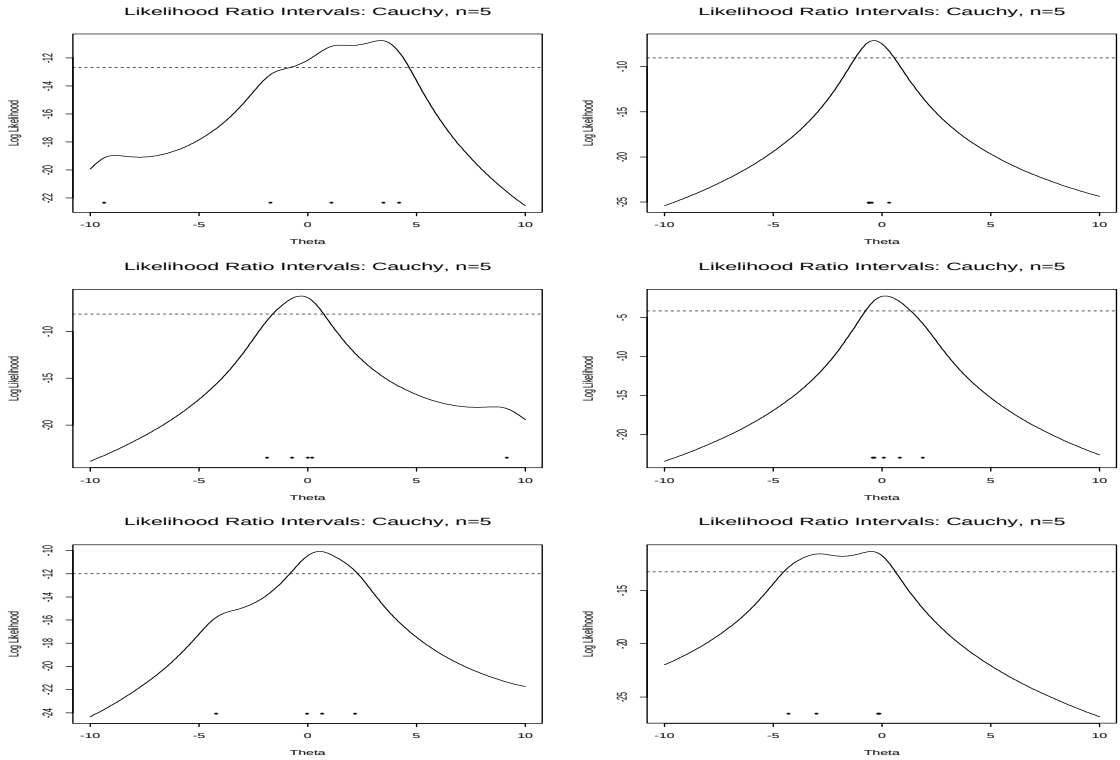
Definition*: The* **Log Likelihood** *function is*

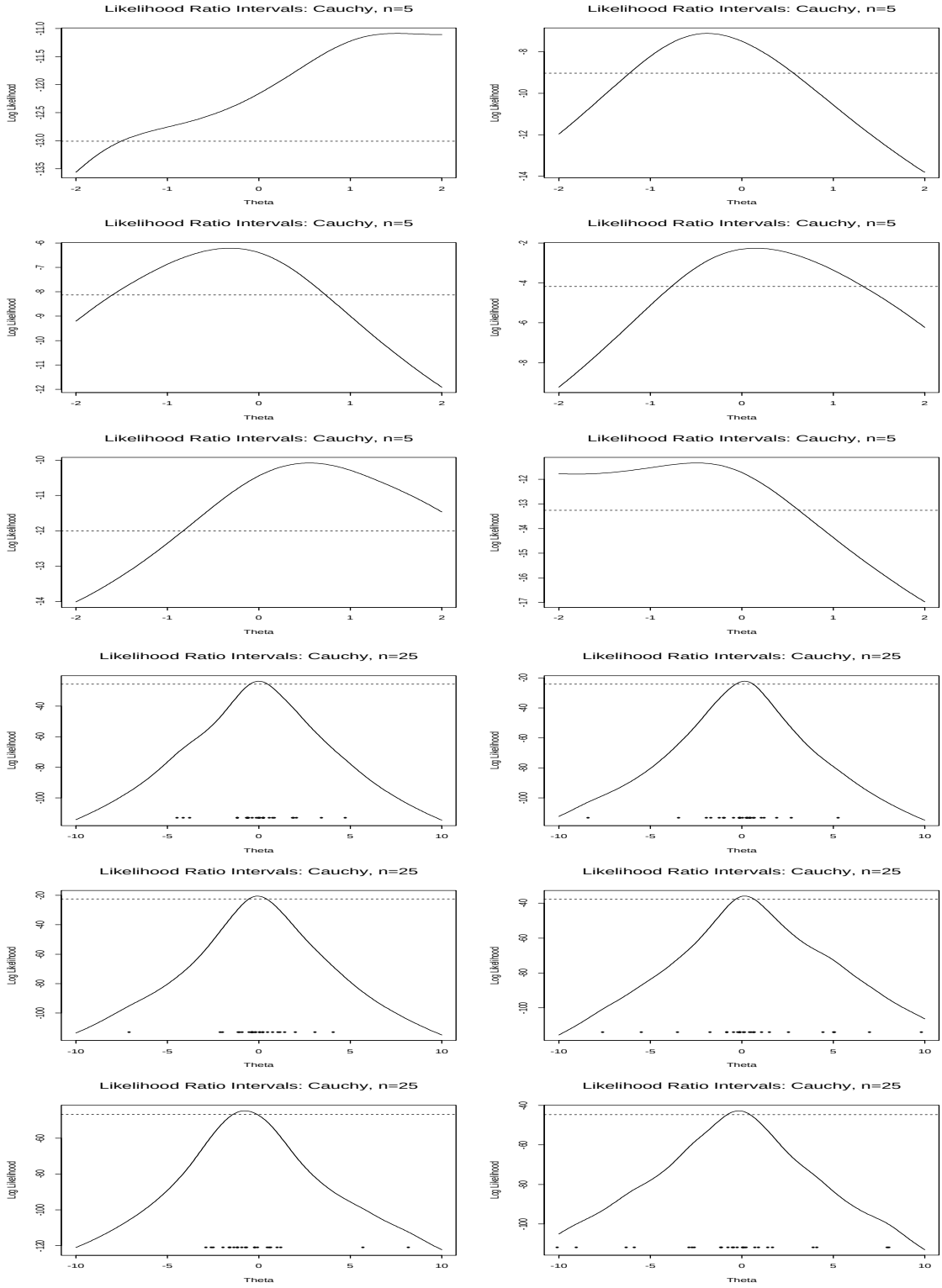
$$\ell(\theta) = \log(L(\theta))) \, .$$

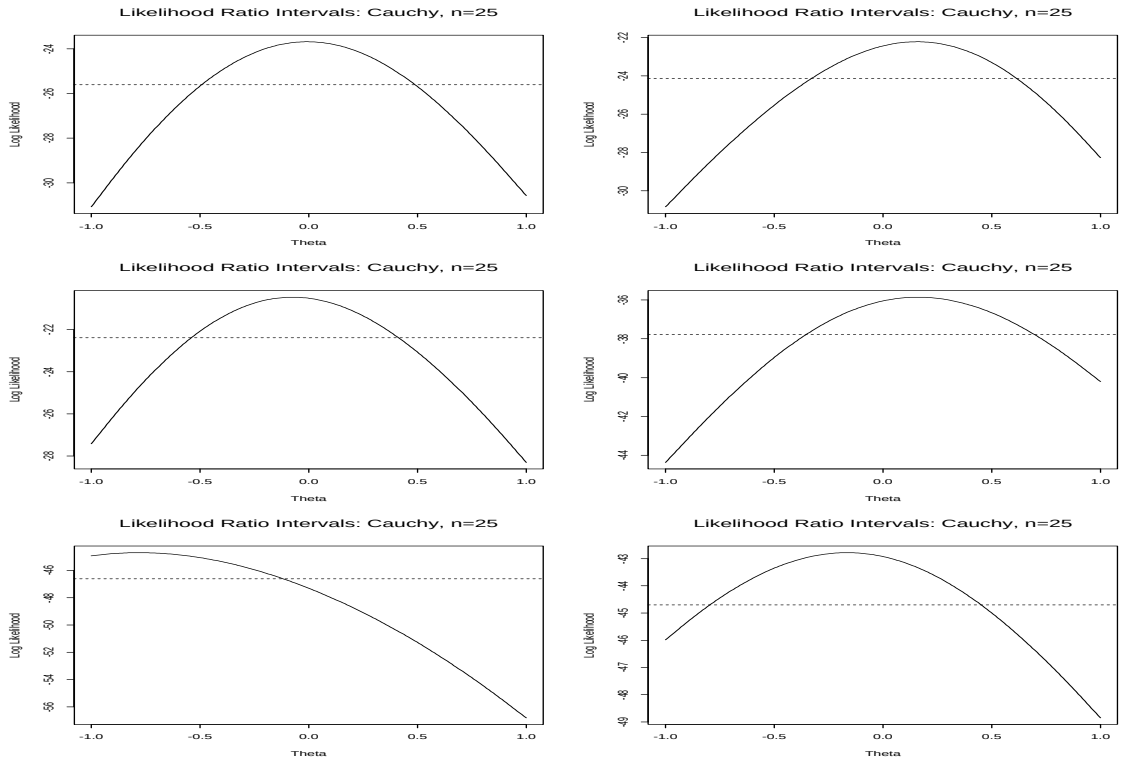
For the Cauchy problem we have

$$\ell(\theta) = -\sum \log(1+(X_i-\theta)^2) - n\log(\pi)$$

Here are the logarithms of the likelihoods plotted above:




Likelihood Ratio Intervals: Cauchy, n=5
Likelihood Ratio Intervals: Cauchy, n=5
Likelihood Ratio Intervals: Cauchy, n=5
Likelihood Ratio Intervals: Cauchy, n=5
Likelihood Ratio Intervals: Cauchy, n=5
Likelihood Ratio Intervals: Cauchy, n=5
Likelihood Ratio Intervals: Cauchy, n=25
Likelihood Ratio Intervals: Cauchy, n=25
Likelihood Ratio Intervals: Cauchy, n=25
Likelihood Ratio Intervals: Cauchy, n=25
Likelihood Ratio Intervals: Cauchy, n=25
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
Theta



I want you to notice the following points:

- *The log likelihood functions with $n = 25$ have pretty smooth shapes which look rather parabolic.*
- *For $n = 5$ there are plenty of local maxima and minima of ℓ.*

You can see that the likelihood will tend to 0 as $|\theta| \to \infty$ so that the maximum of ℓ will occur at a root of ℓ', the derivative of ℓ with respect to θ.

Definition*: The* **Score Function** *is the gradient of ℓ*

$$U(\theta) = \frac{\partial \ell}{\partial \theta}$$

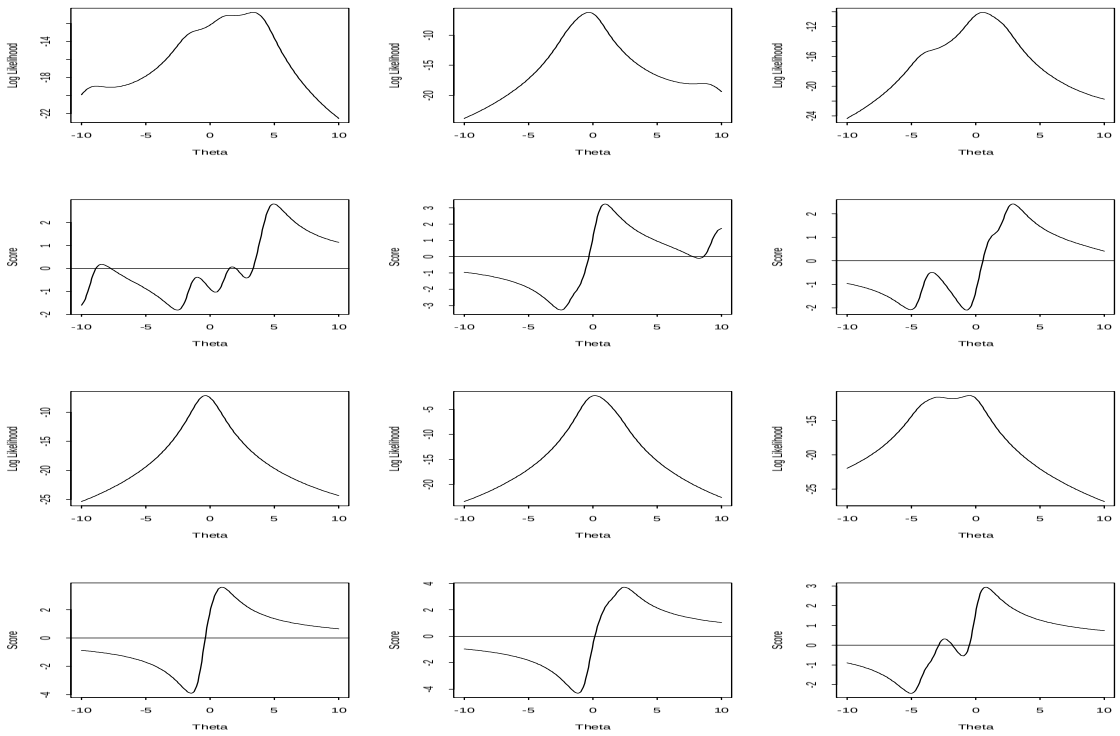
The MLE $\hat{\theta}$ usually solves the **Likelihood Equations**

$$U(\theta) = 0$$

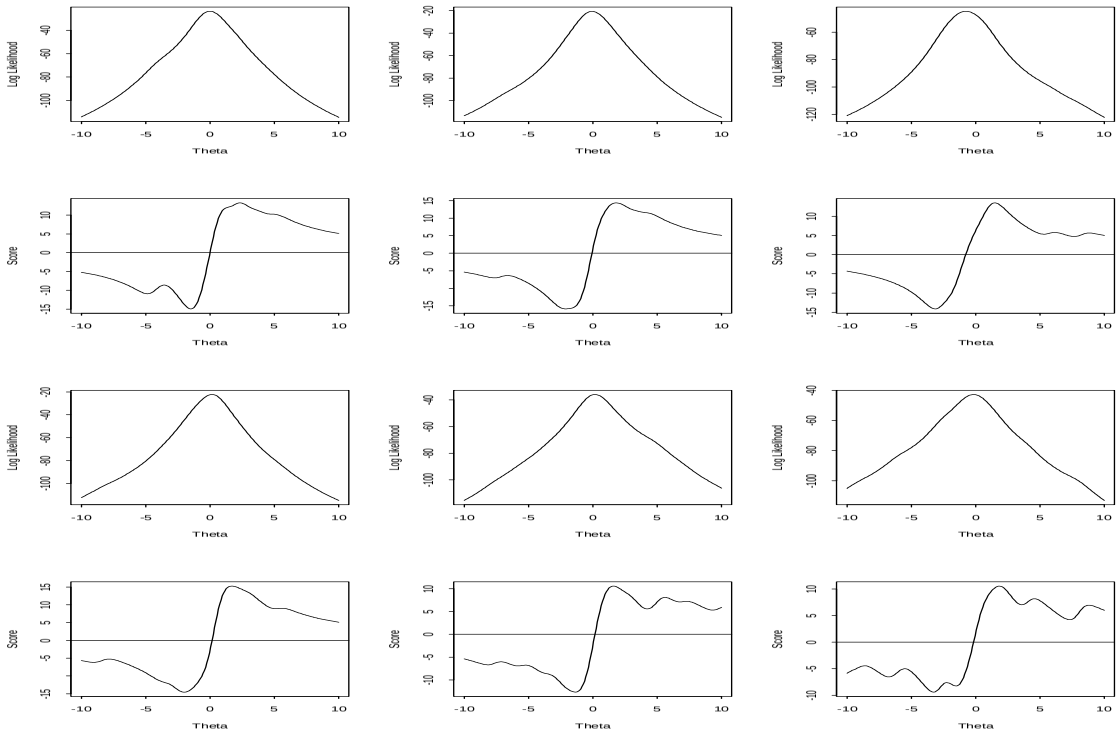
In our Cauchy example we find

$$U(\theta) = \sum \frac{2(X_i - \theta)}{1 + (X_i - \theta)^2}$$

Here are some plots of the score functions for $n = 5$ for our Cauchy data sets. Each score is plotted beneath a plot of the corresponding ℓ.



Notice that there are often multiple roots of the likelihood equations. Here is $n = 25$*:*



The Binomial Distribution

If X has a Binomial(n, θ) distribution then

$$L(\theta) = \begin{pmatrix} n \\ X \end{pmatrix} \theta^X (1-\theta)^{n-X}$$

$$\ell(\theta) = \log \begin{pmatrix} n \\ X \end{pmatrix} + X \log(\theta) + (n - X) \log(1-\theta)$$

$$U(\theta) = \frac{X}{\theta} - \frac{n-X}{1-\theta}$$

The function L is 0 at $\theta = 0$ and at $\theta = 1$ unless $X = 0$ or $X = n$ so for $1 \le X \le n$ the MLE must be found by setting $U = 0$ and getting

$$\hat{\theta} = \frac{X}{n}$$

For $X = n$ the log-likelihood has derivative

$$U(\theta) = \frac{n}{\theta} > 0$$

for all θ so that the likelihood is an increasing function of θ which is maximized at $\hat{\theta} = 1 = X/n$. Similarly when $X = 0$ the maximum is at $\hat{\theta} = 0 = X/n$.

The Normal Distribution

Now we have $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$. There are two parameters $\theta = (\mu, \sigma)$. We find

$$L(\mu, \sigma) = (2\pi)^{-n/2} \sigma^{-n} \exp\{-\sum (X_i - \mu)^2 / (2\sigma^2)\}$$

$$\ell(\mu, \sigma) = -n \log(2\pi)/2 - \frac{\sum (X_i - mu)^2}{2\sigma^2}$$

$$U(\mu, \sigma) = \begin{bmatrix} \frac{\sum (X_i - \mu)}{\sigma^2} \\ \frac{\sum (X_i - \mu)^2}{\sigma^3} - \frac{n}{\sigma} \end{bmatrix}$$

Notice that U is a function with two components because θ has two components.

Setting the likelihood equal to 0 and solving gives

$$\hat{\mu} = \bar{X}$$

and

$$\hat{\sigma} = \sqrt{\frac{\sum (X_i - \bar{X})^2}{n}}$$

You need to check that this is actually a maximum. To do so you compute one more derivative. The matrix H of second derivatives of ℓ is

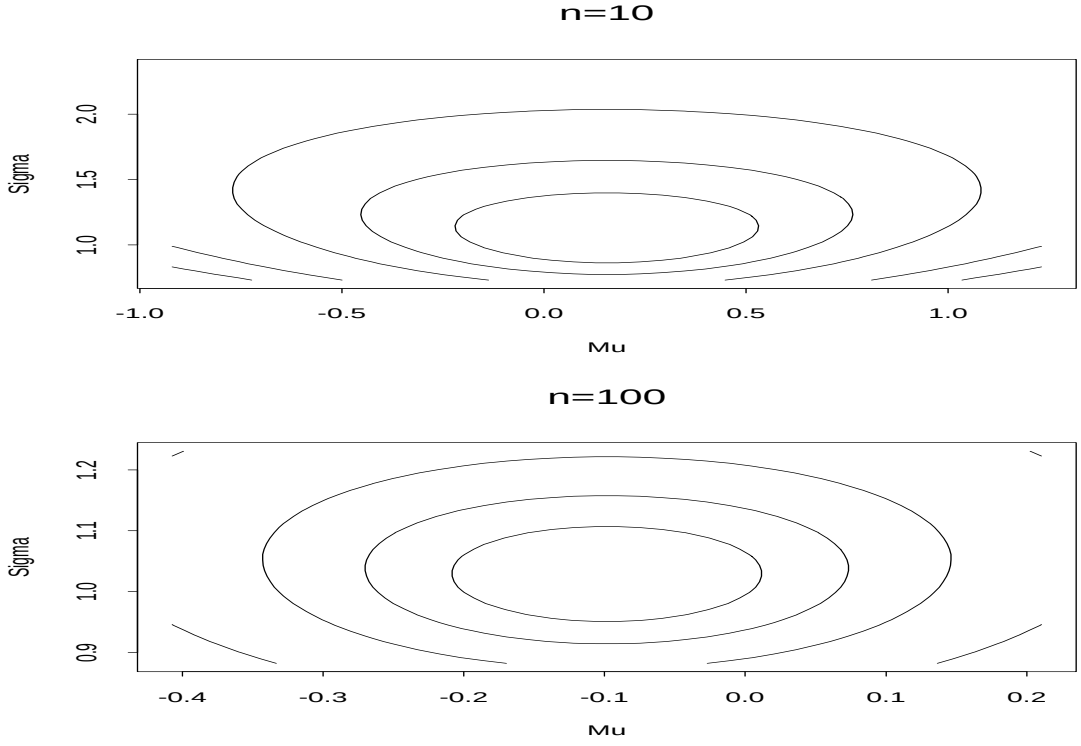
$$\begin{bmatrix} \frac{-n}{\sigma^2} & \frac{-2\sum (X_i - \mu)}{\sigma^3} \\ \frac{-2\sum (X_i - \mu)}{\sigma^3} & \frac{-3\sum (X_i - \mu)^2}{\sigma^4} + \frac{n}{\sigma^2} \end{bmatrix}$$

Plugging in the MLE gives

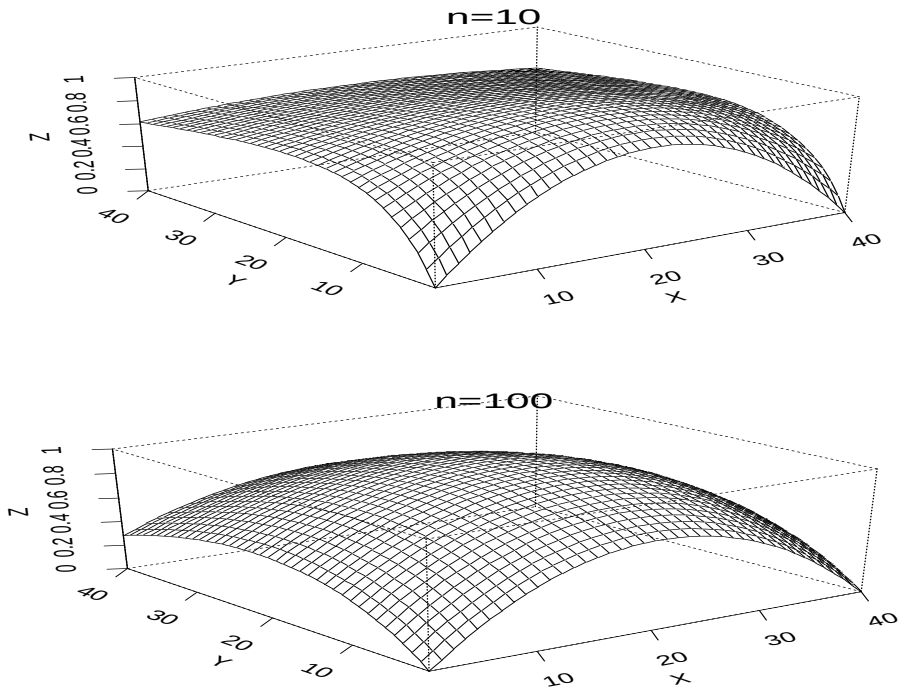
$$H(\hat{\theta}) = \begin{bmatrix} \frac{-n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{-2n}{\hat{\sigma}^2} \end{bmatrix}$$

This matrix is negative definite. Both its eigenvalues are negative. So $\hat{\theta}$ must be a local maximum.

