STAT 830: Statistical Theory I

Richard Lockhart

November 20, 2012

Chapter 1

Introduction

1.1 Statistics versus Probability

Statistics versus Probability

I want to begin this course by discussing the difference between Probability Theory and Statistics. Statisticians use the tools of Probability but reason from effects to causes rather than from causes to effects. I want to try to say that again with a bit more detail but still in a vague sort of way.

The standard view of scientific inference starts with a set of theories which make predictions about the outcomes of an experiment as in the following table:

Theory	Prediction
A	1
B	2
C	3

Now imagine that we actually conduct the experiment and see outcome 2. We **infer** that theory B is correct (or at least that theories A and C are wrong). The question of how much more faith put in B than before is subtle and has been much discussed. As usual theories can easily be falsified – that is, shown to be wrong. But they are only shown to be right in the sense that we try and fail to falsify them. If a theory makes many many correct predictions in many contexts we start to treat it as if it were true; but one wrong prediction demands a rethink.

Now we add **Randomness** to our little table because the outcomes of experiments are not perfectly predictable, even in theory:

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

Now imagine again that we see outcome 2. We now infer that Theory B is probably correct, that Theory A is probably not correct, and that Theory C is wrong. Notice the precision gained, when Theory C absolutely rules out outcome 2 but outcome 2 actually happens – we can rule out theory C.

That leads me to summarize the difference between Probability and Statistics as follows:

- In **Probability Theory**: we construct the table by computing likely outcomes of experiments. We predict what ought to happen if we do the experiment and some specific theory holds.
- In **Statistics** we follow the inverse process. We use the table to draw inferences from outcome of experiment – deciding how sure we are about which theory is correct. In this course we consider the questions: how should we do draw these inferences and how wrong are our inferences likely to be? Notice: our task is hopeless unless different theories make different predictions – see future discussions of *identifiable* models.

I will start the course with Probability and switch after about 5 weeks to statistics.

Chapter 2

Probability

In this section I want to define the basic objects. I am going to give full precise definitions and make lists of various properties – even prove some things rigorously – but then I am going to give examples. In different versions of this course I require more or less understanding of the objects being studied.

Definition: A **Probability Space** (or **Sample Space**) is an ordered triple $(\Omega, \mathcal{F}, P)$ with the following properties:

- Ω is a set (it is the set of all possible outcomes of some experiment); elements of Ω are denoted by the letter ω. They are called elementary outcomes.
- $\mathcal{F}$ is a family of subsets (we call these subsets **events**) of Ω with the property that $\mathcal{F}$ is a σ-field (or Borel field or σ-algebra) – that is $\mathcal{F}$ has the following **closure** properties:

  1. The empty set denoted $\emptyset$ and Ω are members of $\mathcal{F}$.
  2. $A \in \mathcal{F