Here is a contour plot of the normal log likelihood for two data sets with $n = 10$ and $n = 100$.



Here are perspective plots of the same.



Notice that the contours are quite ellipsoidal for the larger sample size.

Examples

N(μ, σ^2)

There is a unique root of the likelihood equations. It is a global maximum.

[Remark: Suppose we had called $\tau = \sigma^2$ the parameter. The score function would still have two components with the first component being the same as before but now the second component is

$$\frac{\partial}{\partial \tau}\ell = \frac{\sum (X_i - \mu)^2}{2\tau^2} - \frac{n}{2\tau}$$

Setting the new likelihood equations equal to 0 still gives

$$\hat{\tau} = \hat{\sigma}^2$$

This is a general **invariance** *(or* **equivariance***) principal. If $\phi = g(\theta)$ is some reparametrization of a model (a one to one relabelling of the parameter values) then $\hat{\phi} = g(\hat{\theta})$. We will see that this does not apply to other estimators.]*

Cauchy: location θ

There is at least 1 root of the likelihood equations but often several more. One of the roots is a global maximum, others, if they exist may be local minima or maxima.

Binomial(n, θ)

If $X = 0$ or $X = n$ there is no root of the likelihood equations; in this case the likelihood is monotone. For other values of X there is a unique root, a global maximum. The global maximum is at $\hat{\theta} = X/n$ even if $X = 0$ or n.

The 2 parameter exponential

The density is

$$f(x;\alpha,\beta) = \frac{1}{\beta} e^{-(x-\alpha)/\beta} 1(x > \alpha)$$

The resulting log-likelihood is $-\infty$ for $\alpha > \min\{X_1, \ldots, X_m n\}$ and otherwise is

$$\ell(\alpha,\beta) = -n\log(\beta) - \sum (X_i - \alpha)/\beta$$

As a function of α this is increasing till α reaches

$$\hat{\alpha} = X_{(1)} = \min\{X_1, \ldots, X_m n\}$$

which gives the MLE of α. Now plug in this value $\hat{\alpha}$ for α and get the so-called profile likelihood for β:

$$\ell_{\textit{profile}}(\beta) = -n\log(\beta) - \sum (X_i - X_{(1)})/\beta$$

Take the β derivative and set it equal to 0 to get

$$\hat{\beta} = \sum (X_i - X_{(1)})/n$$

Notice that the MLE $\hat{\theta} = (\hat{\alpha}, \hat{\beta})$ does not *solve the likelihood equations; we had to look at the edge of the possible parameter space. The parameter α is called a* support *or* truncation *parameter. ML methods behave oddly in problems with such parameters.*

Three parameter Weibull

The density in question is

$$f(x;\alpha,\beta,\gamma) = \frac{1}{\beta}\left(\frac{x-\alpha}{\beta}\right)^{\gamma-1} \exp[-\{(x-\alpha)/\beta\}^{\gamma}] 1(x > \alpha)$$

There are 3 derivatives to take to solve the likelihood equations. Setting the β derivative equal to 0 gives the equation

$$\hat{\beta}(\alpha,\gamma) = \left[\sum (X_i - \alpha)^{\gamma}/n\right]^{1/\gamma}$$

where we use the notation $\hat{\beta}(\alpha,\gamma)$ to indicate that the MLE of β could be found by finding the MLEs of the other two parameters and then plugging in to the formula above. It is not possible to find explicitly the remaining two parameters; numerical methods are needed. However, you can see that putting $\gamma < 1$ and letting $\alpha \to X_{(1)}$ will make the log likelihood go to ∞. The MLE is not uniquely defined, then, since any $\gamma < 1$ and any β will do.

If the true value of γ is more than 1 then the probability that there is a root of the likelihood equations is high; in this case there must be two more roots: a local maximum and a saddle point! For a true value of $\gamma > 0$ the theory we detail below applies to the local maximum and not to the global maximum of the likelihood equations.

Large Sample Theory

We now study the approximate behaviour of $\hat{\theta}$ by studying the function U. Notice first that U is a sum of independent random variables.

Theorem*: If $Y_1, Y_2, \ldots$ are iid with mean μ then*

$$\frac{\sum Y_i}{n} \to \mu$$

This is called the law of large numbers. The strong law says

$$P(\lim \frac{\sum Y_i}{n} = \mu) = 1$$

and the weak law that

$$\lim P(|\frac{\sum Y_i}{n} - \mu| > \epsilon) = 0$$

For iid Y_i the stronger conclusion holds but for our heuristics we will ignore the differences between these notions.

Now suppose that θ_0 is the true value of θ. Then

$$U(\theta)/n \to \mu(\theta)$$

where

$$\begin{aligned}\mu(\theta) =& E_{\theta_0}\left[\frac{\partial \log f}{\partial \theta}(X_i, \theta)\right] \\ =& \int \frac{\partial \log f}{\partial \theta}(x, \theta) f(x, \theta_0) dx\end{aligned}$$

Consider as an example the case of $N(\mu, 1)$ data where

$$U(\mu)/n = \sum (X_i - \mu)/n = \bar{X} - \mu$$

If the true mean is μ_0 then $\bar{X} \to \mu_0$ and

$$U(\mu)/n \to \mu_0 - \mu$$

If we think of a $\mu < \mu_0$ we see that the derivative of $\ell(\mu)$ is likely to be positive so that ℓ increases as we increase μ. For μ more than μ_0 the derivative is probably negative and so ℓ tends to be decreasing for $\mu > 0$. It follows that ℓ is likely to be maximized close to μ_0.

Now we repeat these ideas for a more general case. We study the random variable $\log[f(X_i, \theta)/f(X_i, \theta_0)]$. *You know the inequality*

$$E(X)^2 \leq E(X^2)$$

(because the difference is $Var(X) \geq 0$. *This inequality has the following generalization, called Jensen's inequality. If* g *is a convex function (non-negative second derivative, roughly) then*

$$g(E(x)) \leq E(g(X))$$

The inequality above has $g(x) = x^2$. *We use* $g(x) = -\log(x)$ *which is convex because* $g''(x) = x^{-2} > 0$. *We get*

$$-\log(E_{\theta_0}[f(X_i, \theta)/f(X_i, \theta_0)] \leq E_{\theta_0}[-\log\{f(X_i, \theta)/f(X_i, \theta_0)\}]$$

But

$$\begin{aligned} E_{\theta_0}[f(X_i, \theta)/f(X_i, \theta_0)] &= \int \frac{f(x, \theta)}{f(x, \theta_0)} f(x, \theta_0) dx \\ &= \int f(x, \theta) dx \\ &= 1 \end{aligned}$$

We can reassemble the inequality and this calculation to get

$$E_{\theta_0}[\log\{f(X_i, \theta)/f(X_i, \theta_0)\}] \leq 0$$

It is possible to prove that the inequality is strict unless the θ *and* θ_0 *densities are actually the same. Let* $\mu(\theta) < 0$ *be this expected value. Then for each* θ *we find*

$$\begin{aligned} n^{-1}[\ell(\theta) - \ell(\theta_0)] &= n^{-1} \sum \log[f(X_i, \theta)/f(X_i, \theta_0)] \\ &\to \mu(\theta) \end{aligned}$$

This proves that the likelihood is probably higher at θ_0 *than at any other single* θ. *This idea can often be stretched to prove that the MLE is* **consistent**.

Definition *A sequence* $\hat{\theta}_n$ *of estimators of* θ *is consistent if* $\hat{\theta}_n$ *converges weakly (or strongly) to* θ.

Proto theorem*: In regular problems the MLE* $\hat{\theta}$ *is consistent.*

Now let us study the shape of the log likelihood near the true value of $\hat{\theta}$ *under the assumption that* $\hat{\theta}$ *is a root of the likelihood equations close to* θ_0. *We use Taylor expansion to write, for a 1 dimensional parameter* θ,

$$\begin{aligned} U(\hat{\theta}) &= 0 \\ &= U(\theta_0) + U'(\theta_0)(\hat{\theta} - \theta_0) + U''(\tilde{\theta})(\hat{\theta} - \theta_0)^2/2 \end{aligned}$$

for some $\tilde{\theta}$ between θ_0 and $\hat{\theta}$. (This form of the remainder in Taylor's theorem is not valid for multivariate θ.) The derivatives of U are each sums of n terms and so should be both proportional to n in size. The second derivative is multiplied by the square of the small number $\hat{\theta} - \theta_0$ so should be negligible compared to the first derivative term. If we ignore the second derivative term we get

$$-U'(\theta_0)(\hat{\theta} - \theta_0) \approx U(\theta_0)$$

Now let's look at the terms U and U'.

In the normal case

$$U(\theta_0) = \sum (X_i - \mu_0)$$

has a normal distribution with mean 0 and variance n (SD $\sqrt{n}$). The derivative is simply

$$U'(\mu) = -n$$

and the next derivative U'' is 0. We will analyze the general case by noticing that both U and U' are sums of iid random variables. Let

$$U_i = \frac{\partial \log f}{\partial \theta}(X_i, \theta_0)$$

and

$$V_i = -\frac{\partial^2 \log f}{\partial \theta^2}(X_i, \theta)$$

In general, $U(\theta_0) = \sum U_i$ has mean 0 and approximately a normal distribution. Here is how we check that:

$$\begin{aligned}
E_{\theta_0}(U(\theta_0)) &= nE_{\theta_0}(U_1) \\
&= n \int \frac{\partial \log(f(x,\theta))}{\partial \theta}(x, \theta_0) f(x, \theta_0) dx \\
&= n \int \frac{\partial f / \partial \theta(x, \theta_0)}{f(x, \theta_0)} \theta f(x, \theta_0) dx \\
&= n \int \frac{\partial f}{\partial \theta}(x, \theta_0) dx \\
&= n \frac{\partial}{\partial \theta} \int f(x, \theta) dx \Big|_{\theta=\theta_0} \\
&= n \frac{\partial}{\partial \theta} 1 \\
&= 0
\end{aligned}$$

Notice that I have interchanged the order of differentiation and integration at one point. This step is usually justified by applying the dominated convergence theorem to

the definition of the derivative. The same tactic can be applied by differentiating the identity which we just proved

$$\int \frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta) dx = 0$$

Taking the derivative of both sides with respect to θ and pulling the derivative under the integral sign again gives

$$\int \frac{\partial}{\partial \theta}\left[\frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta)\right] dx = 0$$

Do the derivative and get

$$\begin{aligned} -\int \frac{\partial^2 \log(f)}{\partial \theta^2} f(x,\theta) dx &= \int \frac{\partial \log f}{\partial \theta}(x,\theta) \frac{\partial f}{\partial \theta}(x,\theta) dx \\ &= \int \left[\frac{\partial \log f}{\partial \theta}(x,\theta)\right]^2 f(x,\theta) dx \end{aligned}$$

Definition*: The* **Fisher Information** *is*

$$I(\theta) = -E_\theta(U'(\theta)) = nE_{\theta_0}(V_1)$$

We refer to $\mathcal{I}(\theta_0) = E_{\theta_0}(V_1)$ as the information in 1 observation.

The idea is that I is a measure of how curved the log likelihood tends to be at the true value of θ. Big curvature means precise estimates. Our identity above is

$$I(\theta) = Var_\theta(U(\theta)) = n\mathcal{I}(\theta)$$

Now we return to our Taylor expansion approximation

$$-U'(\theta_0)(\hat{\theta} - \theta_0) \approx U(\theta_0)$$

and study the two appearances of U.

We have shown that $U = \sum U_i$ is a sum of iid mean 0 random variables. The central limit theorem thus proves that

$$n^{-1/2} U(\theta) \Rightarrow N(0, \sigma^2)$$

where $\sigma^2 = Var(U_i) = E(V_i) = \mathcal{I}(\theta)$.

Next observe that

$$-U'(\theta) = \sum V_i$$

where again

$$V_i = -\frac{\partial U_i}{\partial \theta}$$

The law of large numbers can be applied to show

$$-U'(\theta_0)/n \to E_{\theta_0}[V_1] = \mathcal{I}(\theta_0)$$

Now manipulate our Taylor expansion as follows

$$n^{1/2}(\hat{\theta} - \theta_0) \approx \left[\frac{\sum V_i}{n}\right]^{-1} \frac{\sum U_i}{\sqrt{n}}$$

Apply Slutsky's Theorem to conclude that the right hand side of this converges in distribution to $N(0, \sigma^2/\mathcal{I}(\theta)^2)$ which simplifies, because of the identities, to $N(0, 1/\mathcal{I}(\theta)$.

Summary

In regular families:

- *Under strong regularity conditions Jensen's inequality can be used to demonstrate that $\hat{\theta}$ which maximizes ℓ globally is consistent and that this $\hat{\theta}$ is a root of the likelihood equations.*
- *It is generally easier to study ℓ only close to θ_0. For instance define A to be the event that ℓ is concave on the set of θ such that $|\theta - \theta_0| < \delta$ and the likelihood equations have a unique root in that set. Under weaker conditions than the previous case we can prove that there is a $\delta > 0$ such that*

  $$P(A) \to 1$$

  In that case we can prove that the root $\hat{\theta}$ of the likelihood equations mentioned in the definition of A is consistent.
- *Sometimes we can only get an even weaker conclusion. Define B to be the event that $\ell(\theta)$ is concave for $n^{1/2}|\theta - \theta_0| < L$ and there is a unique root of ℓ over this range. Again this root is consistent but there might be other consistent roots of the likelihood equations.*
- *Under any of these scenarios there is a consistent root of the likelihood equations which is definitely the closest to the true value θ_0. This root $\hat{\theta}$ has the property*

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

We usually simply say that the MLE is consistent and asymptotically normal with an asymptotic variance which is the inverse of the Fisher information. This assertion is actually valid for vector valued θ where now I is a matrix with ijth entry

$$I_i j = -E\left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right)$$

Estimating Equations

The same ideas arise in almost any model where estimates are derived by solving some equation. As an example I sketch large sample theory for **Generalized Linear Models**.

Suppose that for $i = 1, \ldots, n$ *we have observations of the numbers of cancer cases* Y_i *in some group of people characterized by values* x_i *of some covariates. You are supposed to think of* x_i *as containing variables like age, or a dummy for sex or average income or ... A parametric regression model for the* Y_i *might postulate that* Y_i *has a Poisson distribution with mean* μ_i *where the mean* μ_i *depends somehow on the covariate values. Typically we might assume that* $g(\mu_i) = \beta 0 + x_i\beta$ *where* g *is a so-called* **link** *function, often for this case* $g(\mu) = \log(\mu)$ *and* $x_i\beta$ *is a matrix product with* x_i *written as a row vector and* μ *a column vector. This is supposed to function as a "linear regression model with Poisson errors". I will do as a special case* $\log(\mu_i) = \beta x_i$ *where* x_i *is a scalar.*

The log likelihood is simply

$$\ell(\beta) = \sum (Y_i \log(\mu_i) - \mu_i)$$

ignoring irrelevant factorials. The score function is, since $\log(\mu_i) = \beta x_i$,

$$U(\beta) = \sum (Y_i x_i - x_i \mu_i) = \sum x_i (Y_i - \mu_i)$$

(Notice again that the score has mean 0 when you plug in the true parameter value.) The key observation, however, is that it is not necessary to believe that Y_i *has a Poisson distribution to make solving the equation* $U = 0$ *sensible. Suppose only that* $\log(E(Y_i)) = x_i\beta$. *Then we have assumed that*

$$E_\beta(U(\beta)) = 0$$

This was the key condition in proving that there was a root of the likelihood equations which was consistent and here it is what is needed, roughly, to prove that the equation $U(\beta) = 0$ *has a consistent root* $\hat{\beta}$. *Ignoring higher order terms in a Taylor expansion will give*

$$V(\beta)(\hat{\beta} - \beta) \approx U(\beta)$$

where $V = -U'$. *In the MLE case we had identities relating the expectation of* V *to the variance of* U. *In general here we have*

$$Var(U) = \sum x_i^2 Var(Y_i)$$

If Y_i *is Poisson with mean* μ_i *(and so* $Var(Y_i) = \mu_i$*) this is*

$$Var(U) = \sum x_i^2 \mu_i$$

Moreover we have

$$V_i = x_i^2 \mu_i$$

and so

$$V(\beta) = \sum x_i^2 \mu_i$$

The central limit theorem (the Lyapunov kind) will show that $U(\beta)$ *has an approximate normal distribution with variance* $\sigma_U^2 = \sum x_i^2 Var(Y_i)$ *and so*

$$\hat{\beta} - \beta \approx N(0, \sigma_U^2/(\sum x_i^2 \mu_i)^2)$$

If $Var(Y_i) = \mu_i$*, as it is for the Poisson case, the asymptotic variance simplifies to* $1/\sum x_i^2 \mu_i$*.*

Notice that other estimating equations are possible. People suggest alternatives very often. If w_i *is any set of deterministic weights (even possibly depending on* μ_i *then we could define*

$$U(\beta) = \sum w_i (Y_i - \mu_i)$$

and still conclude that $U = 0$ *probably has a consistent root which has an asymptotic normal distribution. This idea is being used all over the place these days: see, for example Zeger and Liang's Generalized estimating equations (GEE) which the econometricians call Generalized Method of Moments.*

Problems with maximum likelihood

(a) *In problems with many parameters the approximations don't work very well and maximum likelihood estimators can be far from the right answer. See your homework for the Neyman Scott example where the MLE is not consistent.*

(b) *When there are multiple roots of the likelihood equation you must choose the right root. To do so you might start with a different consistent estimator and then apply some iterative scheme like Newton Raphson to the likelihood equations to find the MLE. It turns out not many steps of NR are generally required if the starting point is a reasonable estimate.*

Finding (good) preliminary Point Estimates

Method of Moments

Basic strategy: set sample moments equal to population moments and solve for the parameters.

Definition*: The* r^{th} *sample moment (about the origin) is*

$$\frac{1}{n} \sum_{i=1}^{n} X_i^r$$

The r^{th} *population moment is*

$$\mathrm{E}(X^r)$$

*(***Central** *moments are*

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

and

$$\mathrm{E}\left[(X-\mu)^r\right] .$$

Large Sample Theory of the MLE

Application of the law of large numbers to the likelihood function:

The log likelihood ratio for θ to θ_0 is

$$\ell(\theta) - \ell(\theta_0) = \sum L_i$$

where $L_i = \log[f(X_i,\theta)/f(X_i,\theta_0)]$. We proved that $\mu_\theta \equiv E_{\theta_0}(L_i) < 0$. Then

$$\frac{\ell(\theta) - \ell(\theta_0)}{n} \to \mu(\theta)$$

so

$$P_{\theta_0}(\ell(\theta) < \ell(\theta_0)) \to 1 .$$

Theorem*: In regular problems the MLE $\hat{\theta}$ is consistent.*

Now let us study the shape of the log likelihood near the true value of $\hat{\theta}$ under the assumption that $\hat{\theta}$ is a root of the likelihood equations close to θ_0. We use Taylor expansion to write, for a 1 dimensional parameter θ,

$$\begin{aligned} U(\hat{\theta}) &= 0 \\ &= U(\theta_0) + U'(\theta_0)(\hat{\theta}-\theta_0) + U''(\tilde{\theta})(\hat{\theta}-\theta_0)^2/2 \end{aligned}$$

for some $\tilde{\theta}$ between θ_0 and $\hat{\theta}$. (This form of the remainder in Taylor's theorem is not valid for multivariate θ.) The derivatives of U are each sums of n terms and so should be both proportional to n in size. The second derivative is multiplied by the square of the small number $\hat{\theta}-\theta_0$ so should be negligible compared to the first derivative term. If we ignore the second derivative term we get

$$-U'(\theta_0)(\hat{\theta}-\theta_0) \approx U(\theta_0)$$

Now let's look at the terms U and U'.

In the normal case

$$U(\theta_0) = \sum (X_i - \mu_0)$$

has a normal distribution with mean 0 and variance n (SD $\sqrt{n}$). The derivative is simply

$$U'(\mu) = -n$$

and the next derivative U'' is 0. We will analyze the general case by noticing that both U and U' are sums of iid random variables. Let

$$U_i = \frac{\partial \log f}{\partial \theta}(X_i, \theta_0)$$

and

$$V_i = -\frac{\partial^2 \log f}{\partial \theta^2}(X_i, \theta)$$

In general, $U(\theta_0) = \sum U_i$ has mean 0 and approximately a normal distribution. Here is how we check that:

$$\begin{aligned}
E_{\theta_0}(U(\theta_0)) &= nE_{\theta_0}(U_1) \\
&= n \int \frac{\partial \log(f(x,\theta))}{\partial \theta}(x, \theta_0) f(x, \theta_0) dx \\
&= n \int \frac{\partial f / \partial \theta(x, \theta_0)}{f(x, \theta_0)} \theta f(x, \theta_0) dx \\
&= n \int \frac{\partial f}{\partial \theta}(x, \theta_0) dx \\
&= n \frac{\partial}{\partial \theta} \int f(x, \theta) dx \Big|_{\theta=\theta_0} \\
&= n \frac{\partial}{\partial \theta} 1 \\
&= 0
\end{aligned}$$

Notice that I have interchanged the order of differentiation and integration at one point. This step is usually justified by applying the dominated convergence theorem to the definition of the derivative. The same tactic can be applied by differentiating the identity which we just proved

$$\int \frac{\partial \log f}{\partial \theta}(x, \theta) f(x, \theta) dx = 0$$

Taking the derivative of both sides with respect to θ and pulling the derivative under the integral sign again gives

$$\int \frac{\partial}{\partial \theta} \left[\frac{\partial \log f}{\partial \theta}(x, \theta) f(x, \theta) \right] dx = 0$$

Do the derivative and get

$$\begin{aligned}
-\int \frac{\partial^2 \log(f)}{\partial \theta^2} f(x, \theta) dx &= \int \frac{\partial \log f}{\partial \theta}(x, \theta) \frac{\partial f}{\partial \theta}(x, \theta) dx \\
&= \int \left[\frac{\partial \log f}{\partial \theta}(x, \theta) \right]^2 f(x, \theta) dx
\end{aligned}$$

Definition: *The* **Fisher Information** *is*

$$I(\theta) = -E_\theta(U'(\theta)) = nE_{\theta_0}(V_1)$$

We refer to $\mathcal{I}(\theta_0) = E_{\theta_0}(V_1)$ *as the information in 1 observation.*

The idea is that I *is a measure of how curved the log likelihood tends to be at the true value of* θ*. Big curvature means precise estimates. Our identity above is*

$$I(\theta) = Var_\theta(U(\theta)) = n\mathcal{I}(\theta)$$

Now we return to our Taylor expansion approximation

$$-U'(\theta_0)(\hat{\theta} - \theta_0) \approx U(\theta_0)$$

and study the two appearances of U*.*

We have shown that $U = \sum U_i$ *is a sum of iid mean 0 random variables. The central limit theorem thus proves that*

$$n^{-1/2}U(\theta) \Rightarrow N(0, \sigma^2)$$

where $\sigma^2 = Var(U_i) = E(V_i) = \mathcal{I}(\theta)$*.*

Next observe that

$$-U'(\theta) = \sum V_i$$

where again

$$V_i = -\frac{\partial U_i}{\partial \theta}$$

The law of large numbers can be applied to show

$$-U'(\theta_0)/n \to E_{\theta_0}[V_1] = \mathcal{I}(\theta_0)$$

Now manipulate our Taylor expansion as follows

$$n^{1/2}(\hat{\theta} - \theta_0) \approx \left[\frac{\sum V_i}{n}\right]^{-1} \frac{\sum U_i}{\sqrt{n}}$$

Apply Slutsky's Theorem to conclude that the right hand side of this converges in distribution to $N(0, \sigma^2/\mathcal{I}(\theta)^2)$ *which simplifies, because of the identities, to* $N(0, 1/\mathcal{I}(\theta)$*.*

Summary

In regular families:

- *Under strong regularity conditions Jensen's inequality can be used to demonstrate that* $\hat{\theta}$ *which maximizes* ℓ *globally is consistent and that this* $\hat{\theta}$ *is a root of the likelihood equations.*

- *It is generally easier to study ℓ only close to θ_0. For instance define A to be the event that ℓ is concave on the set of θ such that $|\theta - \theta_0| < \delta$ and the likelihood equations have a unique root in that set. Under weaker conditions than the previous case we can prove that there is a $\delta > 0$ such that*

  $$P(A) \to 1$$

  In that case we can prove that the root $\hat{\theta}$ of the likelihood equations mentioned in the definition of A is consistent.
- *Sometimes we can only get an even weaker conclusion. Define B to be the event that $\ell(\theta)$ is concave for $n^{1/2}|\theta - \theta_0| < L$ and there is a unique root of ℓ over this range. Again this root is consistent but there might be other consistent roots of the likelihood equations.*
- *Under any of these scenarios there is a consistent root of the likelihood equations which is definitely the closest to the true value θ_0. This root $\hat{\theta}$ has the property*

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

We usually simply say that the MLE is consistent and asymptotically normal with an asymptotic variance which is the inverse of the Fisher information. This assertion is actually valid for vector valued θ where now I is a matrix with $ijth$ entry

$$I_i j = -E\left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}\right)$$

Estimating Equations

The same ideas arise in almost any model where estimates are derived by solving some equation. As an example I sketch large sample theory for **Generalized Linear Models**.

Suppose that for $i = 1, \ldots, n$ we have observations of the numbers of cancer cases Y_i in some group of people characterized by values x_i of some covariates. You are supposed to think of x_i as containing variables like age, or a dummy for sex or average income or ... A parametric regression model for the Y_i might postulate that Y_i has a Poisson distribution with mean μ_i where the mean μ_i depends somehow on the covariate values. Typically we might assume that $g(\mu_i) = \beta 0 + x_i \beta$ where g is a so-called **link** *function, often for this case $g(\mu) = \log(\mu)$ and $x_i \beta$ is a matrix product with x_i written as a row vector and μ a column vector. This is supposed to function as a "linear regression model with Poisson errors". I will do as a special case $\log(\mu_i) = \beta x_i$ where x_i is a scalar.*

The log likelihood is simply

$$\ell(\beta) = \sum (Y_i \log(\mu_i) - \mu_i)$$

ignoring irrelevant factorials. The score function is, since $\log(\mu_i) = \beta x_i$,

$$U(\beta) = \sum (Y_i x_i - x_i \mu_i) = \sum x_i (Y_i - \mu_i)$$

(Notice again that the score has mean 0 when you plug in the true parameter value.) The key observation, however, is that it is not necessary to believe that Y_i has a Poisson distribution to make solving the equation $U = 0$ sensible. Suppose only that $\log(E(Y_i)) = x_i\beta$. Then we have assumed that

$$E_\beta(U(\beta)) = 0$$

This was the key condition in proving that there was a root of the likelihood equations which was consistent and here it is what is needed, roughly, to prove that the equation $U(\beta) = 0$ has a consistent root $\hat{\beta}$. Ignoring higher order terms in a Taylor expansion will give

$$V(\beta)(\hat{\beta} - \beta) \approx U(\beta)$$

where $V = -U'$. In the MLE case we had identities relating the expectation of V to the variance of U. In general here we have

$$Var(U) = \sum x_i^2 Var(Y_i)$$

If Y_i is Poisson with mean μ_i (and so $Var(Y_i) = \mu_i$) this is

$$Var(U) = \sum x_i^2 \mu_i$$

Moreover we have

$$V_i = x_i^2 \mu_i$$

and so

$$V(\beta) = \sum x_i^2 \mu_i$$

The central limit theorem (the Lyapunov kind) will show that $U(\beta)$ has an approximate normal distribution with variance $\sigma_U^2 = \sum x_i^2 Var(Y_i)$ and so

$$\hat{\beta} - \beta \approx N(0, \sigma_U^2/(\sum x_i^2 \mu_i)^2)$$

If $Var(Y_i) = \mu_i$, as it is for the Poisson case, the asymptotic variance simplifies to $1/\sum x_i^2 \mu_i$.

Notice that other estimating equations are possible. People suggest alternatives very often. If w_i is any set of deterministic weights (even possibly depending on μ_i then we could define

$$U(\beta) = \sum w_i(Y_i - \mu_i)$$

and still conclude that $U = 0$ probably has a consistent root which has an asymptotic normal distribution. This idea is being used all over the place these days: see, for example Zeger and Liang's Generalized estimating equations (GEE) which the econometricians call Generalized Method of Moments.

Problems with maximum likelihood

(a) In problems with many parameters the approximations don't work very well and maximum likelihood estimators can be far from the right answer. See your homework for the Neyman Scott example where the MLE is not consistent.

(b) When there are multiple roots of the likelihood equation you must choose the right root. To do so you might start with a different consistent estimator and then apply some iterative scheme like Newton Raphson to the likelihood equations to find the MLE. It turns out not many steps of NR are generally required if the starting point is a reasonable estimate.

Finding (good) preliminary Point Estimates

Method of Moments

Basic strategy: set sample moments equal to population moments and solve for the parameters.

Definition*: The r^{th} sample moment (about the origin) is*

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r$$

The r^{th} population moment is

$$\mathrm{E}(X^r)$$

*(***Central** *moments are*

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

and

$$\mathrm{E}\left[(X-\mu)^r\right].$$

If we have p parameters we can estimate the parameters $\theta_1, \ldots, \theta_p$ by solving the system of p equations:

$$\mu_1 = \bar{X}$$
$$\mu_2' = \overline{X^2}$$

and so on to

$$\mu_p' = \overline{X^p}$$

You need to remember that the population moments μ_k' will be formulas involving the parameters.

Gamma Example

The Gamma(α, β) density is

$$f(x;\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\frac{x}{\beta}\right]1(x>0)$$

and has

$$\mu_1 = \alpha\beta$$

and

$$\mu_2' = \alpha\beta^2$$

This gives the equations

$$\alpha\beta = \overline{X}$$
$$\alpha\beta^2 = \overline{X^2}$$

Divide the second by the first to find the method of moments estimate of β is

$$\tilde{\beta} = \overline{X^2}/\overline{X}$$

Then from the first equation get

$$\tilde{\alpha} = \overline{X}/\tilde{\beta} = (\overline{X})^2/\overline{X^2}$$

The equations are much easier to solve than the likelihood equations which involve the function

$$\psi(\alpha) = \frac{d}{d\alpha}\log(\Gamma(\alpha))$$

called the digamma function.

Large Sample Theory of the MLE

Theorem*: Under suitable regularity conditions there is a unique consistent root $\hat{\theta}$ of the likelihood equations. This root has the property*

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

In general (not just iid cases)

$$\sqrt{I(\theta_0)}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{I(\hat{\theta})}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{V(\theta_0)}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{V(\hat{\theta})}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$

where $V = -\ell''$ is the so-called observed information*, the negative second derivative of the log-likelihood.*

Note*: If the square roots are replaced by matrix square roots we can let θ be vector valued and get $MVN(0, I)$ as the limit law.*

Why bother with all these different forms? We actually use the limit laws to test hypotheses and compute confidence intervals. We test $H_o : \theta - \theta_0$ using one of the four

quantities as the test statistic. To find confidence intervals we use the quantities as pivots. *For example the second and fourth limits above lead to confidence intervals*

$$\hat{\theta} \pm z_{\alpha/2}/\sqrt{I(\hat{\theta})}$$

and

$$\hat{\theta} \pm z_{\alpha/2}/\sqrt{V(\hat{\theta})}$$

respectively. The other two are more complicated. For iid $N(0, \sigma^2)$ data we have

$$V(\sigma) = \frac{3\sum X_i^2}{\sigma^4} - \frac{n}{\sigma^2}$$

and

$$I(\sigma) = \frac{2n}{\sigma^2}$$

The first line above then justifies confidence intervals for σ computed by finding all those σ for which

$$\left| \frac{\sqrt{2n}(\hat{\sigma} - \sigma}{\sigma} \right| \leq z_{\alpha/2}$$

A similar interval can be derived from the third expression, though this is much more complicated.

Estimating Equations

An estimating equation is unbiased if

$$E_\theta(U(\theta)) = 0$$

Theorem*: Suppose $\hat{\theta}$ is a consistent root of the unbiased estimating equation*

$$U(\theta) = 0.$$

Let $V = -U'$. Suppose there is a sequence of constants $B(\theta)$ such that

$$V(\theta)/B(\theta) \to 1$$

and let

$$A(\theta) = Var_\theta(U(\theta))$$

and

$$C(\theta) = B(\theta)A^{-1}(\theta)B(\theta).$$

Then

$$\sqrt{C(\theta_0)}(\hat{\theta} - \theta_0) \Rightarrow N(0, 1)$$

$$\sqrt{C(\hat{\theta})}(\hat{\theta} - \theta_0) \Rightarrow N(0, 1)$$

There are other ways to estimate A, B and C which all lead to the same conclusions and there are multivariate extensions for which the square roots are matrix square roots.

Problems with maximum likelihood

(a) *In problems with many parameters the approximations don't work very well and maximum likelihood estimators can be far from the right answer. See your homework for the Neyman Scott example where the MLE is not consistent.*

(b) *When there are multiple roots of the likelihood equation you must choose the right root. To do so you might start with a different consistent estimator and then apply some iterative scheme like Newton Raphson to the likelihood equations to find the MLE. It turns out not many steps of NR are generally required if the starting point is a reasonable estimate.*

Finding (good) preliminary Point Estimates

Method of Moments

Basic strategy: set sample moments equal to population moments and solve for the parameters.

Definition*: The r^{th} sample moment (about the origin) is*

$$\frac{1}{n}\sum_{i=1}^{n} X_i^r$$

The r^{th} population moment is

$$\mathrm{E}(X^r)$$

*(***Central** *moments are*

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

and

$$\mathrm{E}\left[(X-\mu)^r\right] .$$

If we have p parameters we can estimate the parameters $\theta_1, \ldots, \theta_p$ by solving the system of p equations:

$$\mu_1 = \bar{X}$$

$$\mu_2' = \overline{X^2}$$

and so on to

$$\mu_p' = \overline{X^p}$$

You need to remember that the population moments μ_k' will be formulas involving the parameters.

Gamma Example

The Gamma(α, β) density is

$$f(x;\alpha,\beta) = \frac{1}{\beta\Gamma(\alpha)}\left(\frac{x}{\beta}\right)^{\alpha-1}\exp\left[-\frac{x}{\beta}\right]1(x>0)$$

and has

$$\mu_1 = \alpha\beta$$

and

$$\mu_2' = \alpha(\alpha+1)\beta^2.$$

This gives the equations

$$\alpha\beta = \overline{X}$$
$$\alpha(\alpha+1)\beta^2 = \overline{X^2}$$

or

$$\alpha\beta = \overline{X}$$
$$\alpha\beta^2 = \overline{X^2} - \overline{X}^2$$

Divide the second equation by the first to find the method of moments estimate of β is

$$\tilde{\beta} = (\overline{X^2} - \overline{X}^2)/\overline{X}$$

Then from the first equation get

$$\tilde{\alpha} = \overline{X}/\tilde{\beta} = (\overline{X})^2/(\overline{X^2} - \overline{X}^2)$$

These equations are much easier to solve than the likelihood equations. The latter involve the function

$$\psi(\alpha) = \frac{d}{d\alpha}\log(\Gamma(\alpha))$$

called the digamma function. The score function in this problem has components

$$U_\beta = \frac{\sum X_i}{\beta^2} - n\alpha/\beta$$

and

$$U_\alpha = -n\psi(\alpha) + \sum \log(X_i) - n\log(\beta)$$

You can solve for β in terms of α to leave you trying to find a root of the equation

$$-n\psi(\alpha) + \sum \log(X_i) - n\log(\sum X_i/(n\alpha)) = 0$$

To use Newton Raphson on this you begin with the preliminary estimate $\hat{\alpha}_1 = \tilde{\alpha}$ and then compute iteratively

$$\hat{\alpha}_{k+1} = \frac{\overline{\log(X)} - \psi(\hat{\alpha}_k) - \log(\overline{X}/\hat{\alpha}_k}{1/\alpha - \psi'(\hat{\alpha}_k)}$$

until the sequence converges. Computation of ψ', the trigamma function, requires special software. Web sites like netlib *and* statlib *are good sources for this sort of thing.*

Optimality theory for point estimates

Why bother doing the Newton Raphson steps? Why not just use the method of moments estimates? The answer is that the method of moments estimates are not usually as close to the right answer as the MLEs.

Rough principle*: A good estimate $\hat{\theta}$ of θ is usually close to θ_0 if θ_0 is the true value of θ. Closer estimates, more often, are better estimates.*

This principle must be quantified if we are to "prove" that the MLE is a good estimate. In the Neyman Pearson spirit we measure average closeness.

Definition*: The* **Mean Squared Error** *(MSE) of an estimator $\hat{\theta}$ is the* **function**

$$MSE(\theta) = E_\theta[(\hat{\theta} - \theta)^2]$$

Standard identity:

$$MSE = \mathrm{Var}_\theta(\hat{\theta}) + Bias^2_{\hat{\theta}}(\theta)$$

where the bias is defined as

$$Bias_{\hat{\theta}}(\theta) = E_\theta(\hat{\theta}) - \theta\,.$$

Primitive example*: I take a coin from my pocket and toss it 6 times. I get $HTHTTT$. The MLE of the probability of heads is*

$$\hat{p} = X/n$$

where X is the number of heads. In this case I get $\hat{p} = \frac{1}{3}$.

An alternative estimate is $\tilde{p} = \frac{1}{2}$. That is, $\tilde{p}$ ignores the data and guesses the coin is fair. The MSEs of these two estimators are

$$MSE_{MLE} = \frac{p(1-p)}{6}$$

and

$$MSE_{0.5} = (p - 0.5)^2$$

If p is between 0.311 and 0.689 then the second MSE is smaller than the first. For this reason I would recommend use of $\tilde{p}$ for sample sizes this small.

Now suppose I did the same experiment with a thumbtack. The tack can land point up (U) or tipped over (O). If I get $UOUOOO$ how should I estimate p the probability of U? The mathematics is identical to the above but it seems clear that there is less reason to think $\tilde{p}$ is better than $\hat{p}$ since there is less reason to believe $0.311 \le p \le 0.689$ than with a coin.

Unbiased Estimation

The problem above illustrates a general phenomenon. An estimator can be good for some values of θ and bad for others. When comparing $\hat{\theta}$ and $\tilde{\theta}$, two estimators of θ we will say that $\hat{\theta}$ is better than $\tilde{\theta}$ if it has uniformly *smaller MSE:*

$$MSE_{\hat{\theta}}(\theta) \leq MSE_{\tilde{\theta}}(\theta)$$

for **all** *θ. Normally we also require that the inequality be strict for at least one θ.*

The definition raises the question of the existence of a best *estimate – one which is better than every other estimator. There is no such estimate. Suppose $\hat{\theta}$ were such a best estimate. Fix a θ^* in Θ and let $\tilde{p} \equiv \theta^*$. Then the MSE of $\tilde{p}$ is 0 when $\theta = \theta^*$. Since $\hat{\theta}$ is better than $\tilde{p}$ we must have*

$$MSE_{\hat{\theta}}(\theta^*) = 0$$

so that $\hat{\theta} = \theta^*$ with probability equal to 1. This makes $\hat{\theta} = \tilde{\theta}$. If there are actually two different possible values of θ this gives a contradiction; so no such $\hat{\theta}$ exists.

Principle of Unbiasedness*: A good estimate is unbiased, that is,*

$$E_\theta(\hat{\theta}) \equiv 0\,.$$

WARNING: In my view the Principle of Unbiasedness is a load of hog wash.

For an unbiased estimate the MSE is just the variance.

Definition*: An estimator $\hat{\phi}$ of a parameter $\phi = \phi(\theta)$ is* **Uniformly Minimum Variance Unbiased** *(UMVU) if, whenever $\tilde{\phi}$ is an unbiased estimate of ϕ we have*

$$\mathrm{Var}_\theta(\hat{\phi}) \leq \mathrm{Var}_\theta(\tilde{\phi})$$

We call $\hat{\phi}$ the UMVUE. ('E' is for Estimator.)

The point of having $\phi(\theta)$ is to study problems like estimating μ when you have two parameters like μ and σ for example.

Cramér Rao Inequality

If $\phi(\theta) = \theta$ we can derive some information from the identity

$$E_\theta(T) \equiv \theta$$

When we worked with the score function we derived some information from the identity

$$\int f(x,\theta)dx \equiv 1$$

by differentiation and we do the same here. If $T = T(X)$ is some function of the data X which is unbiased for θ then

$$E_\theta(T) = \int T(x) f(x,\theta)dx \equiv \theta$$

Differentiate both sides to get

$$\begin{aligned} 1 &= \frac{d}{d\theta}\int T(x)f(x,\theta)dx \\ &= \int T(x)\frac{\partial}{\partial\theta}f(x,\theta)dx \\ &= \int T(x)\frac{\partial}{\partial\theta}\log(f(x,\theta))f(x,\theta)dx \\ &= E_\theta(T(X)U(\theta)) \end{aligned}$$

where U is the score function. Since we already know that the score has mean 0 we see that

$$\mathrm{Cov}_\theta(T(X),U(\theta)) = 1$$

Now remember that correlations are between -1 and 1 or

$$1 = |\mathrm{Cov}_\theta(T(X),U(\theta))| \leq \sqrt{\mathrm{Var}_\theta(T)\mathrm{Var}_\theta(U(\theta))}$$

Squaring gives the inequality

$$\mathrm{Var}_\theta(T) \geq \frac{1}{I(\theta)}$$

which is called the Cramér Rao Lower Bound. The inequality is strict unless the correlation is 1 which would require that

$$U(\theta) = A(\theta)T(X) + B(\theta)$$

for some non-random constants A and B (which might depend on θ.) This would prove that

$$\ell(\theta) = A^*(\theta)T(X) + B^*(\theta) + C(X)$$

for some further constants A^* and B^* and finally

$$f(x,\theta) = h(x)e^{A*(\theta)T(x)+B^*(\theta)}$$

for $h = e^C$.

Summary of Implications

- *You can recognize a UMVUE sometimes. If $\mathrm{Var}_\theta(T(X)) \equiv 1/I(\theta)$ then $T(X)$ is the UMVUE. For instance in the $N(\mu,1)$ example the Fisher information is n and $\mathrm{Var}(\overline{X}) = 1/n$ so that $\overline{X}$ is the UMVUE of μ.*
- *In an asymptotic sense the MLE is nearly optimal: it is nearly unbiased and (approximate) variance nearly $1/I(\theta)$.*
- *Good estimates are highly correlated with the score function.*
- *Densities of the exponential form given above are somehow special. (The form is called an exponential family.)*

- *For most problems the inequality will be strict. It is strict unless The score is an affine function of a statistic T and T (possibility divided by some constant which doesn't depend on θ) is unbiased for θ.*

What can we do to find UMVUEs when the CRLB is a strict inequality?

Example*: Suppose X has a Binomial(n, p) distribution. The score function is*

$$U(p) = \frac{1}{p(1-p)}X - \frac{n}{1-p}$$

Thus the CRLB will be strict unless $T = cX$ for some c. If we are trying to estimate p then choosing $c = n^{-1}$ does give an unbiased estimate $\hat{p} = X/n$ and $T = X/n$ achieves the CRLB so it is UMVU.

A different tactic proceeds as follows. Suppose $T(X)$ is some unbiased function of X. Then we have

$$E_p(T(X) - X/n) \equiv 0$$

because $\hat{p} = X/n$ is also unbiased. If $h(k) = T(k) - k/n$ then

$$E_p(h(X)) = \sum_{k=0}^{n} h(k)\binom{n}{k}p^k(1-p)^{n-k} \equiv 0$$

The left hand side of the $\equiv$ sign is a polynomial function of p as is the right. Thus if the left hand side is expanded out the coefficient of each power p^k is 0. The constant term occurs only in the term $k = 0$ and its coefficient is

$$h(0)\binom{n}{0} = h(0)$$

Thus $h(0) = 0$. Now $p^1 = p$ occurs only in the term $k = 1$ with coefficient $nh(1)$ so $h(1) = 0$. Since the terms with $k = 0$ or 1 are 0 the quantity p^2 occurs only in the term with $k = 2$ with coefficient

$$n(n-1)h(2)/2$$

so $h(2) = 0$. We can continue in this way to see that in fact $h(k) = 0$ for each k and so the only *unbiased function of X is X/n.*

Now a Binomial random variable is just of sum of n iid Bernoulli(p) random variables. If $Y_1, \ldots, Y_n$ are iid Bernoulli(p) then $X = \sum Y_i$ is Binomial(n, p). Could we do better by than $\hat{p} = X/n$ by trying $T(Y_1, \ldots, Y_n)$ for some other function T?

Let's consider the case $n = 2$ so that there are 4 possible values for Y_1, Y_2. If $h(Y_1, Y_2) = T(Y_1, Y_2) - [Y_1 + Y_2]/2$ then again

$$E_p(h(Y_1, Y_2)) \equiv 0$$

and we have

$$E_p(h(Y_1, Y_2)) = h(0,0)(1-p)^2 + [h(1,0) + h(0,1)]p(1-p) + h(1,1)p^2$$

This can be rewritten in the form

$$\sum_{k=0}^{n} w(k)\binom{n}{k}p^k(1-p)^{n-k}$$

where $w(0) = h(0,0)$, $w(1) = [h(1,0) + h(0,1)]/2$ *and* $w(2) = h(1,1)$. *Just as before it follows that* $w(0) = w(1) = w(2) = 0$. *This argument can be used to prove that for any unbiased estimate* $T(Y_1, \ldots, Y_n)$ *we have that the average value of* $T(y_1, \ldots, y_n)$ *over vectors* $y_1, \ldots, y_n$ *which have exactly* k *1s and* $n - k$ *0s is* k/n. *Now let's look at the variance of* T*:*

$$\begin{aligned}
\text{Var(T)} =& E_p([T(Y_1, \ldots, Y_n) - p]^2) \\
=& E_p([T(Y_1, \ldots, Y_n) - X/n + X/n - p]^2) \\
=& E_p([T(Y_1, \ldots, Y_n) - X/n]^2) \\
&+ 2E_p([T(Y_1, \ldots, Y_n) - X/n][X/n - p]) \\
&+ E_p([X/n - p]^2)
\end{aligned}$$

I claim that the cross product term is 0 which will prove that the variance of T *is the variance of* X/n *plus a non-negative quantity (which will be positive unless* $T(Y_1, \ldots, Y_n) \equiv X/n$*). We can compute the cross product term by writing*

$$\begin{aligned}
&E_p([T(Y_1, \ldots, Y_n) - X/n][X/n - p]) \\
&\qquad = \sum_{y_1, \ldots, y_n} [T(y_1, \ldots, y_n) - \sum y_i/n][\sum y_i/n - p]p^{\sum y_i}(1-p)^{n - \sum y_i}
\end{aligned}$$

We can do the sum by summing over those $y_1, \ldots, y_n$ *whose sum is an integer* x *and then summing over* x. *We get*

$$\begin{aligned}
&E_p([T(Y_1, \ldots, Y_n) - X/n][X/n - p]) \\
&\quad = \sum_{x=0}^{n} \sum_{\sum y_i = x} [T(y_1, \ldots, y_n) - \sum y_i/n][\sum y_i/n - p]p^{\sum y_i}(1-p)^{n - \sum y_i} \\
&\qquad = \sum_{x=0}^{n} \left[\sum_{\sum y_i = x} [T(y_1, \ldots, y_n) - x/n] \right] [x/n - p]p^x(1-p)^{n-x}
\end{aligned}$$

We have already shown that the sum in [] *is 0!*

This long, algebraically involved, method of proving that $\hat{p} = X/n$ *is the UMVUE of* p *is one special case of a general tactic.*

To get more insight I begin by rewriting

$$E_p(T(Y_1,\ldots,Y_n))$$
$$\sum_{x=0}^{n}\sum_{\sum y_i=x} T(y_1,\ldots,y_n)P(Y_1=y_1,\ldots,Y_n=y_n)$$
$$=\sum_{x=0}^{n}\sum_{\sum y_i=x} T(y_1,\ldots,y_n)P(Y_1=y_1,\ldots,Y_n=y_n|X=x)P(X=x)$$
$$=\sum_{x=0}^{n}\frac{\sum_{\sum y_i=x}T(y_1,\ldots,y_n)}{\binom{n}{x}}\binom{n}{x}p^x(1-p)^{n-x}$$

Notice that the large fraction in this formula is the average value of T over values of y when $\sum y_i$ is held fixed at x. Notice that the weights in this average do not depend on p. Notice that this average is actually

$$E(T(Y_1,\ldots,Y_n|X=x))$$
$$\sum_{y_1,\ldots,y_n} T(y_1,\ldots,y_n)P(Y_1=y_1,\ldots,Y_n=y_n|X=x)$$

Notice that the conditional probabilities do not depend on p. In a sequence of Binomial trials if I tell you that 5 of 17 were heads and the rest tails the actual trial numbers of the 5 Heads are chosen at random from the 17 possibilities; all of the 17 choose 5 possibilities have the same chance and this chance does not depend on p.

Notice that in this problem with data $Y_1,\ldots,Y_n$ the log likelihood is

$$\ell(p)=\sum Y_i\log(p)-(n-\sum Y_i)\log(1-p)$$

and

$$U(p)=\frac{1}{p(1-p)}X-\frac{n}{1-p}$$

as before. Again we see that the CRLB will be a strict inequality except for multiples of X. Since the only unbiased multiple of X is $\hat{p}=X/n$ we see again that $\hat{p}$ is UMVUE for p.

The Binomial(n,p) example

If $Y_1,\ldots,Y_n$ are iid Bernoulli(p) then $X=\sum Y_i$ is Binomial(n,p). We used various algebraic tactics to arrive at the following conclusions:

- *The log likelihood is a function of X only and not the actual values of $Y_1,\ldots,Y_n$.*
- *There is only one function of X, namely, $\hat{p}=X/n$ which is an unbiased estimate of p.*

- *If $T(Y_1, \ldots, Y_n)$ is an unbiased estimate of p then the average value of $T(y_1, \ldots, y_n)$ over those $y_1, \ldots, y_n$ for which $\sum y_i = x$ is x/n.*
- *The conditional distribution of T given $\sum Y_i = x$ does not depend on p.*
- *If $T(Y_1, \ldots, Y_n)$ is an unbiased estimate of p then*

$$\mathrm{Var}(T) = \mathrm{Var}(\hat{p}) + E[(T - \hat{p})^2]$$

- *$\hat{p}$ is the UMVUE of p.*

This long, algebraically involved, method of proving that $\hat{p} = X/n$ is the UMVUE of p is one special case of a general tactic.

Sufficiency

In the binomial situation the conditional distribution of the data $Y_1, \ldots, Y_n$ given X is the same for all values of θ; we say this conditional distribution is **free** *of θ.*

Definition*: A statistic $T(X)$ is sufficient for the model $\{P_\theta; \theta \in \Theta\}$ if the conditional distribution of the data X given $T = t$ is free of θ.*

Intuition*: Why do the data tell us about θ? Because different values of θ give different distributions to X. If two different values of θ correspond to the same joint density or CDF for X then we cannot, even in principle, distinguish these two values of θ by examining X. We extend this notion to the following. If two values of θ give the same conditional distribution of X given T then observing T in addition to X does not improve our ability to distinguish the two values.*

Mathematically Precise version of this intuition*: If $T(X)$ is a sufficient statistic then we can do the following. If $S(X)$ is any estimate or confidence interval or whatever for a given problem but we only know the value of T then:*

- *Generate an observation X^* (via some sort of Monte Carlo program) from the conditional distribution of X given T.*
- *Use $S(X^*)$ instead of $S(X)$. Then $S(X^*)$ has the same performance characteristics as $S(X)$ because the distribution of X^* is the same as that of X.*

You can carry out the first step **only** *if the statistic T is sufficient; otherwise you need to know the true value of θ to generate X^*.*

Example 1*: If $Y_1, \ldots, Y_n$ are iid Bernoulli(p) then given $\sum Y_i = y$ the indexes of the y successes have the same chance of being any one of the $\binom{n}{y}$ possible subsets of $\{1, \ldots, n\}$. This chance does not depend on p so $T(Y_1, \ldots, Y_n) = \sum Y_i$ is a sufficient statistic.*

Example 2*: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ then the joint distribution of $X_1, \ldots, X_n, \overline{X}$ is multivariate normal with mean vector whose entries are all μ and variance covariance matrix which can be partitioned as*

$$\begin{bmatrix} I_{n\times n} & \mathbf{1}_n/n \\ \mathbf{1}_n^t/n & 1/n \end{bmatrix}$$

where $\mathbf{1}_n$ is a column vector of n 1s and $I_{n\times n}$ is an $n \times n$ identity matrix.

You can now compute the conditional means and variances of X_i given $\overline{X}$ and use the fact that the conditional law is multivariate normal to prove that the conditional distribution of the data given $\overline{X} = x$ is multivariate normal with mean vector all of whose entries are x and variance-covariance matrix given by $I_{n\times n} - \mathbf{1}_n\mathbf{1}_n^t/n$. Since this does not depend on μ we find that $\overline{X}$ is sufficient.

WARNING: Whether or not a statistic is sufficient depends on the density function and on Θ.

Rao Blackwell Theorem

Theorem*: Suppose that $S(X)$ is a sufficient statistic for some model $\{P_\theta, \theta \in \Theta\}$. If T is an estimate of some parameter $\phi(\theta)$ then:*

(a) $E(T|S)$ is a statistic.

(b) $E(T|S)$ has the same bias as T; if T is unbiased so is $E(T|S)$.

(c) $\mathrm{Var}_\theta(E(T|S)) \leq \mathrm{Var}_\theta(T)$ and the inequality is strict unless T is a function of S.

(d) The MSE of $E(T|S)$ is no more than that of T.

Proof*: It will be useful to review conditional distributions a bit more carefully at this point. The abstract definition of conditional expectation is this:*

Definition*: $E(Y|X)$ is any function of X such that*

$$E\left[R(X)E(Y|X)\right] = E\left[R(X)Y\right]$$

for any function $R(X)$.

Definition*: $E(Y|X = x)$ is a function $g(x)$ such that*

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

Fact*: If X, Y has joint density $f_{X,Y}(x, y)$ and conditional density $f(y|x)$ then*

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

satisfies these definitions.

Proof of Fact*:*

$$\begin{aligned} E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\ &= \int\int R(x)yf(y|x)f_X(x)f_{Y|X}(y|x)dydx \\ &= \int\int R(x)yf_{X,Y}(x,y)dydx \\ &= E(R(X)Y) \end{aligned}$$

You should simply think of $E(Y|X)$ as being what you get when you average Y holding X fixed. It behaves like an ordinary expected value but where functions of X only are like constants.

Proof of the Rao Blackwell Theorem

Step 1: The definition of sufficiency is that the conditional distribution of X given S does not depend on θ. This means that $E(T(X)|S)$ does not depend on θ.

Step 2: This step hinges on the following identity (called Adam's law by Jerzy Neyman – he used to say it comes before all the others)

$$E[E(Y|X)] = E(Y)$$

which is just the definition of $E(Y|X)$ with $R(X) \equiv 1$.

From this we deduce that

$$E_\theta[E(T|S)] = E_\theta(T)$$

so that $E(T|S)$ and T have the same bias. If T is unbiased then

$$E_\theta[E(T|S)] = E_\theta(T) = \phi(\theta)$$

so that $E(T|S)$ is unbiased for ϕ.

Step 3: This relies on the following very useful decomposition. (In regression courses we say that the total sum of squares is the sum of the regression sum of squares plus the residual sum of squares.)

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

The conditional variance means

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

This identity is just a matter of squaring out the right hand side

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

and

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

Adding these together gives

$$E[Y^2 - 2YE[Y|X] + 2(E[Y|X])^2 - 2E(Y)E[Y|X] + E^2(Y)]$$

The middle term actually simplifies. First, remember that $E(Y|X)$ is a function of X so can be treated as a constant when holding X fixed. This means

$$E[Y|X]E[Y|X] = E[YE(Y|X)|X]$$

and taking expectations gives

$$E[(E[Y|X])^2] = E[E[YE(Y|X)|X]] = E[YE(Y|X)]$$

This makes the middle term above cancel with the second term. Moreover the fourth term simplifies

$$E[E(Y)E[Y|X]] = E(Y)E[E[Y|X]] = E^2(Y)$$

so that

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

We apply this to the Rao Blackwell theorem to get

$$\mathrm{Var}_\theta(T) = \mathrm{Var}_\theta(E(T|S)) + E[(T - E(T|S))^2]$$

The second term is non negative so that the variance of $E(T|S)$ must be no more than that of T and will be strictly less unless $T = E(T|S)$. This would mean that T is already a function of S. Adding the squares of he biases of T (or of $E(T|S)$ which is the same) gives the inequality for mean squared error.

Examples*:*

In the binomial problem $Y_1(1 - Y_2)$ is an unbiased estimate of $p(1 - p)$. We improve this by computing

$$E(Y_1(1 - Y_2)|X)$$

We do this in two steps. First compute

$$E(Y_1(1 - Y_2)|X = x)$$

Notice that the random variable $Y_1(1 - Y_2)$ is either 1 or 0 so its expected value is just

the probability it is equal to 1:

$$\begin{aligned}
E(Y_1(1-Y_2)|X=x) &= P(Y_1(1-Y_2)=1|X=x) \\
&= P(Y_1=1,Y_2=0|Y_1+Y_2+\cdots+Y_n=x) \\
&= \frac{P(Y_1=1,Y_2=0,Y_1+Y_2+\cdots+Y_n=x)}{P(Y_1+Y_2+\cdots+Y_n=x)} \\
&= \frac{P(Y_1=1,Y_2=0,Y_3+Y_4+\cdots+Y_n=x-1)}{\binom{n}{x}p^x(1-p)^{n-x}} \\
&= \frac{p(1-p)\binom{n-2}{x-1}p^{x-1}(1-p)^{(n-1)-(x-1)}}{\binom{n}{x}p^x(1-p)^{n-x}} \\
&= \frac{\binom{n-2}{x-1}}{\binom{n}{x}} \\
&= \frac{x(n-x)}{n(n-1)}
\end{aligned}$$

This is simply $n\hat{p}(1-\hat{p})/(n-1)$ (which can be bigger than $1/4$ which is the maximum value of $p(1-p)$).

Example*: If $X_1,\ldots,X_n$ are iid $N(\mu,1)$ then $\bar{X}$ is sufficient and X_1 is an unbiased estimate of μ. Now*

$$\begin{aligned}
E(X_1|\bar{X}) &= E[X_1-\bar{X}+\bar{X}|\bar{X}] \\
&= E[X_1-\bar{X}|\bar{X}]+\bar{X} \\
&= \bar{X}
\end{aligned}$$

which is the UMVUE.

Finding Sufficient statistics

In the binomial example the log likelihood (at least the part depending on the parameters) was seen above to be a function of X (and not of the original data $Y_1,\ldots,Y_n$ as well). In the normal example the log likelihood is, ignoring terms which don't contain μ,

$$\ell(\mu) = \mu\sum X_i - n\mu^2/2 = n\mu\bar{X} - n\mu^2/2\,.$$

These are examples of the **Factorization Criterion***:*

Theorem*: If the model for data X has density $f(x,\theta)$ then the statistic $S(X)$ is sufficient if and only if the density can be factored as*

$$f(x,\theta) = g(s(x),\theta)h(x)$$

The theorem is proved by finding a statistic $T(x)$ such that X is a one to one function of the pair S, T and applying the change of variables to the joint density of S and T. If the density factors then you get

$$f_{S,T}(s,t) = g(s,\theta)h(x(s,t))$$

from which we see that the conditional density of T given $S = s$ does not depend on θ. Thus the conditional distribution of (S,T) given S does not depend on θ and finally the conditional distribution of X given S does not depend on θ. Conversely if S is sufficient then the conditional density of T given S has no θ in it and the joint density of S, T is

$$f_S(s,\theta)f_{T|S}(t|s)$$

Apply the change of variables formula to get the density of X to be

$$f_S(s(x),\theta)f_{T|S}(t(x)|s(x))J(x)$$

where J is the Jacobian. This factors.

Example*: If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then the joint density is*

$$(2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum X_i^2/(2\sigma^2) + \mu\sum X_i/\sigma^2 - n\mu^2/(2\sigma^2)\}$$

which is evidently a function of

$$\sum X_i^2, \sum X_i$$

This pair is a sufficient statistic. You can write this pair as a bijective function of $\bar{X}, \sum(X_i - \bar{X})^2$ so that this pair is also sufficient.

Completeness

In the Binomial(n,p) example I showed that there is only one function of X which is unbiased. The Rao Blackwell theorem shows that a UMVUE, if it exists, will be a function of any sufficient statistic. Can there be more than one such function? Generally the answer is yes but for some models like the binomial the answer is no.

Definition*: A statistic T is complete for a model $P_\theta; \theta \in \Theta$ if*

$$E_\theta(h(T)) = 0$$

for all θ implies $h(T) = 0$.

We have already seen that X is complete in the Binomial(n,p) model. In the $N(\mu, 1)$ model suppose

$$E_\mu(h(\bar{X})) \equiv 0$$

Since $\bar{X}$ has a $N(\mu, 1/n)$ distribution we find that

$$\int_{-infty}^{\infty} h(x)e^{-x^2/2}e^{\mu x}dx \equiv 0$$

This is the so called Laplace transform of the function $h(x)e^{-x^2/2}$. It is a theorem that a Laplace transform is 0 if and only if the function is 0 (because you can invert the transform). Hence $h \equiv 0$.

How to Prove Completeness

There is only one general tactic. Suppose X has density

$$f(x,\theta) = h(x)\exp\{\sum_1^p a_i(\theta)S_i(x) + c(\theta)\}$$

If the range of the function $(a_1(\theta),\ldots,a_p(\theta))$ (as θ varies over Θ contains a (hyper-) rectangle in R^p then the statistic

$$(S_1(X),\ldots,S_p(X))$$

is complete and sufficient.

You prove the sufficiency by the factorization criterion and the completeness using the properties of Laplace transforms and the fact that the joint density of $S_1,\ldots,S_p$

$$g(s_1,\ldots,s_p;\theta) = h^*(s)\exp\{\sum a_k(\theta)s_k + c^*(\theta)\}$$

Example*: In the $N(\mu,\sigma^2)$ model the density has the form*

$$\frac{1}{\sqrt{2\pi}}\exp\left\{\left(-\frac{1}{2\sigma^2}\right) + \left(\frac{\mu}{\sigma^2}\right)x - \frac{\mu^2}{2\sigma^2}\right\}$$

which is an exponential family with

$$h(x) = \frac{1}{\sqrt{2\pi}}$$

$$a_1(\theta) = -\frac{1}{2\sigma^2}$$

$$S_1(x) = x^2$$

$$a_2(\theta) = \frac{\mu}{\sigma^2}$$

$$S_2(x) = x$$

and

$$c(\theta) = -\frac{\mu^2}{2\sigma^2}$$

It follows that

$$(\sum X_i^2, \sum X_i)$$

is a complete sufficient statistic.

Remark: The statistic $(s^2,\bar{X})$ is a one to one function of $(\sum X_i^2, \sum X_i)$ so it must be complete and sufficient, too. Any function of the latter statistic can be rewritten as a function of the former and vice versa.

The Lehmann-Scheffé Theorem

Theorem: *If S is a complete sufficient statistic for some model and $h(S)$ is an unbiased estimate of some parameter $\phi(\theta)$ then $h(S)$ is the UMVUE of $\phi(\theta)$.*

Proof: *Suppose T is another unbiased estimate of ϕ. According to the Rao-Blackwell theorem T is improved by $E(T|S)$ so if $h(S)$ is not UMVUE then there must exist another function $h^*(S)$ which is unbiased and whose variance is smaller than that of $h(S)$ for some value of θ. But*

$$E_\theta(h^*(S) - h(S)) \equiv 0$$

so, in fact $h^*(S) = h(S)$.

Example: *In the $N(\mu, \sigma^2)$ example the random variable $(n-1)s^2/\sigma^2$ has a χ^2_{n-1} distribution. It follows that*

$$E\left[\frac{\sqrt{n-1}s}{\sigma}\right] = \int_0^\infty x^{1/2}\left(\frac{x}{2}\right)^{(n-1)/2-1} e^{-x/2}\frac{dx}{2\Gamma((n-1)/2)}$$

Make the substitution $y = x/2$ and get

$$E(s) = \frac{\sigma}{\sqrt{n-1}}\frac{\sqrt{2}}{|gamma((n-1)/2)}\int_0^\infty y^{n/2-1}e^{-y}dy$$

Hence

$$E(s) = \sigma\frac{\sqrt{2(n-1)}\Gamma(n/2)}{\sqrt{n-1}\Gamma((n-1)/2)}$$

The UMVUE of σ is then

$$s\frac{\sqrt{n-1}\Gamma((n-1)/2)}{\sqrt{2(n-1)}\Gamma(n/2)}$$

by the Lehmann-Scheffé theorem.

Criticism of Unbiasedness

(a) *The UMVUE can be* **inadmissible for squared error loss** *meaning that there is a (biased, of course) estimate whose MSE is smaller for every parameter value. An example is the UMVUE of $\phi = p(1-p)$ which is $\hat{\phi} = n\hat{p}(1-\hat{p})/(n-1)$. The MSE of*

$$\tilde{\phi} = \min(\hat{\phi}, 1/4)$$

is smaller than that of $\hat{\phi}$.

(b) *There are examples where unbiased estimation is impossible. The log odds in a Binomial model is $\phi = \log(p/(1-p))$. Since the expectation of any function of the data is a polynomial function of p and since ϕ is* **not** *a polynomial function of p there is no unbiased estimate of ϕ*

(c) The UMVUE of σ is not the square root of the UMVUE of σ^2. This method of estimation does not have the parameterization equivariance that maximum likelihood does.

(d) Unbiasedness is irrelevant (unless you plan to average together many estimators). The property is an average over possible values of the estimate in which positive errors are allowed to cancel negative errors. An exception to this criticism is that if you plan to average a number of estimators to get a single estimator then it is a problem if all the estimators have the same bias. In assignment 5 you have the one way layout example in which the MLE of the residual variance averages together many biased estimates and so is very badly biased. That assignment shows that the solution is not really to insist on unbiasedness but to consider an alternative to averaging for putting the individual estimates together.

Minimal Sufficiency

In any model the statistic $S(X) \equiv X$ is sufficient. In any iid model the vector of order statistics $X_{(1)}, \ldots, X_{(n)}$ is sufficient. In the $N(\mu, 1)$ model then we have three possible sufficient statistics:

(a) $S_1 = (X_1, \ldots, X_n)$.

(b) $S_2 = (X_{(1)}, \ldots, X_{(n)})$.

(c) $S_3 = \bar{X}$.

Notice that I can calculate S_3 from the values of S_1 or S_2 but not vice versa and that I can calculate S_2 from S_1 but not vice versa. It turns out that $\bar{X}$ is a **minimal** *sufficient statistic meaning that it is a function of any other sufficient statistic. (You can't collapse the data set any more without losing information about μ.)*

To recognize minimal sufficient statistics you look at the likelihood function:

Fact*: If you fix some particular θ^* then the log likelihood ratio function*

$$\ell(\theta) - \ell(\theta^*)$$

is minimal sufficient. WARNING: the function is the statistic.

The subtraction of $\ell(\theta^*)$ gets rid of those irrelevant constants in the log-likelihood. For instance in the $N(\mu, 1)$ example we have

$$\ell(\mu) = -n \log(2\pi)/2 - \sum X_i^2/2 + \mu \sum X_i - n\mu^2/2$$

This depends on $\sum X_i^2$ which is not needed for the sufficient statistic. Take $\mu^* = 0$ and get

$$\ell(\mu) - \ell(\mu^*) = \mu \sum X_i - n\mu^2/2$$

This function of μ is minimal sufficient. Notice that from $\sum X_i$ you can compute this minimal sufficient statistic and vice versa. Thus $\sum X_i$ is also minimal sufficient.

FACT*: A complete sufficient statistic is also minimal sufficient.*

Hypothesis Testing

Hypothesis testing is a statistical problem where you must choose, on the basis of data X, between two alternatives. We formalize this as the problem of choosing between two hypotheses*: $H_o : \theta \in \Theta_0$ or $H_1 : \theta \in \Theta_1$ where Θ_0 and Θ_1 are a partition of the model $P_\theta; \theta \in \Theta$. That is $\Theta_0 \cup \Theta_1 = \Theta$ and $\Theta_0 \cap \Theta_1 =$.*

A rule for making the required choice can be described in two ways:

(a) In terms of the set

$$C = \{X : \text{we choose } \Theta_1 \text{ if we observe } X\}$$

called the rejection *or* critical *region of the test.*

(b) In terms of a function $\phi(x)$ which is equal to 1 for those x for which we choose Θ_1 and 0 for those x for which we choose Θ_0.

For technical reasons which will come up soon I prefer to use the second description. However, each ϕ corresponds to a unique rejection region $R_\phi = \{x : \phi(x) = 1\}$.

The Neyman Pearson approach to hypothesis testing which we consider first treats the two hypotheses asymmetrically. The hypothesis H_o is referred to as the null *hypothesis (because traditionally it has been the hypothesis that some treatment has no effect).*

Definition*: The power function of a test ϕ (or the corresponding critical region R_ϕ) is*

$$\pi(\theta) = P_\theta(X \in R_\phi) = E_\theta(\phi(X))$$

We are interested here in **optimality** *theory, that is, the problem of finding the best ϕ. A good ϕ will evidently have $\pi(\theta)$ small for $\theta \in \Theta_0$ and large for theta $\in \Theta_1$. There is generally a trade off which can be made in many ways, however.*

Simple versus Simple testing

Finding a best test is easiest when the hypotheses are very precise.

Definition*: A hypothesis H_i is* **simple** *if Θ_i contains only a single value θ_i.*

The simple versus simple testing problem arises when we test $\theta = \theta_0$ against $\theta = \theta_1$ so that Θ has only two points in it. This problem is of importance as a technical tool, not because it is a realistic situation.

Suppose that the model specifies that if $\theta = \theta_0$ then the density of X is $f_0(x)$ and if $\theta = \theta_1$ then the density of X is $f_1(x)$. How should we choose ϕ? To answer the question we begin by studying the problem of minimizing the total error probability.

We define a **Type I error** *as the error made when $\theta = \theta_0$ but we choose H_1, that is, $X \in R_\phi$. The other kind of error, when $\theta = \theta_1$ but we choose H_0 is called a* **Type II error***. We define the level of a simple versus simple test to be*

$$\alpha = P_{\theta_0}(\text{We make a Type I error})$$

or

$$\alpha = P_{\theta_0}(X \in R_\phi) = E_{\theta_0}(\phi(X))$$

The other error probability is denoted β *and defined as*

$$\beta = P_{\theta_1}(X \notin R_\phi) = E_{\theta_1}(1 - \phi(X))$$

Suppose we want to minimize $\alpha + \beta$*, the total error probability. We want to minimize*

$$E_{\theta_0}(\phi(X)) + E_{\theta_1}(1 - \phi(X)) = \int [\phi(x) f_0(x) + (1 - \phi(x)) f_1(x)] dx$$

The problem is to choose, for each x*, either the value 0 or the value 1, in such a way as to minimize the integral. But for each* x *the quantity*

$$\phi(x) f_0(x) + (1 - \phi(x)) f_1(x)$$

can be chosen either to be $f_0(x)$ *or* $f_1(X)$*. To make it small we take* $\phi(x) = 1$ *if* $f_1(x) > f_0(x)$ *and* $\phi(x) = 0$ *if* $f_1(x) < f_0(x)$*. It makes no difference what we do for those* x *for which* $f_1(x) = f_0(x)$*. Notice that we can divide both sides of these inequalities to rephrase the condition in terms of the* **likelihood ration** $f_1(x)/f_0(x)$*.*

Theorem*: For each fixed* λ *the quantity* $\beta + \lambda\alpha$ *is minimized by any* ϕ *which has*

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda \end{cases}$$

Neyman and Pearson suggested that in practice the two kinds of errors might well have unequal consequences. They suggested that rather than minimize any quantity of the form above you pick the more serious kind of error, label it **Type I** *and require your rule to hold the probability* α *of a Type I error to be no more than some prespecified level* α_0*. (This value* α_0 *is typically* 0.05 *these days, chiefly for historical reasons.)*

The Neyman and Pearson approach is then to minimize beta subject to the constraint $\alpha \leq \alpha_0$*. Usually this is really equivalent to the constraint* $\alpha = \alpha_0$ *(because if you use* $\alpha < \alpha_0$ *you could make* R *larger and keep* $\alpha \leq \alpha_0$ *but make* β *smaller. For discrete models, however, this may not be possible.*

Example*: Suppose* X *is Binomial*(n, p) *and either* $p = p_0 = 1/2$ *or* $p = p_1 = 3/4$*. If* R *is any critical region (so* R *is a subset of* $\{0, 1, \ldots, n\}$*) then*

$$P_{1/2}(X \in R) = \frac{k}{2^n}$$

for some integer k*. If we want* $\alpha_0 = 0.05$ *with say* $n = 5$ *for example we have to recognize that the possible values of* α *are* $0, 1/32 = 0.03125, 2/32 = 0.0625$ *and so on. For* $\alpha_0 = 0.05$ *we must use one of three rejection regions:* R_1 *which is the empty set,* R_2 *which is the set* $x = 0$ *or* R_3 *which is the set* $x = 5$*. These three regions have alpha equal to 0, 0.3125 and 0.3125 respectively and* β *equal to 1,* $1 - (1/4)^5$ *and* $1 - (3/4)^5$

respectively so that R_3 minimizes β subject to $\alpha < 0.05$. If we raise α_0 slightly to 0.0625 then the possible rejection regions are R_1, R_2, R_3 and a fourth region $R_4 = R_2 \cup R_3$. The first three have the same α and β as before while R_4 has $\alpha = \alpha_0 = 0.0625$ an $\beta = 1 - (3/4)^5 - (1/4)^5$. Thus R_4 is optimal! The trouble is that this region says if all the trials are failures we should choose $p = 3/4$ rather than $p = 1/2$ even though the latter makes 5 failures much more likely than the former.

The problem in the example is one of discreteness. Here's how we get around the problem. First we expand the set of possible values of ϕ to include numbers between 0 and 1. Values of $\phi(x)$ between 0 and 1 represent the chance that we choose H_1 given that we observe x; the idea is that we actually toss a (biased) coin to decide! This tactic will show us the kinds of rejection regions which are sensible. In practice we then restrict our attention to levels α_0 for which the best ϕ is always either 0 or 1. In the binomial example we will insist that the value of α_0 be either 0 or $P_{\theta_0}(X \geq 5)$ or $P_{\theta_0}(X \geq 4)$ or ...

Here is a smaller example. There are 4 possible values of X and 2^4 possible rejection regions. Here is a table of the levels for each possible rejection region R:

R	α
{}	0
{3}, {0}	1/8
{0,3}	2/8
{1}, {2}	3/8
{0,1}, {0,2}, {1,3}, {2,3}	4/8
{0,1,3}, {0,2,3}	5/8
{1,2}	6/8
{0,1,3}, {0,2,3}	7/8
{0,1,2,3}	

The best level 2/8 test has rejection region $\{0, 3\}$ and $\beta = 1-[(3/4)^3+(1/4)^3] = 36/64$. If, instead, we permit randomization then we will find that the best level test rejects when $X = 3$ and, when $X = 2$ tosses a coin which has chance 1/3 of landing heads, then rejects if you get heads. The level of this test is $1/8 + (1/3)(3/8) = 2/8$ and the probability of a Type II error is $\beta = 1 - [(3/4)^3 + (1/3)(3)(3/4)^2(1/4)] = 28/64$.

Definition*: A hypothesis test is a function $\phi(x)$ whose values are always in $[0, 1]$. If we observe $X = x$ then we choose H_1 with conditional probability $\phi(X)$. In this case we have*

$$\pi(\theta) = E_\theta(\phi(X))$$

$$\alpha = E_0(\phi(X))$$

and

$$\beta = E_1(\phi(X))$$

The Neyman Pearson Lemma

Theorem: *In testing f_0 against f_1 the probability β of a type II error is minimized, subject to $\alpha \leq \alpha_0$ by the test function:*

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda \\ \gamma & \frac{f_1(x)}{f_0(x)} = \lambda \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda \end{cases}$$

where λ is the largest constant such that

$$P_0(\frac{f_1(x)}{f_0(x)} \geq \lambda) \geq \alpha_0$$

and

$$P_0(\frac{f_1(x)}{f_0(x)} \leq \lambda) \geq 1 - \alpha_0$$

and where γ is any number chosen so that

$$E_0(\phi(X)) = P_0(\frac{f_1(x)}{f_0(x)} > \lambda) + \gamma P_0(\frac{f_1(x)}{f_0(x)} = \lambda) = \alpha_0$$

The value of γ is unique if $P_0(\frac{f_1(x)}{f_0(x)} = \lambda) > 0$.

Example: *In the Binomial(n, p) with $p_0 = 1/2$ and $p_1 = 3/4$ the ratio f_1/f_0 is*

$$3^x 2^{-n}$$

Now if $n = 5$ then this ratio must be one of the numbers 1, 3, 9, 27, 81, 243 divided by 32. Suppose we have $\alpha = 0.05$. The value of λ must be one of the possible values of f_1/f_0. If we try $\lambda = 343/32$ then

$$P_0(3^X 2^{-5} \geq 343/32) = P_0(X = 5) = 1/32 < 0.05$$

and

$$P_0(3^X 2^{-5} \geq 81/32) = P_0(X \geq 4) = 6/32 > 0.05$$

This means that $\lambda = 81/32$. Since

$$P_0(3^X 2^{-5} > 81/32) = P_0(X = 5) = 1/32$$

we must solve

$$P_0(X = 5) + \gamma P_0(X = 4) = 0.05$$

for γ and find

$$\gamma = \frac{0.05 - 1/32}{5/32} = 0.12$$

NOTE: No-one ever uses this procedure. Instead the value of α_0 used in discrete problems is chosen to be a possible value of the rejection probability when $\gamma = 0$ (or $\gamma = 1$).

When the sample size is large you can come very close to any desired α_0 with a non-randomized test.

If $\alpha_0 = 6/32$ then we can either take λ to be 343/32 and $\gamma = 1$ or $\lambda = 81/32$ and $\gamma = 0$. However, our definition of λ in the theorem makes $\lambda = 81/32$ and $\gamma = 0$.

When the theorem is used for continuous distributions it can be the case that the CDF of $f_1(X)/f_0(X)$ has a flat spot where it is equal to $1-\alpha_0$. This is the point of the word "largest" in the theorem.

Example*: If $X_1, \ldots, X_n$ are iid $N(\mu, 1)$ and we have $\mu_0 = 0$ and $\mu_1 > 0$ then*

$$\frac{f_1(X_1, \ldots, X_n)}{f_0(X_1, \ldots, X_n)} = \exp\{\mu_1 \sum X_i - n\mu_1^2/2 - \mu_0 \sum X_i + n\mu_2^2/2\}$$

which simplifies to

$$\exp\{\mu_1 \sum X_i - n\mu_1^2/2\}$$

We now have to choose λ so that

$$P_0(\exp\{\mu_1 \sum X_i - n\mu_1^2/2\} > \lambda) = \alpha_0$$

We can make it equal because in this case $f_1(X)/f_0(X)$ has a continuous distribution. Rewrite the probability as

$$P_0(\sum X_i > [\log(\lambda) + n\mu_1^2/2]/\mu_1) = 1 - \Phi([\log(\lambda) + n\mu_1^2/2]/[n^{1/2}\mu_1])$$

If z_α is notation for the usual upper α critical point of the normal distribution then we find

$$z_{\alpha_0} = [\log(\lambda) + n\mu_1^2/2]/[n^{1/2}\mu_1]$$

which you can solve to get a formula for λ in terms of z_{α_0}, n and μ_1.

The rejection region looks complicated: reject if a complicated statistic is larger than λ which has a complicated formula. But in calculating λ we re-expressed the rejection region in terms of

$$\frac{\sum X_i}{\sqrt{n}} > z_{\alpha_0}$$

The key feature is that this rejection region is the same for any $\mu_1 > 0$. [WARNING: in the algebra above I used $\mu_1 > 0$.] This is why the Neyman Pearson lemma is a lemma!

Definition*: In the general problem of testing Θ_0 against Θ_1 the level of a test function ϕ is*

$$\alpha = \sup_{\theta \in \Theta_0} E_\theta(\phi(X))$$

The power function is

$$\pi(\theta) = E_\theta(\phi(X))$$

A test ϕ^* is a Uniformly Most Powerful level α_0 test if

(a) ϕ^* *has level* $\alpha \leq \alpha_o$

(b) If ϕ *has level* $\alpha \leq \alpha_0$ *then for every* $\theta \in \Theta_1$ *we have*

$$E_\theta(\phi(X)) \leq E_\theta(\phi^*(X))$$

Application of the NP lemma*: In the* $N(\mu, 1)$ *model consider* $\Theta_1 = \{\mu > 0\}$ *and* $\Theta_0 = \{0\}$ *or* $\Theta_0 = \{\mu \leq 0\}$*. The UMP level* α_0 *test of* $H_0 : \mu \in \Theta_0$ *against* $H_1 : \mu \in \Theta_1$ *is*

$$\phi(X_1, \ldots, X_n) = 1(n^{1/2}\bar{X} > z_{\alpha_0})$$

Proof*: For either choice of* Θ_0 *this test has level* α_0 *because for* $\mu \leq 0$ *we have*

$$\begin{aligned} P_\mu(n^{1/2}\bar{X} > z_{\alpha_0}) &= P_\mu(n^{1/2}(\bar{X} - \mu) > z_{\alpha_0} - n^{1/2}\mu) \\ &= P(N(0,1) > z_{\alpha_0} - n^{1/2}\mu) \\ &\leq P(N(0,1) > z_{\alpha_0}) \\ &= \alpha_0 \end{aligned}$$

(Notice the use of $\mu \leq 0$*. The central point is that the critical point is determined by the behaviour on the edge of the null hypothesis.)*

Now if ϕ *is any other level* α_0 *test then we have*

$$E_0(\phi(X_1, \ldots, X_n)) \leq \alpha_0$$

Fix a $\mu > 0$*. According to the NP lemma*

$$E_\mu(\phi(X_1, \ldots, X_n)) \leq E_\mu(\phi_\mu(X_1, \ldots, X_n))$$

where ϕ_μ *rejects if* $f_\mu(X_1, \ldots, X_n)/f_0(X_1, \ldots, X_n) > \lambda$ *for a suitable* λ*. But we just checked that this test had a rejection region of the form*

$$n^{1/2}\bar{X} > z_{\alpha_0}$$

which is the rejection region of ϕ^**. The NP lemma produces the same test for every* $\mu > 0$ *chosen as an alternative. So we have shown that* $\phi_\mu = \phi^*$ *for any* $\mu > 0$*.*

Proof of the Neyman Pearson lemma*: Given a test* ϕ *with level strictly less than* α_0 *we can define the test*

$$\phi^*(x) = \frac{1 - \alpha_0}{1 - \alpha}\phi(x) + \frac{\alpha_0 - \alpha}{1 - \alpha}$$

has level α_0 *and* β *smaller than that of* ϕ*. Hence we may assume without loss that* $\alpha = \alpha_0$ *and minimize* β *subject to* $\alpha = \alpha_0$*. However, the argument which follows doesn't actually need this.*

Lagrange Multipliers

Suppose you want to minimize $f(x)$ subject to $g(x) = 0$. Consider first the function

$$h_\lambda(x) = f(x) + \lambda g(x)$$

If x_λ minimizes h_λ then for any other x

$$f(x_\lambda) \leq f(x) + \lambda[g(x) - g(x_\lambda)]$$

Now suppose you can find a value of λ such that the solution x_λ has $g(x_\lambda) = 0$. Then for any x we have

$$f(x_\lambda) \leq f(x) + \lambda g(x)$$

and for any x satisfying the constraint $g(x) = 0$ we have

$$f(x_\lambda) \leq f(x)$$

This proves that for this special value of λ the quantity x_λ minimizes $f(x)$ subject to $g(x) = 0$.

Notice that to find x_λ you set the usual partial derivatives equal to 0; then to find the special x_λ you add in the condition $g(x\lambda) = 0$.

Proof of NP lemma

For each $\lambda > 0$ we have seen that ϕ_λ minimizes $\lambda\alpha + \beta$ where $\phi_\lambda = 1(f_1(x)/f_0(x) \geq \lambda)$.

As λ increases the level of ϕ_λ decreases from 1 when $\lambda = 0$ to 0 when $\lambda = \infty$. There is thus a value λ_0 where for $\lambda < \lambda_0$ the level is less than α_0 while for $\lambda > \lambda_0$ the level is at least α_0. Temporarily let $\delta = P_0(f_1(X)/f_0(X) = \lambda_0)$. If $\delta = 0$ define $\phi = \phi_\lambda$. If $\delta > 0$ define

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda_0 \\ \gamma & \frac{f_1(x)}{f_0(x)} = \lambda_0 \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda_0 \end{cases}$$

where $P_0(f_1(X)/f_0(X) < \lambda_0) + \gamma\delta = \alpha_0$. You can check that $\gamma \in [0, 1]$.

Now ϕ has level α_0 and according to the theorem above minimizes $\alpha + \lambda_0\beta$. Suppose ϕ^* is some other test with level $\alpha^* \leq \alpha_0$. Then

$$\lambda_0\alpha_\phi + \beta_\phi \leq \lambda_0\alpha_{\phi^*} + \beta_{\phi^*}$$

We can rearrange this as

$$\beta_{\phi^*} \geq \beta_\phi + (\alpha_\phi - \alpha_{\phi^*})\lambda_0$$

Since

$$\alpha_{\phi^*} \leq \alpha_0 = \alpha_\phi$$

the second term is non-negative and

$$\beta_{\phi^*} \geq \beta_\phi$$

which proves the Neyman Pearson Lemma.

Definition*: In the general problem of testing Θ_0 against Θ_1 the level of a test function ϕ is*

$$\alpha = \sup_{\theta \in \Theta_0} E_\theta(\phi(X))$$

The power function is

$$\pi(\theta) = E_\theta(\phi(X))$$

A test ϕ^* is a Uniformly Most Powerful level α_0 test if

(a) ϕ^* has level $\alpha \leq \alpha_o$

(b) If ϕ has level $\alpha \leq \alpha_0$ then for every $\theta \in \Theta_1$ we have

$$E_\theta(\phi(X)) \leq E_\theta(\phi^*(X))$$

Application of the NP lemma*: In the $N(\mu, 1)$ model consider $\Theta_1 = \{\mu > 0\}$ and $\Theta_0 = \{0\}$ or $\Theta_0 = \{\mu \leq 0\}$. The UMP level α_0 test of $H_0 : \mu \in \Theta_0$ against $H_1 : \mu \in \Theta_1$ is*

$$\phi(X_1, \ldots, X_n) = 1(n^{1/2}\bar{X} > z_{\alpha_0})$$

Proof*: For either choice of Θ_0 this test has level α_0 because for $\mu \leq 0$ we have*

$$\begin{aligned} P_\mu(n^{1/2}\bar{X} > z_{\alpha_0}) &= P_\mu(n^{1/2}(\bar{X} - \mu) > z_{\alpha_0} - n^{1/2}\mu) \\ &= P(N(0,1) > z_{\alpha_0} - n^{1/2}\mu) \\ &\leq P(N(0,1) > z_{\alpha_0}) \\ &= \alpha_0 \end{aligned}$$

(Notice the use of $\mu \leq 0$. The central point is that the critical point is determined by the behaviour on the edge of the null hypothesis.)

Now if ϕ is any other level α_0 test then we have

$$E_0(\phi(X_1, \ldots, X_n)) \leq \alpha_0$$

Fix a $\mu > 0$. According to the NP lemma

$$E_\mu(\phi(X_1, \ldots, X_n)) \leq E_\mu(\phi_\mu(X_1, \ldots, X_n))$$

where ϕ_μ rejects if $f_\mu(X_1, \ldots, X_n)/f_0(X_1, \ldots, X_n) > \lambda$ for a suitable λ. But we just checked that this test had a rejection region of the form

$$n^{1/2}\bar{X} > z_{\alpha_0}$$

which is the rejection region of ϕ^*. The NP lemma produces the same test for every $\mu > 0$ chosen as an alternative. So we have shown that $\phi_\mu = \phi^*$ for any $\mu > 0$.

This phenomenon is somewhat general. What happened was this. For any $\mu > \mu_0$ the likelihood ratio f_μ/f_0 is an increasing function of $\sum X_i$. The rejection region of the

NP test is thus always a region of the form $\sum X_i > k$. The value of the constant k is determined by the requirement that the test have level α_0 and this depends on μ_0 not on μ_1.

Definition*: The family $f_\theta; \theta \in \Theta \subset R$ has monotone likelihood ratio with respect to a statistic $T(X)$ if for each $\theta_1 > \theta_0$ the likelihood ratio $f_{\theta_1}(X)/f_{\theta_0}(X)$ is a monotone increasing function of $T(X)$.*

Theorem*: For a monotone likelihood ratio family the Uniformly Most Powerful level α test of $\theta \leq \theta_0$ (or of $\theta = \theta_0$) against the alternative $\theta > \theta_0$ is*

$$\phi(x) = \begin{cases} 1 & T(x) > t_\alpha \\ \gamma & T(X) = t_\alpha \\ 0 & T(x) < t_\alpha \end{cases}$$

where $P_0(T(X) > t_\alpha) + \gamma P_0(T(X) = t_\alpha) = \alpha_0$.

A typical family where this will work is a one parameter exponential family. In almost any other problem the method doesn't work and there is no uniformly most powerful test. For instance to test $\mu = \mu_0$ against the two sided alternative $\mu \neq \mu_0$ there is no UMP level α test. If there were its power at $\mu > \mu_0$ would have to be as high as that of the one sided level α test and so its rejection region would have to be the same as that test, rejecting for large positive values of $\bar{X} - \mu_0$. But it also has to have power as good as the one sided test for the alternative $\mu < \mu_0$ and so would have to reject for large negative values of $\bar{X} - \mu_0$. This would make its level too large.

The favourite test is the usual 2 sided test which rejects for large values of $|\bar{X} - \mu_0|$ with the critical value chosen appropriately. This test maximizes the power subject to two constraints: first, that the level be α and second that the test have power which is minimized at $\mu = \mu_0$. This second condition is really that the power on the alternative be larger than it is on the null.

Definition*: A test ϕ of Θ_0 against Θ_1 is unbiased level α if it has level α and, for every $\theta \in \Theta_1$ we have*

$$\pi(\theta) \geq \alpha\,.$$

When testing a point null hypothesis like $\mu = \mu_0$ this requires that the power function be minimized at μ_0 which will mean that if π is differentiable then

$$\pi'(\mu_0) = 0$$

We now apply that condition to the $N(\mu, 1)$ problem. If ϕ is any test function then

$$\pi'(\mu) = \frac{\partial}{\partial \mu} \int \phi(x) f(x, \mu) dx$$

We can differentiate this under the integral and use

$$\frac{\partial f(x, \mu)}{\partial \mu} = \sum (x_i - \mu) f(x, \mu)$$

to get the condition

$$\int \phi(x)\bar{x}f(x,\mu_0)dx = \mu_0\alpha_0$$

Consider the problem of minimizing $\beta(\mu)$ *subject to the two constraints* $E_{\mu_0}(\phi(X)) = \alpha_0$ *and* $E_{\mu_0}(\bar{X}\phi(X)) = \mu_0\alpha_0$. *Now fix two values* $\lambda_1 > 0$ *and* λ_2 *and minimize*

$$\lambda_1\alpha + \lambda_2 E_{\mu_0}[(\bar{X} - \mu_0)\phi(X)] + \beta$$

The quantity in question is just

$$\int [\phi(x)f_0(x)(\lambda_1 + \lambda_2(\bar{X} - \mu_0)) + (1 - \phi(x))f_1(x)]dx$$

As before this is minimized by

$$\phi(x) = \begin{cases} 1 & \frac{f_1(x)}{f_0(x)} > \lambda_1 + \lambda_2(\bar{X} - \mu_0) \\ 0 & \frac{f_1(x)}{f_0(x)} < \lambda_1 + \lambda_2(\bar{X} - \mu_0) \end{cases}$$

The likelihood ratio f_1/f_0 *is simply*

$$\exp\{n(\mu_1 - \mu_0)\bar{X} + n(\mu_0^2 - \mu_1^2)/2\}$$

and this exceeds the linear function

$$\lambda_1 + \lambda_2(\bar{X} - \mu_0)$$

for all $\bar{X}$ *sufficiently large or small. That is, the quantity*

$$\lambda_1\alpha + \lambda_2 E_{\mu_0}[(\bar{X} - \mu_0)\phi(X)] + \beta$$

is minimized by a rejection region of the form

$$\{\bar{X} > K_U\} \cup \{\bar{X} < K_L\}$$

To satisfy the constraints we adjust K_U *and* K_L *to get level* α *and* $\pi'(\mu_0) = 0$. *The second condition shows that the rejection region is symmetric about* μ_0 *and then we discover that the test rejects for*

$$\sqrt{n}|\bar{X} - \mu_0| > z_{\alpha/2}$$

Now you have to mimic the Neyman Pearson lemma proof to check that if λ_1 *and* λ_2 *are adjusted so that the unconstrained problem has the rejection region given then the resulting test minimizes* β *subject to the two constraints.*

A test ϕ^* *is a Uniformly Most Powerful Unbiased level* α_0 *test if*

(a) ϕ^* *has level* $\alpha \leq \alpha_0$.

(b) ϕ^* *is unbiased.*

(c) If ϕ has level $\alpha \leq \alpha_0$ then for every $\theta \in \Theta_1$ we have

$$E_\theta(\phi(X)) \leq E_\theta(\phi^*(X))$$

Conclusion*: The two sided z test which rejects if*

$$|Z| > z_{\alpha/2}$$

where

$$Z = n^{1/2}(\bar{X} - \mu_0)$$

is the uniformly most powerful unbiased test of $\mu = \mu_0$ against the two sided alternative $\mu \neq \mu_0$.

Nuisance Parameters

What good can be said about the t-test? It's UMPU.

Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ and that we want to test $\mu = \mu_0$ or $\mu \leq \mu_0$ against $\mu > \mu_0$. Notice that the parameter space is two dimensional and that the boundary between the null and alternatives is

$$\{(\mu, \sigma); \mu = \mu_0, \sigma > 0\}$$

If a test has $\pi(\mu, \sigma) \leq \alpha$ for all $\mu \leq \mu_0$ and $\pi(\mu, \sigma) \geq \alpha$ for all $\mu > \mu_0$ then we must have $\pi(\mu_0, \sigma) = \alpha$ for all σ because the power function of any test must be continuous. (This actually uses the dominated convergence theorem; the power function is an integral.)

Now think of $\{(\mu, \sigma); \mu = \mu_0\}$ as a parameter space for a model. For this parameter space you can check that

$$S = \sum (X_i - \mu_0)^2$$

is a complete sufficient statistic. Remember that the definitions of both completeness and sufficiency depend on the parameter space. Suppose $\phi(\sum X_i, S)$ is an unbiased level α test. Then we have

$$E_{\mu_0,\sigma}(\phi(\sum X_i, S)) = \alpha$$

for all σ. Condition on S and get

$$E_{\mu_0,\sigma}[E(\phi(\sum X_i, S)|S)] = \alpha$$

for all σ. Sufficiency guarantees that

$$g(S) = E(\phi(\sum X_i, S)|S)$$

is a statistic and completeness that

$$g(S) \equiv \alpha$$

Now let us fix a single value of σ and a $\mu_1 > \mu_0$. To make our notation simpler I take $\mu_0 = 0$. Our observations above permit us to condition on $S = s$. Given $S = s$ we have a level α test which is a function of $\bar{X}$.

If we maximize the conditional power of this test for each s then we will maximize its power. What is the conditional model given $S = s$? That is, what is the conditional distribution of $\bar{X}$ given $S = s$? The answer is that the joint density of $\bar{X}, S$ is of the form

$$f_{\bar{X},S}(t,s) = h(s,t)\exp\{\theta_1 t + \theta_2 s + c(\theta_1,\theta_2)\}$$

where $\theta_1 = n\mu/\sigma^2$ and $\theta_2 = -1/\sigma^2$.

This makes the conditional density of $\bar{X}$ given $S = s$ of the form

$$f_{\bar{X}|s}(t|s) = h(s,t)\exp\{\theta_1 t + c^*(\theta_1,s)\}$$

Notice the disappearance of θ_2. Notice that the null hypothesis is actually $\theta_1 = 0$. This permits the application of the NP lemma to the conditional family to prove that the UMP unbiased test must have the form

$$\phi(\bar{X},S) = 1(\bar{X} > K(S))$$

where $K(S)$ is chosen to make the conditional level α. The function $x \mapsto x/\sqrt{a-x^2}$ is increasing in x for each a so that we can rewrite ϕ in the form

$$\phi(\bar{X},S) = 1(n^{1/2}\bar{X}/\sqrt{n[S/n - \bar{X}^2]/(n-1)} > K^*(S))$$

for some K^*. The quantity

$$T = \frac{n^{1/2}\bar{X}}{\sqrt{n[S/n - \bar{X}^2]/(n-1)}}$$

is the usual t statistic and is exactly independent of S (see Theorem 6.1.5 on page 262 in Casella and Berger). This guarantees that

$$K^*(S) = t_{n-1,\alpha}$$

and makes our UMPU test the usual t test.

Optimal tests

- *A good test has $\pi(\theta)$ large on the alternative and small on the null.*
- *For one sided one parameter families with MLR a UMP test exists.*
- *For two sided or multiparameter families the best to be hoped for is UMP Unbiased or Invariant or Similar.*
- *Good tests are found as follows:*

(a) Use the NP lemma to determine a good rejection region for a simple alternative.

(b) Try to express that region in terms of a statistic whose definition does not depend on the specific alternative.

(c) If this fails impose an additional criterion such as unbiasedness. Then mimic the NP lemma and again try to simplify the rejection region.

Likelihood Ratio tests

For general composite hypotheses optimality theory is not usually successful in producing an optimal test. instead we look for heuristics to guide our choices. The simplest approach is to consider the likelihood ratio

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)}$$

and choose values of $\theta_1 \in \Theta_1$ and $\theta_0 \in \Theta_0$ which are reasonable estimates of θ assuming respectively the alternative or null hypothesis is true. The simplest method is to make each θ_i a maximum likelihood estimate, but maximized only over Θ_i.

Example 1*: In the $N(\mu, 1)$ problem suppose we want to test $\mu \leq 0$ against $\mu > 0$. (Remember there is a UMP test.) The log likelihood function is*

$$-n(\bar{X} - \mu)^2/2$$

If $\bar{X} > 0$ then this function has its global maximum in Θ_1 at $\bar{X}$. Thus $\hat{\mu}_1$ which maximizes $\ell(\mu)$ subject to $\mu > 0$ is $\bar{X}$ if $\bar{X} > 0$. When $\bar{X} \leq 0$ the maximum of $\ell(\mu)$ over $\mu > 0$ is on the boundary, at $\hat{\mu}_1 = 0$. (Technically this is in the null but in this case $\ell(0)$ is the supremum of the values $\ell(\mu)$ for $\mu > 0$. Similarly, the estimate $\hat{\mu}_0$ will be $\bar{X}$ if $\bar{X} \leq 0$ and 0 if $\bar{X} > 0$. It follows that

$$\frac{f_{\theta_1}(X)}{f_{\theta_0}(X)} = \exp\{\ell(\hat{\mu}_1) - \ell(\hat{\mu}_0)\}$$

which simplifies to

$$\exp\{n\bar{X}|\bar{X}|/2\}$$

This is a monotone increasing function of $\bar{X}$ so the rejection region will be of the form $\bar{X} > K$. To get the level right the test will have to reject if $n^{1/2}\bar{X} > z_\alpha$. Notice that the log likelihood ratio statistic

$$\lambda \equiv 2\log(\frac{f_{\hat{\mu}_1}(X)}{f_{\hat{\mu}_0}(X)} = n\bar{X}|\bar{X}|$$

as a simpler statistic.

Example 2*: In the $N(\mu, 1)$ problem suppose we make the null $\mu = 0$. Then the value of $\hat{\mu}_0$ is simply 0 while the maximum of the log-likelihood over the alternative $\mu \neq 0$ occurs at $\bar{X}$. This gives*

$$\lambda = n\bar{X}^2$$

which has a χ^2_1 distribution. This test leads to the rejection region $\lambda > (z_{\alpha/2})^2$ which is the usual UMPU test.

Example 3*: For the $N(\mu, \sigma^2)$ problem testing $\mu = 0$ against $\mu \neq 0$ we must find two estimates of μ, σ^2. The maximum of the likelihood over the alternative occurs at the global mle $\bar{X}, \hat{\sigma}^2$. We find*

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -n/2 - n\log(\hat{\sigma})$$

We also need to maximize ℓ over the null hypothesis. Recall

$$\ell(\mu, \sigma) = -\frac{1}{2\sigma^2}\sum(X_i - \mu)^2 - n\log(\sigma)$$

On the null hypothesis we have $\mu = 0$ and so we must find $\hat{\sigma}_0$ by maximizing

$$\ell(0, \sigma) = -\frac{1}{2\sigma^2}\sum X_i^2 - n\log(\sigma)$$

This leads to

$$\hat{\sigma}_0^2 = \sum X_i^2/n$$

and

$$\ell(0, \hat{\sigma}_0) = -n/2 - n\log(\hat{\sigma}_0)$$

This gives

$$\lambda = -n\log(\hat{\sigma}^2/\hat{\sigma}_0^2)$$

Since

$$\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2} = \frac{\sum(X_i - \bar{X})^2}{\sum(X_i - \bar{X})^2 + n\bar{X}^2}$$

we can write

$$\lambda = n\log(1 + t^2/(n-1))$$

where

$$t = \frac{n^{1/2}\bar{X}}{s}$$

is the usual t statistic. The likelihood ratio test thus rejects for large values of $|t|$ which gives the usual test.

Notice that if n is large we have

$$\lambda \approx n[1 + t^2/(n-1) + O(n^{-2})] \approx t^2\,.$$

Since the t statistic is approximately standard normal if n is large we see that

$$\lambda = 2[\ell(\hat{\theta}_1) - \ell(\hat{\theta}_0)]$$

has nearly a χ^2_1 distribution.

This is a general phenomenon when the null hypothesis being tested is of the form $\phi = 0$. Here is the general theory. Suppose that the vector of $p + q$ parameters θ can

be partitioned into $\theta = (\phi, \gamma)$ with ϕ a vector of p parameters and γ a vector of q parameters. To test $\phi = \phi_0$ we find two MLEs of θ. First the global mle $\hat{\theta} = (\hat{\phi}, \hat{\gamma})$ maximizes the likelihood over $\Theta_1 = \{\theta : \phi \neq \phi_0\}$ (because typically the probability that $\hat{\phi}$ is exactly ϕ_0 is 0).

Now we maximize the likelihood over the null hypothesis, that is we find $\hat{\theta}_0 = (\phi_0, \hat{\gamma}_0)$ to maximize

$$\ell(\phi_0, \gamma)$$

The log-likelihood ratio statistic is

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

Now suppose that the true value of θ is ϕ_0, γ_0 (so that the null hypothesis is true). The score function is a vector of length $p + q$ and can be partitioned as $U = (U_\phi, U_\gamma)$. The Fisher information matrix can be partitioned as

$$\begin{bmatrix} I_{\phi\phi} & I_{\phi\gamma} \\ I_{\phi\gamma} & I_{\gamma\gamma} \end{bmatrix}.$$

According to our large sample theory for the mle we have

$$\hat{\theta} \approx \theta + I^{-1}U$$

and

$$\hat{\gamma}_0 \approx \gamma_0 + I_{\gamma\gamma}^{-1} U_\gamma$$

If you carry out a two term Taylor expansion of both $\ell(\hat{\theta})$ and $\ell(\hat{\theta}_0)$ around θ_0 you get

$$\ell(\hat{\theta}) \approx \ell(\theta_0) + U^t I^{-1} U + \frac{1}{2} U^t I^{-1} V(\theta) I^{-1} U$$

where V is the second derivative matrix of ℓ. Remember that $V \approx -I$ and you get

$$2[\ell(\hat{\theta}) - \ell(\theta_0)] \approx U^t I^{-1} U\,.$$

A similar expansion for $\hat{\theta}_0$ gives

$$2[\ell(\hat{\theta}_0) - \ell(\theta_0)] \approx U_\gamma^t I_{\gamma\gamma}^{-1} U_\gamma\,.$$

If you subtract these you find that

$$2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

can be written in the approximate form

$$U^t M U$$

for a suitable matrix M. It is now possible to use the general theory of the distribution of $X^t M X$ where X is $MVN(0, \Sigma)$ to demonstrate that

Theorem*: The log-likelihood ratio statistic*

$$\lambda = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

has, under the null hypothesis, approximately a χ^2_p distribution.

Aside:

Theorem*: Suppose that $X \sim MVN(0, \Sigma)$ with Σ non-singular and M is a symmetric matrix. If $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ then $X^t M X$ has a χ^2 distribution with degrees of freedom $\nu = trace(M\Sigma)$.*

Proof*: We have $X = AZ$ where $AA^t = \Sigma$ and Z is standard multivariate normal. So $X^t M X = Z^t A^t M A Z$. Let $Q = A^t M A$. Since $AA^t = \Sigma$ the condition in the theorem is actually*

$$AQQA^t = AQA^t$$

Since Σ is non-singular so is A. Multiply by A^{-1} on the left and $(A^t)^{-1}$ on the right to discover $QQ = Q$.

The matrix Q is symmetric and so can be written in the form $P\Lambda P^t$ where Λ is a diagonal matrix containing the eigenvalues of Q and P is an orthogonal matrix whose columns are the corresponding orthonormal eigenvectors. It follows that we can rewrite

$$Z^t Q Z = (P^t Z)^t \Lambda (PZ)$$

The variable $W = P^t Z$ is multivariate normal with mean 0 and variance covariance matrix $P^t P = I$; that is, W is standard multivariate normal. Now

$$W^t \Lambda W = \sum \lambda_i W_i^2$$

We have established that the general distribution of any quadratic form $X^t M X$ is a linear combination of χ^2 variables. Now go back to the condition $QQ = Q$. If λ is an eigenvalue of Q and $v \neq 0$ is a corresponding eigenvector then $QQv = Q(\lambda v) = \lambda Q v = \lambda^2 v$ but also $QQv = Qv = \lambda v$. Thus $\lambda(1-\lambda)v = 0$. It follows that either $\lambda = 0$ or $\lambda = 1$. This means that the weights in the linear combination are all 1 or 0 and that $X^t M X$ has a χ^2 distribution with degrees of freedom, ν, equal to the number of λ_i which are equal to 1. This is the same as the sum of the λ_i so

$$\nu = trace(\Lambda)$$

But

$$\begin{aligned} trace(M\Sigma) &= trace(MAA^t) \\ &= trace(A^t M A) \\ &= trace(Q) \\ &= trace(P\Lambda P^t) \\ &= trace(\Lambda P^t P) \\ &= trace(\Lambda) \end{aligned}$$

In the application Σ is $\mathcal{I}$ the Fisher information and $M = \mathcal{I}^{-1} - J$ where

$$J = \begin{bmatrix} 0 & 0 \\ 0 & I_{\gamma\gamma}^{-1} \end{bmatrix}$$

It is easy to check that $M\Sigma$ becomes

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$$

where I is a $p \times p$ identity matrix. It follows that $M\Sigma M\Sigma = M\Sigma$ and that $trace(M\Sigma) = p$).

Confidence Sets

A level β confidence set for a parameter $\phi(\theta)$ is a random subset C, of the set of possible values of ϕ such that for each θ we have

$$P_\theta(\phi(\theta) \in C) \geq \beta$$

Confidence sets are very closely connected with hypothesis tests:

Suppose C is a level $\beta = 1 - \alpha$ confidence set for ϕ. To test $\phi = \phi_0$ we consider the test which rejects if $\phi \notin C$. This test has level α. Conversely, suppose that for each ϕ_0 we have available a level α test of $\phi = \phi_0$ who rejection region is say R_{ϕ_0}. Then if we define $C = \{\phi_0 : \phi = \phi_0$ is not rejected$\}$ we get a level $1 - \alpha$ confidence for ϕ. The usual t test gives rise in this way to the usual t confidence intervals

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$$

which you know well.

Confidence sets from Pivots

Definition*: A* **pivot** *(or pivotal quantity) is a function $g(\theta, X)$ whose distribution is the same for all θ. (As usual the θ in the pivot is the same θ as the one being used to calculate the distribution of $g(\theta, X)$.*

Pivots can be used to generate confidence sets as follows. Pick a set A in the space of possible values for g. Let $\beta = P_\theta(g(\theta, X) \in A)$; since g is pivotal β is the same for all θ. Now given a data set X solve the relation

$$g(\theta, X) \in A$$

to get

$$\theta \in C(X, A)\,.$$

Example: *The quantity*

$$(n-1)s^2/\sigma^2$$

is a pivot in the $N(\mu, \sigma^2)$ *model. It has a* χ^2_{n-1} *distribution. Given* $\beta = 1-\alpha$ *consider the two points* $\chi^2_{n-1,1-\alpha/2}$ *and* $\chi^2_{n-1,\alpha/2}$. *Then*

$$P(\chi^2_{n-1,1-\alpha/2} \leq (n-1)s^2/\sigma^2 \leq \chi^2_{n-1,\alpha/2}) = \beta$$

for all μ, σ. *We can solve this relation to get*

$$P(\frac{(n-1)^{1/2}s}{\chi_{n-1,\alpha/2}} \leq \sigma \leq \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta$$

so that the interval from $(n-1)^{1/2}s/\chi_{n-1,\alpha/2}$ *to* $(n-1)^{1/2}s/\chi_{n-1,1-\alpha/2}$ *is a level* $1-\alpha$ *confidence interval.*

In the same model we also have

$$P(\chi^2_{n-1,1-\alpha} \leq (n-1)s^2/\sigma^2) = \beta$$

which can be solved to get

$$P(\sigma \leq \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta$$

This gives a level $1-\alpha$ *interval* $(0, (n-1)^{1/2}s/\chi_{n-1,1-\alpha})$. *The right hand end of this interval is usually called a confidence upper bound.*

In general the interval from $(n-1)^{1/2}s/\chi_{n-1,\alpha_1}$ *to* $(n-1)^{1/2}s/\chi_{n-1,1-\alpha_2}$ *has level* $\beta = 1-\alpha_1-\alpha_2$. *For a fixed value of* β *we can minimize the length of the resulting interval numerically. This sort of optimization is rarely used. See your homework for an example of the method.*

Decision Theory and Bayesian Methods

Example: *I get up in the morning and must decide between 4 modes of transportation to work:*

- B = *Ride my bike.*
- C = *Take the car.*
- T = *Use public transit.*
- H = *Stay home.*

Costs to me depend on the weather that day: R = *Rain or* S = *Sun.*

Ingredients of a Decision Problem: *No data case.*

- *Decision space* $D = \{B, C, T, H\}$ *of possible actions I might take.*
- *Parameter space* $\Theta = \{R, S\}$ *of possible "states of nature".*

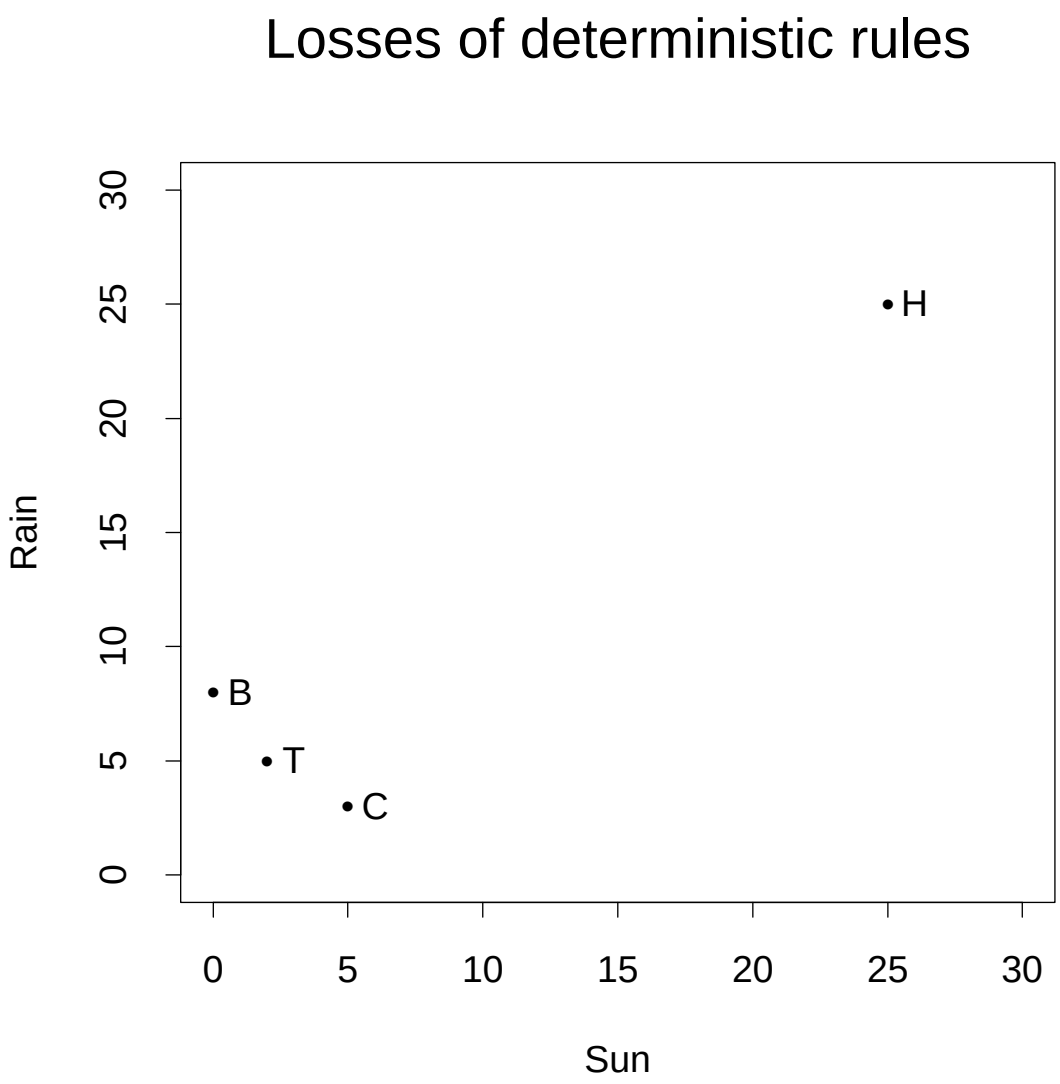
- *Loss function $L = L(d, \theta)$ which is the loss I incur if I do d and θ is the true state of nature.*

In the example we might use the following table for L:

	C	B	T	H
R	3	8	5	25
S	5	0	2	25

Notice that if it rains I will be glad if I drove. If it is sunny I will be glad if I rode my bike. In any case staying at home is expensive.

In general we study this problem by comparing various functions of θ. In this problem a function of θ has only two values, one for rain and one for sun and we can plot any such function as a point in the plane. We do so to indicate the geometry of the problem before stating the general theory.



Statistical Decision Theory

Statistical problems have another ingredient, the data. We observe X a random variable taking values in say $\mathcal{X}$. We may make our decision d depend on X. A **decision rule** *is a function $\delta(X)$ from $\mathcal{X}$ to D. We will want $L(\delta(X), \theta)$ to be small for all θ. Since X is random we quantify this by averaging over X and compare procedures δ in terms of the* **risk function**

$$R_\delta(\theta) = E_\theta(L(\delta(X), \theta))$$

To compare two procedures we must compare two functions of θ and pick "the smaller one". But typically the two functions will cross each other and there won't be a unique 'smaller one'.

Example*: In estimation theory to estimate a real parameter θ we used $D = \Theta$,*

$$L(d, \theta) = (d - \theta)^2$$

and find that the risk of an estimator $\hat{\theta}(X)$ is

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

which is just the Mean Squared Error of $\hat{\theta}$. We have already seen that there is no unique best estimator in the sense of MSE. How do we compare risk functions in general?

- **Minimax methods** *choose δ to minimize the worst case risk:*

  $$\sup\{R_\delta(\theta); \theta \in \Theta)\}.$$

  We call δ^* minimax if

  $$\sup_\theta R_{\delta^*}(\theta) = \inf_\delta \sup_\theta R_\delta(\theta)$$

  Usually the suprema and infima are achieved and we write max *for* sup *and* min *for* inf*. This is the source of "minimax".*

- **Bayes methods** *choose δ to minimize an average*

  $$r_\pi(\delta) = \int R_\delta(\theta)\pi(\theta)d\theta$$

  for a suitable density π. We call π a **prior** *density and r the* **Bayes** *risk of δ for the prior π.*

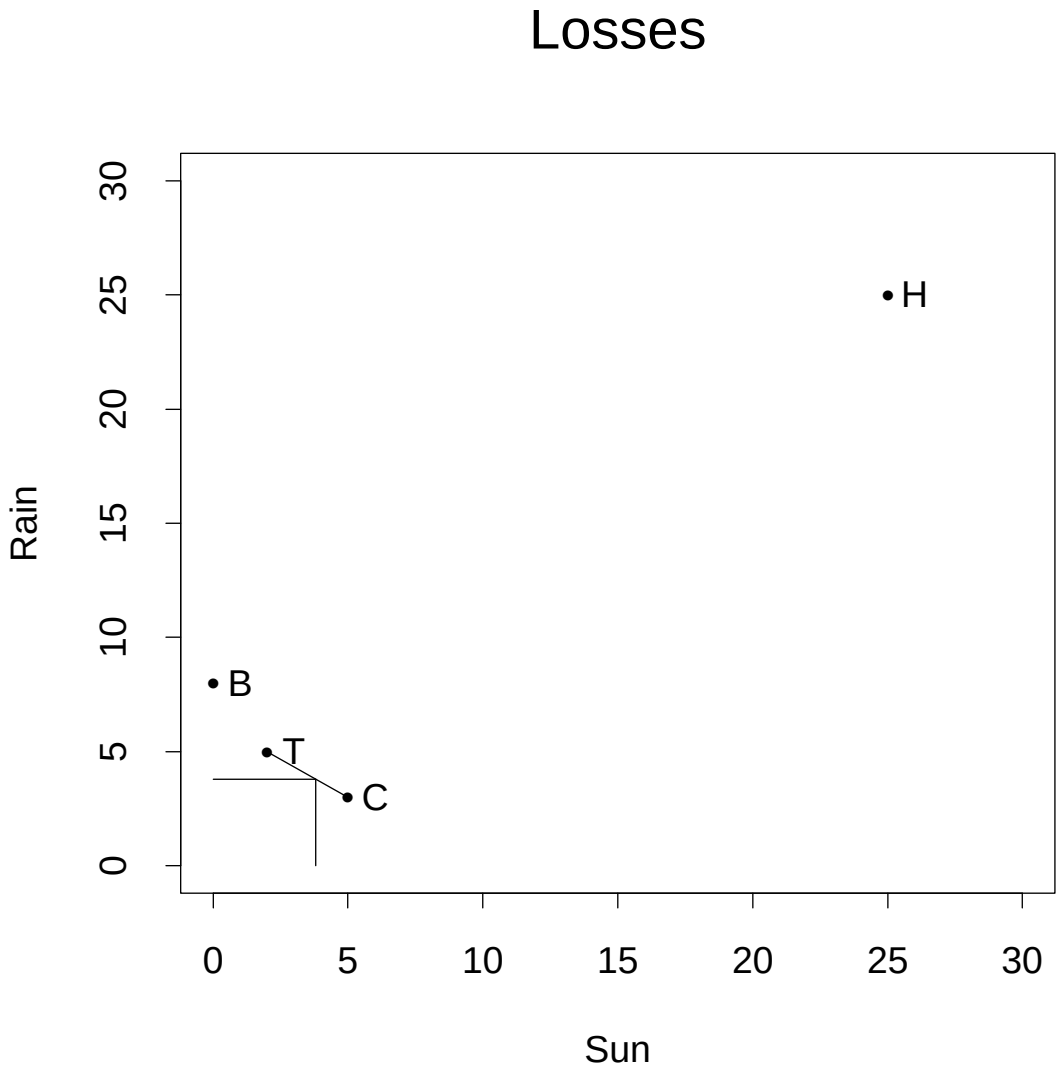
Example*: For my transportation problem there is no data so the only possible (non-randomized) decisions are the four possible actions B, C, T, H. For B and T the worst case is rain. For the other two actions Rain and Sun are equivalent. We have the following table:*

	C	B	T	H
R	3	8	5	25
S	5	0	2	25
Maximum	5	8	5	25

The smallest maximum arises for taking my car. The minimax action is to take my car or public transit.

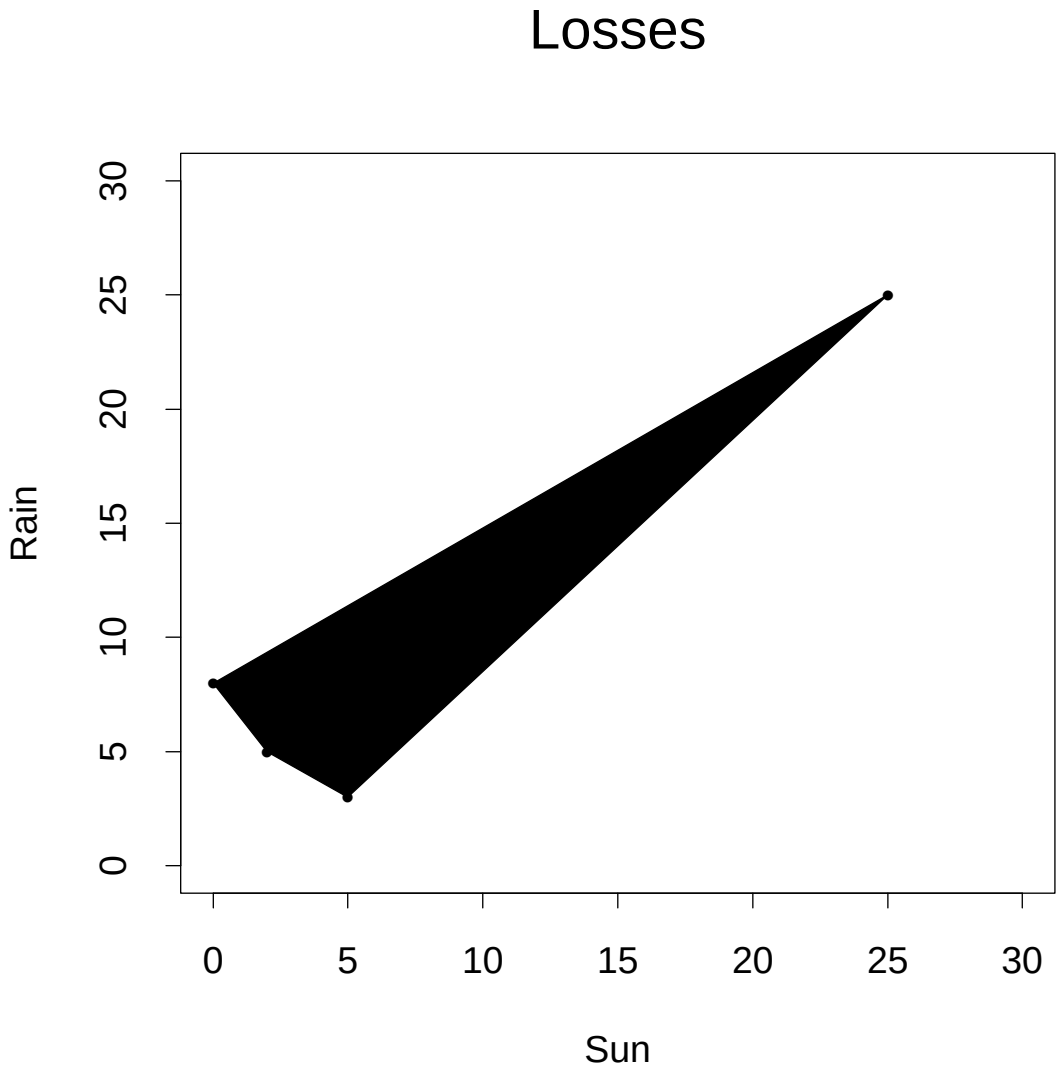
Now imagine each morning I toss a coin with probability λ of getting Heads and take my car if I get Heads, otherwise taking transit. Now in the long run my average daily loss for this procedure would be $3\lambda + 5(1-\lambda)$ when it rains and $5\lambda + 2(1-\lambda)$ when it is Sunny. I will call this procedure d_λ and add it to my graph for each value of λ. Notice that on the graph varying λ from 0 to 1 gives a straight line running from $(3, 5)$ to $(5, 2)$. The two losses are equal when $\lambda = 3/5$. For smaller λ the worst case risk is for sun while for larger λ the worst case risk is for rain.

On the graph below I have added the loss functions for each d_λ, (a straight line) and the set of (x, y) pairs for which $\min(x, y) = 3.8$; this is the worst case risk for d_λ when $\lambda = 3/5$.



The figure then shows that $d_{3/5}$ is actually the minimax procedure when randomized procedures are permitted.

In general we might consider using a 4 sided coin where we took action B with probability λ_B, C with probability λ_C and so on. The loss function of such a procedure is a convex combination of the losses of the four basic procedures making the set of risks achievable with the aid of randomization look like the following:



The use of randomization in general decision problems permits us to assume that the set of possible risk functions is convex. This is an important technical conclusion; it permits us to prove many of the basic results of decision theory.

Studying the graph we can see that many of the points in the picture correspond to bad decision procedures. Regardless of whether or not it rains taking my car to work has a lower loss than staying home; we call the decision to stay home inadmissible.

Definition*: A decision rule δ is* **inadmissible** *if there is a rule δ^* such that*

$$R_{\delta^*}(\theta) \leq R_{\delta}(\theta)$$

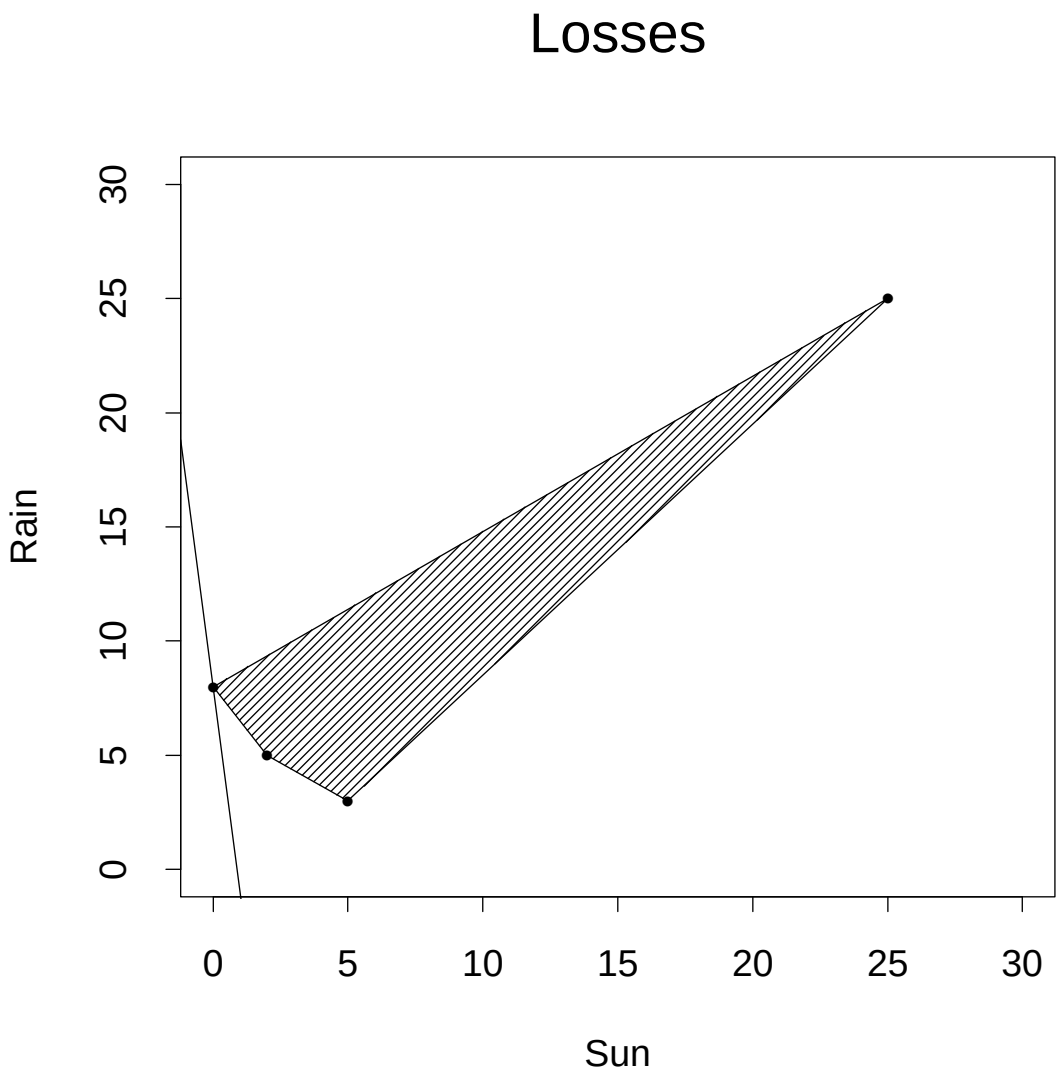
for all θ and there is at least one value of θ where the inequality is strict. A rule which is not inadmissible is called **admissible***.*

The admissible procedures have risks on the lower left of the graphs above. That is, the two lines connecting B to T and T to C are the admissible procedures.

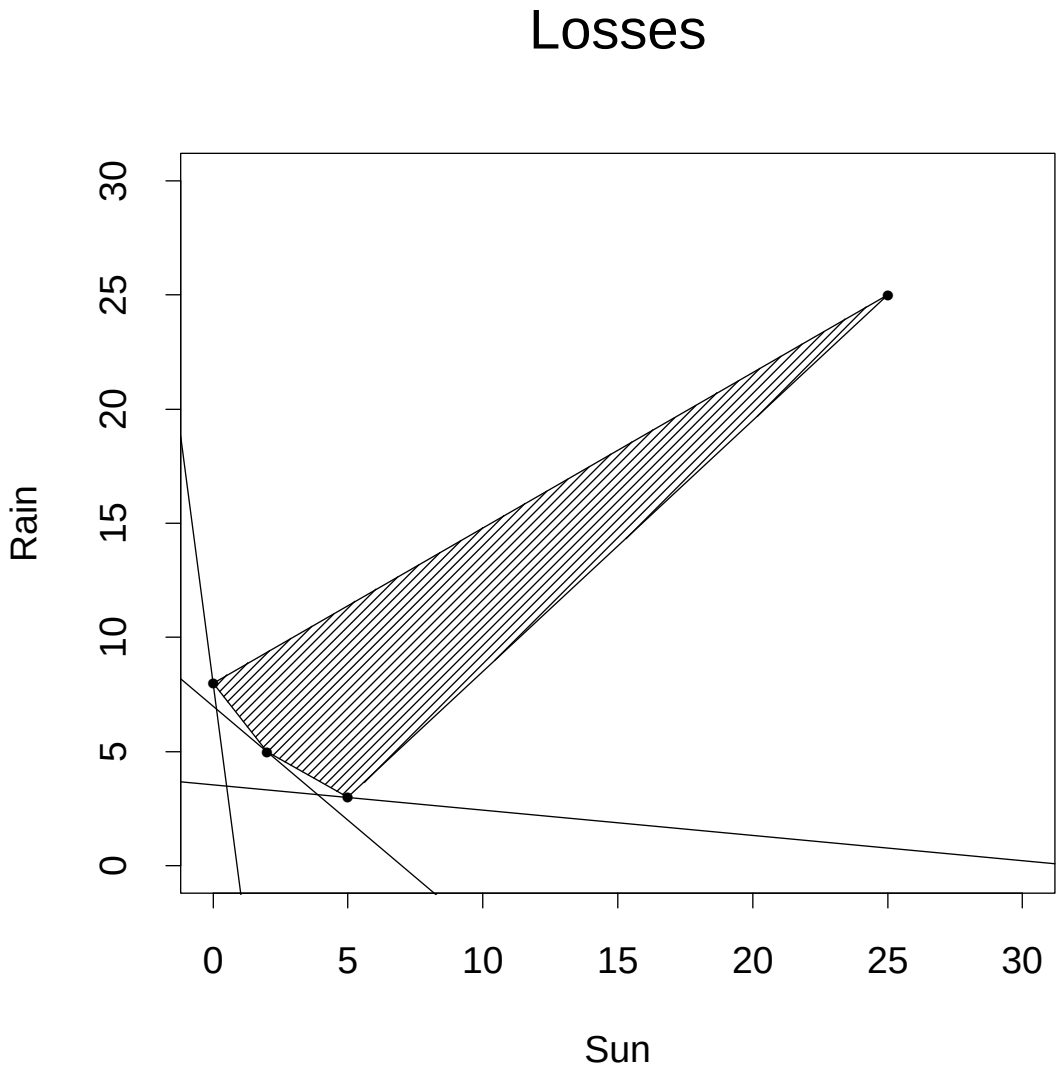
There is a connection between Bayes procedures and admissible procedures. A prior distribution in our example problem is specified by two probabilities, π_R and π_S which add up to 1. If $L = (L_R, L_S)$ is the risk function for some procedure then the Bayes risk is

$$r_\pi = \pi_R L_R + \pi_S L_S$$

Consider the set of L such that this Bayes risk is equal to some constant. On our picture this is a straight line with slope $-\pi_R/\pi_S$. Consider now three priors $\pi_1 = (0.9, 0.1)$, $\pi_2 = (0.5, 0.5)$ and $\pi_3 = (0.1, 0.9)$. For say π_1 imagine a line with slope -9 =0.9/0.1 starting on the far left of the picture and sliding right until it bumps into the convex set of possible losses in the previous picture. It does so at point B as shown in the next graph. Sliding this line to the right corresponds to making the value of r_π larger and larger so that when it just touches the convex set we have found the Bayes procedure.

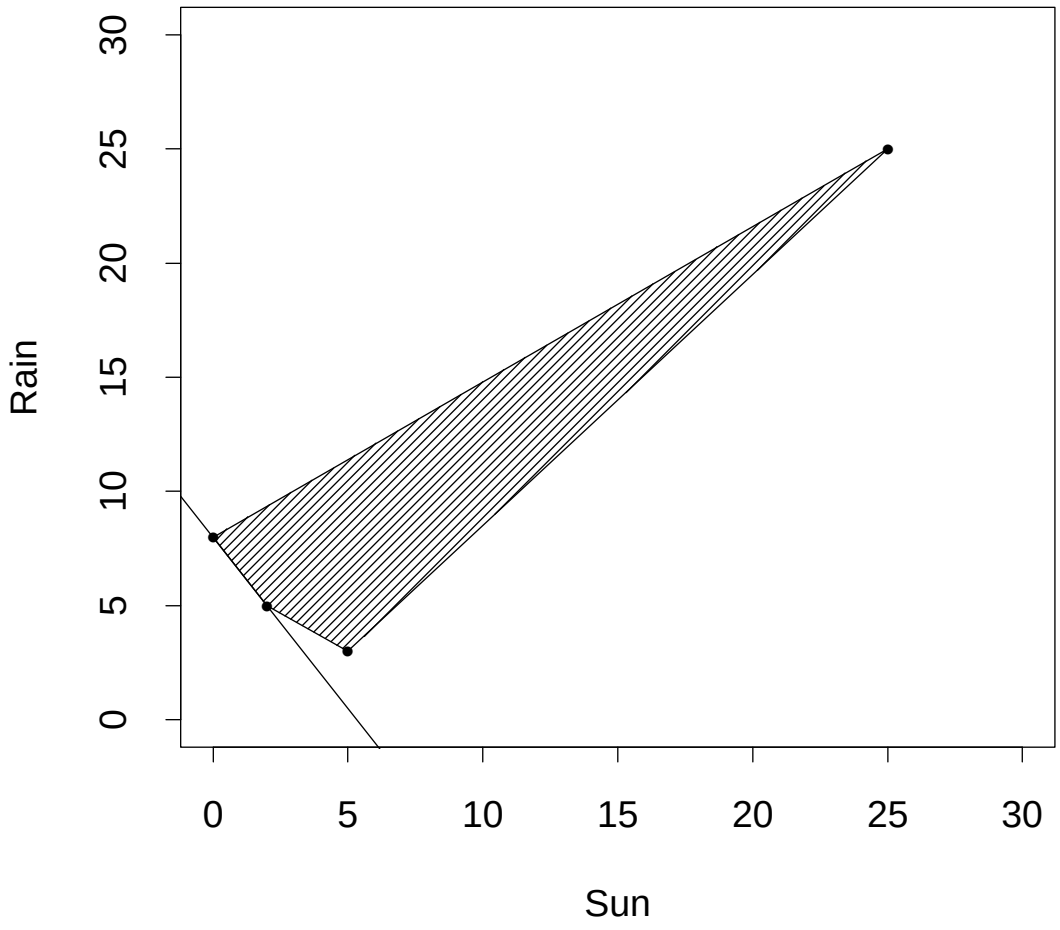


Here is a picture showing the same lines for the three priors above.



We see that the Bayes procedure for π_1 (you are pretty sure it will be sunny) is to ride your bike. If it's a toss up between R and S you take the bus. If R is very likely you take your car.

The special prior $(0.6, 0.4)$ *produces the line shown here:*



You can see that any point on the line connecting B to T is Bayes for this prior.

The ideas here can be used to prove the following general facts:

- *Every admissible procedure is Bayes. (Proof uses the Separating Hyperplane Theorem in Functional Analysis.)*
- *Every Bayes procedure is admissible. (Proof: If δ is Bayes for π but not admissible there is a δ^* such that*

  $$R_{\delta^*}(\theta) \leq R_\delta(\theta)$$

  Multiply by the prior density and integrate to deduce

  $$r_\pi(\delta^*) \leq r_\pi(\delta)$$

  If there is a θ for which the inequality involving R is strict and if the density of π is positive at that θ then the inequality for r_π is strict which would contradict

the hypothesis that δ is Bayes for π. Notice that this theorem actually requires the extra hypothesis about the positive density.)

- *A minimax procedure is admissible. (Actually there can be several minimax procedures and the claim is that at least one of them is admissible. When the parameter space is infinite it might happen that set of possible risk functions is not closed; if not then we have to replace the notion of admissible by some notion of nearly admissible.)*
- *The minimax procedure has constant risk. Actually the admissible minimax procedure is Bayes for some π and its risk is constant on the set of θ for which the prior density is positive.*

Bayesian estimation

Now let's focus on the problem of estimation of a 1 dimensional parameter. Mean Squared Error corresponds to using

$$L(d, \theta) = (d - \theta)^2 .$$

The risk function of a procedure (estimator) $\hat{\theta}$ is

$$R_{\hat{\theta}}(\theta) = E_\theta[(\hat{\theta} - \theta)^2]$$

Now consider using a prior with density $\pi(\theta)$. The Bayes risk of $\hat{\theta}$ is

$$\begin{aligned} r_\pi &= \int R_{\hat{\theta}}(\theta)\pi(\theta)d\theta \\ &= \int\int (\hat{\theta}(x) - \theta)^2 f(x;\theta)\pi(\theta)dxd\theta \end{aligned}$$

Decision Theory and Bayesian Methods

- *Decision space is the set of possible actions I might take. We assume that it is convex, typically by expanding a basic decision space D to the space $\mathcal{D}$ of all probability distributions on D.*
- *Parameter space Θ of possible "states of nature".*
- *Loss function $L = L(d, \theta)$ which is the loss I incur if I do d and θ is the true state of nature.*
- *We call δ^* minimax if*

$$\max_\theta L(\delta^*, \theta) = \min_\delta \max_\theta L(\delta, \theta) .$$

- *A* **prior** *is a probability distribution π on Θ,.*

- *The Bayes risk of a decision δ for a prior π is*

$$r_\pi(\delta) = E_\pi(L(d,\theta)) = \int L(\delta,\theta)\pi(\theta)d\theta$$

if the prior has a density. For finite parameter spaces Θ the integral is a sum.

- *A decision δ^* is Bayes for a prior π if*

$$r_\pi(\delta^*) \leq r_\pi(\delta)$$

for any decision δ.

- *When the parameter space is infinite we sometimes have to consider prior "densities" which are not really densities because they have integral equal to ∞. A positive function on Θ is called a proper prior if it has a finite integral; in this case we divide through by that integral to get a density. A positive function on Θ whose integral is infinite is an* **improper** *prior density.*

- *A decision δ is* **inadmissible** *if there is a decision δ^* such that*

$$L(\delta^*,\theta) \leq L(\delta,\theta)$$

for all θ and there is at least one value of θ where the inequality is strict. A decision which is not inadmissible is called **admissible**.

- *Every admissible procedure is Bayes, perhaps only for an improper prior. (Proof uses the Separating Hyperplane Theorem in Functional Analysis.)*

- *Every Bayes procedure with finite Bayes risk (for a prior which has a density which is positive for all θ) is admissible. (Proof: If δ is Bayes for π but not admissible there is a δ^* such that*

$$L(\delta^*,\theta) \leq L(\delta,\theta)$$

Multiply by the prior density and integrate to deduce

$$r_\pi(\delta^*) \leq r_\pi(\delta)$$

If there is a θ for which the inequality involving L is strict and if the density of π is positive at that θ then the inequality for r_π is strict which would contradict the hypothesis that δ is Bayes for π. Notice that this theorem actually requires the extra hypothesis about the positive density and that the risk functions of δ and δ^* be continuous.)

- *A minimax procedure is admissible. (Actually there can be several minimax procedures and the claim is that at least one of them is admissible. When the parameter space is infinite it might happen that set of possible risk functions is not closed; if not then we have to replace the notion of admissible by some notion of nearly admissible.)*

- *The minimax procedure has constant risk. Actually the admissible minimax procedure is Bayes for some π and its risk is constant on the set of θ for which the prior density is positive.*

Statistical Decision Theory

Statistical problems have another ingredient, the data. We observe X a random variable taking values in say $\mathcal{X}$. We may make our decision d depend on X. A **decision rule** *is a function $\delta(X)$ from $\mathcal{X}$ to D. We will want $L(\delta(X), \theta)$ to be small for all θ. Since X is random we quantify this by averaging over X and compare procedures δ in terms of the* **risk function**

$$R_\delta(\theta) = E_\theta(L(\delta(X), \theta))$$

To compare two procedures we must compare two functions of θ and pick "the smaller one". But typically the two functions will cross each other and there won't be a unique 'smaller one'.

Example*: In estimation theory to estimate a real parameter θ we used $D = \Theta$,*

$$L(d, \theta) = (d - \theta)^2$$

and find that the risk of an estimator $\hat{\theta}(X)$ is

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

which is just the Mean Squared Error of $\hat{\theta}$. We have already seen that there is no unique best estimator in the sense of MSE. How do we compare risk functions in general? We extend **minimax** *and* **Bayes** *to risks rather than just losses.*

- **Minimax methods** *choose δ to minimize the worst case risk:*

  $$\sup\{R_\delta(\theta); \theta \in \Theta)\}.$$

  We call δ^* minimax if

  $$\sup_\theta R_{\delta^*}(\theta) = \inf_\delta \sup_\theta R_\delta(\theta)$$

  Usually the suprema and infima are achieved and we write max *for* sup *and* min *for* inf*. This is the source of "minimax".*

- **Bayes methods** *choose δ to minimize an average*

  $$r_\pi(\delta) = \int R_\delta(\theta)\pi(\theta)d\theta$$

  for a suitable density π. We call π a **prior** *density and r the* **Bayes** *risk of δ for the prior π.*

Definition: *A decision rule δ is* **inadmissible** *if there is a rule δ^* such that*

$$R_{\delta^*}(\theta) \leq R_\delta(\theta)$$

for all θ and there is at least one value of θ where the inequality is strict. A rule which is not inadmissible is called **admissible**.

Bayesian estimation

Now let's focus on the problem of estimation of a 1 dimensional parameter. Mean Squared Error corresponds to using

$$L(d, \theta) = (d - \theta)^2 .$$

The risk function of a procedure (estimator) $\hat{\theta}$ is

$$R_{\hat{\theta}}(\theta) = E_\theta[(\hat{\theta} - \theta)^2]$$

Now consider using a prior with density $\pi(\theta)$. The Bayes risk of $\hat{\theta}$ is

$$\begin{aligned} r_\pi &= \int R_{\hat{\theta}}(\theta)\pi(\theta)d\theta \\ &= \int\int (\hat{\theta}(x) - \theta)^2 f(x;\theta)\pi(\theta)dxd\theta \end{aligned}$$

How should we choose $\hat{\theta}$ to minimize r_π? The solution lies in recognizing that $f(x;\theta)\pi(\theta)$ is really a joint density

$$\int\int (x;\theta)\pi(\theta)dxd\theta = 1$$

For this joint density the conditional density of X given θ is just the model $f(x;\theta)$. From now on I write the model as $f(x|\theta)$ to emphasize this fact. We can now compute r_π a different way by factoring the joint density a different way:

$$f(x|\theta)\pi(\theta) = \pi(\theta|x)f(x)$$

where now $f(x)$ is the marginal density of x and $\pi(\theta|x)$ denotes the conditional density of θ given X. We call $\pi(\theta|x)$ the **posterior density.** *It is found via Bayes theorem (which is why this is Bayesian statistics):*

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{\int f(x|\phi)\pi(\phi)d\phi}$$

With this notation we can write

$$r_\pi(\hat{\theta}) = \int \left[\int (\hat{\theta}(x) - \theta)^2 \pi(\theta|x)d\theta\right] f(x)dx$$

Now we can choose $\hat{\theta}(x)$ separately for each x to minimize the quantity in square brackets (as in the NP lemma). The quantity in square brackets is a quadratic function of $\hat{\theta}(x)$ and can be seen to be minimized by

$$\hat{\theta}(x) = \int \theta \pi(\theta|x) d\theta$$

which is

$$E(\theta|X)$$

and is called the **posterior expected mean** *of θ.*

Example*: Consider first the problem of estimating a normal mean μ. Imagine, for example that μ is the true speed of sound. I think this is around 330 metres per second and am pretty sure that I am within 30 metres per second of the truth with that guess. I might summarize my opinion by saying that I think μ has a normal distribution with mean ν =330 and standard deviation $\tau = 10$. That is, I take a prior density π for μ to be $N(\nu, \tau^2)$. Before I make any measurements my best guess of μ minimizes*

$$\int (\hat{\mu} - \mu)^2 \frac{1}{\tau\sqrt{2\pi}} \exp\{-(\mu - \nu)^2/(2\tau^2)\} d\mu$$

This quantity is minimized by the prior mean of μ, namely,

$$\hat{\mu} = E_\pi(\mu) = \int \mu \pi(\mu) d\mu = \nu \, .$$

Now suppose we collect 25 measurements of the speed of sound. I will assume that the relationship between the measurements and μ is that the measurements are unbiased and that the standard deviation of the measurement errors is $\sigma = 15$ which I assume that we know. Thus the model is that conditional on μ $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$. The joint density of the data and μ is then

$$(2\pi)^{-n/1} \sigma^{-n} \exp\{-\sum (X_i - \mu)^2/(2\sigma^2)\} (2\pi)^{-1/2} \tau^{-1} \exp\{-(\mu - \nu)^2/\tau^2\}$$

This is a multivariate normal joint density for $(X_1, \ldots, X_n, \mu)$ so the conditional distribution of θ given $X_1, \ldots, X_n$ is normal. Standard multivariate normal formulas can be used to calculate the conditional means and variances. Alternatively the exponent in the joint density is of the form

$$-\frac{1}{2} \left[\mu^2/\gamma^2 - 2\mu\psi/\gamma^2\right]$$

plus terms not involving μ where

$$\frac{1}{\gamma^2} = \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)$$

and

$$\frac{\psi}{\gamma^2} = \frac{\sum X_i}{\sigma^2} + \frac{\nu}{\tau^2}$$

This means that the conditional density of μ given the data is $N(\psi, \gamma^2)$. In other words the posterior mean of μ is

$$\frac{\frac{n}{\sigma^2}\bar{X} + \frac{1}{\tau^2}\nu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

which is a weighted average of the prior mean ν and the sample mean $\bar{X}$. Notice that the weight on the data is large when n is large or σ is small (precise measurements) and small when τ is small (precise prior opinion).

Improper priors*: When the density does not integrate to 1 we can still follow the machinery of Bayes' formula to derive a posterior. For instance in the $N(\mu, \sigma^2)$ example consider the prior density $\pi(\theta) \equiv 1$. This "density" integrates to infty but using Bayes' theorem to compute the posterior would give*

$$\pi(\theta|X) = \frac{(2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i - \mu)^2/(2\sigma^2)\}}{\int (2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i - \nu)^2/(2\sigma^2)\}d\nu}$$

It is easy to see that this cancels to the limit of the case previously done when $\tau \to \infty$ giving a $N(\bar{X}, \sigma^2/n)$ density. That is, the Bayes estimate of μ for this improper prior is $\bar{X}$.

Admissibility*: ayes procedures for proper priors are admissible. It follows that for each $w \in (0, 1)$ and each real ν the estimate*

$$w\bar{X} + (1 - w)\nu$$

is admissible. That this is also true for $w = 1$, that is, that $\bar{X}$ is admissible is much harder to prove.

Minimax estimation*: The risk function of $\bar{X}$ is simply σ^2/n. That is, the risk function is constant since it does not depend on μ. Were $\bar{X}$ Bayes for a proper prior this would prove that $\bar{X}$ is minimax. In fact this is also true but hard to prove.*

Example*: Suppose that given p X has a Binomial(n, p) distribution. We will give p a Beta(α, β) prior density*

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1 - p)^{\beta-1}$$

The joint "density" of X and p is

$$\binom{n}{X}p^X(1 - p)^{n-X}\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1 - p)^{\beta-1}$$

so that the posterior density of p given X is of the form

$$cp^{X+\alpha-1}(1 - p)^{n-X+\beta-1}$$

for a suitable normalizing constant c. But this is a Beta$(X + \alpha, n - X + \beta)$ density. The mean of a Beta(α, β) distribution is $\alpha/(\alpha + \beta)$. Thus the Bayes estimate of p is

$$\frac{X + \alpha}{n + \alpha + \beta} = w\hat{p} + (1 - w)\frac{\alpha}{\alpha + \beta}$$

where $\hat{p} = X/n$ is the usual mle. Notice that this is again a weighted average of the prior mean and the mle. Notice also that the prior is proper for $\alpha > 0$ and $\beta > 0$. To get $w = 1$ we take $\alpha = \beta = 0$ and use the improper prior

$$\frac{1}{p(1-p)}$$

Again we learn that each $w\hat{p} + (1-w)p_o$ is admissible for $w \in (0,1)$. Again it is true that $\hat{p}$ is admissible but that our theorem is not adequate to prove this fact.

The risk function of $w\hat{p} + (1-w)p_0$ is

$$R(p) = E[(w\hat{p} + (1-w)p_0 - p)^2]$$

which is

$$w^2 Var(\hat{p}) + (wp + (1-w)p - p)^2 = w^2 p(1-p)/n + (1-w)^2(p-p_0)^2$$

This risk function will be constant if the coefficients of both p^2 and of p in the risk are 0. The coefficient of p is

$$-w^2/n + (1-w)^2$$

so $w = n^{1/2}/(1+n^{1/2})$. The coefficient of p is then

$$w^2/n - 2p_0(1-w)^2$$

which will vanish if $2p_0 = 1$ or $p_0 = 1/2$. Working backwards we find that to get these values for w and p_0 we require $\alpha = \beta$. Moreover the equation

$$w^2/(1-w)^2 = n$$

gives

$$n/(\alpha+\beta) = \sqrt{n}$$

or $\alpha = \beta = \sqrt{n}/2$. The minimax estimate of p is

$$\frac{\sqrt{n}}{1+\sqrt{n}}\hat{p} + \frac{1}{1+\sqrt{n}}\frac{1}{2}$$

Example*: Now suppose that $X_1, \ldots, X_n$ are iid $MVN(\mu, \Sigma)$ with Σ known. Consider as the improper prior for μ which is constant. The posterior density of μ given X is then $MVN(\bar{X}, \Sigma/n)$.*

For multivariate estimation it is common to extend the notion of squared error loss by defining

$$L(\hat{\theta}, \theta) = \sum(\hat{\theta}_i - \theta_i)^2 = (\hat{\theta} - \theta)^t(\hat{\theta} - \theta)\,.$$

For this loss function the risk is the sum of the MSEs of the individual components and the Bayes estimate is the posterior mean again. Thus $\bar{X}$ is Bayes for an improper prior in this problem. It turns out that $\bar{X}$ is minimax; its risk function is the constant $trace(\Sigma)/n$. If the dimension p of θ is 1 or 2 then $\bar{X}$ is also admissible but if $p \geq 3$ then it is inadmissible. This fact was first demonstrated by James and Stein who produced an estimate which is better, in terms of this risk function, for every μ. The "improved" estimator, called the James Stein estimator is essentially never used.

Statistical Decision Theory: examples

Example: *In estimation theory to estimate a real parameter θ we used $D = \Theta$,*

$$L(d, \theta) = (d - \theta)^2$$

and find that the risk of an estimator $\hat{\theta}(X)$ is

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

which is just the Mean Squared Error of $\hat{\theta}$.

The Bayes estimate of θ is $E(\theta|X)$, the posterior mean of θ.

Example: *In $N(\mu, \sigma^2)$ model with σ known a common prior is $\mu \sim N(\nu, \tau^2)$. The resulting posterior distribution is Normal with posterior mean*

$$E(\mu|X) = \frac{\frac{n}{\sigma^2}\bar{X} + \frac{1}{\tau^2}\nu}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$$

and posterior variance $1/(n/\sigma^2 + 1/\tau^2)$.

Improper priors: *When the density does not integrate to 1 we can still follow the machinery of Bayes' formula to derive a posterior. For instance in the $N(\mu, \sigma^2)$ example consider the prior density $\pi(\theta) \equiv 1$. This "density" integrates to ∞ but using Bayes' theorem to compute the posterior would give*

$$\pi(\theta|X) = \frac{(2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i - \mu)^2/(2\sigma^2)\}}{\int(2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i - \nu)^2/(2\sigma^2)\}d\nu}$$

It is easy to see that this cancels to the limit of the case previously done when $\tau \to \infty$ giving a $N(\bar{X}, \sigma^2/n)$ density. That is, the Bayes estimate of μ for this improper prior is $\bar{X}$.

Admissibility: *Bayes procedures with finite Bayes risk and continuous risk functions are admissible. It follows that for each $w \in (0, 1)$ and each real ν the estimate*

$$w\bar{X} + (1 - w)\nu$$

is admissible. That this is also true for $w = 1$, that is, that $\bar{X}$ is admissible, is much harder to prove.

Minimax estimation: *The risk function of $\bar{X}$ is simply σ^2/n. That is, the risk function is constant since it does not depend on μ. Were $\bar{X}$ Bayes for a proper prior this would prove that $\bar{X}$ is minimax. In fact this is also true but hard to prove.*

Example: *Suppose that given p X has a Binomial(n, p) distribution. We will give p a Beta(α, β) prior density*

$$\pi(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)}p^{\alpha-1}(1 - p)^{\beta-1}$$

The joint "density" of X and p is

$$\binom{n}{X} p^X (1-p)^{n-X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}$$

so that the posterior density of p given X is of the form

$$c p^{X+\alpha-1}(1-p)^{n-X+\beta-1}$$

for a suitable normalizing constant c. But this is a Beta$(X+\alpha, n-X+\beta)$ density. The mean of a Beta(α,β) distribution is $\alpha/(\alpha+\beta)$. Thus the Bayes estimate of p is

$$\frac{X+\alpha}{n+\alpha+\beta} = w\hat{p} + (1-w)\frac{\alpha}{\alpha+\beta}$$

where $\hat{p} = X/n$ is the usual mle. Notice that this is again a weighted average of the prior mean and the mle. Notice also that the prior is proper for $\alpha > 0$ and $\beta > 0$. To get $w = 1$ we take $\alpha = \beta = 0$ and use the improper prior

$$\frac{1}{p(1-p)}$$

Again we learn that each $w\hat{p} + (1-w)p_o$ is admissible for $w \in (0,1)$. Again it is true that $\hat{p}$ is admissible but that our theorem is not adequate to prove this fact.

The risk function of $w\hat{p} + (1-w)p_0$ is

$$R(p) = E[(w\hat{p} + (1-w)p_0 - p)^2]$$

which is

$$w^2 Var(\hat{p}) + (wp + (1-w)p - p)^2 = w^2 p(1-p)/n + (1-w)^2(p-p_0)^2$$

This risk function will be constant if the coefficients of both p^2 and of p in the risk are 0. The coefficient of p is

$$-w^2/n + (1-w)^2$$

so $w = n^{1/2}/(1+n^{1/2})$. The coefficient of p is then

$$w^2/n - 2p_0(1-w)^2$$

which will vanish if $2p_0 = 1$ or $p_0 = 1/2$. Working backwards we find that to get these values for w and p_0 we require $\alpha = \beta$. Moreover the equation

$$w^2/(1-w)^2 = n$$

gives

$$n/(\alpha+\beta) = \sqrt{n}$$

or $\alpha = \beta = \sqrt{n}/2$. *The minimax estimate of* p *is*

$$\frac{\sqrt{n}}{1+\sqrt{n}}\hat{p} + \frac{1}{1+\sqrt{n}}\frac{1}{2}$$

Example*: Now suppose that* $X_1, \ldots, X_n$ *are iid* $MVN(\mu, \Sigma)$ *with* Σ *known. Consider as the improper prior for* μ *which is constant. The posterior density of* μ *given* X *is then* $MVN(\bar{X}, \Sigma/n)$.

For multivariate estimation it is common to extend the notion of squared error loss by defining

$$L(\hat{\theta}, \theta) = \sum (\hat{\theta}_i - \theta_i)^2 = (\hat{\theta} - \theta)^t(\hat{\theta} - \theta)\,.$$

For this loss function the risk is the sum of the MSEs of the individual components and the Bayes estimate is the posterior mean again. Thus $\bar{X}$ *is Bayes for an improper prior in this problem. It turns out that* $\bar{X}$ *is minimax; its risk function is the constant* $trace(\Sigma)/n$. *If the dimension* p *of* θ *is 1 or 2 then* $\bar{X}$ *is also admissible but if* $p \geq 3$ *then it is inadmissible. This fact was first demonstrated by James and Stein who produced an estimate which is better, in terms of this risk function, for every* μ. *The "improved" estimator, called the James Stein estimator, is essentially never used.*

Hypothesis Testing and Decision Theory

One common decision analysis of hypothesis testing takes $D = \{0, 1\}$ *and* $L(d, \theta) = 1(make\ an\ error)$ *or more generally* $L(0, \theta) = \ell_1 1(\theta \in \Theta_1)$ *and* $L(1, \theta) = \ell_2 1(\theta \in \Theta_0)$ *for two positive constants* ℓ_1 *and* ℓ_2. *We make the decision space convex by allowing a decision to be a probability measure on* D. *Any such measure can be specified by* $\delta = P(reject)$ *so* $\mathcal{D} = [0, 1]$. *The loss function of* $\delta \in [0, 1]$ *is*

$$L(\delta, \theta) = (1 - \delta)\ell_1 1(\theta \in \Theta_1) + \delta\ell_0 1(\theta \in \Theta_0)\,.$$

Simple hypotheses*: A prior is just two numbers* π_0 *and* π_1 *which are non-negative and sum to 1. A procedure is a map from the data space to* $\mathcal{D}$ *which is exactly what a test function was. The risk function of a procedure* $\phi(X)$ *is a pair of numbers:*

$$R_\phi(\theta_0) = E_0(L(\delta, \theta_0))$$

and

$$R_\phi(\theta_1) = E_1(L(\delta, \theta_1))$$

We find

$$R_\phi(\theta_0) = \ell_0 E_0(\phi(X)) = \ell_0\alpha$$

and

$$R_\phi(\theta_1) = \ell_1 E_1(1 - \phi(X)) = \ell_1\beta$$

The Bayes risk of ϕ *is*

$$\pi_0\ell_0\alpha + \pi_1\ell_1\beta$$

We saw in the hypothesis testing section that this is minimized by

$$\phi(X) = 1(f_1(X)/f_0(X) > \pi_1\ell_1/(\pi_0\ell_0))$$

which is a likelihood ratio test. These tests are Bayes and admissible. The risk is constant if $\beta\ell_1 = \alpha\ell_0$; you can use this to find the minimax test in this context.

Composite hypotheses*:*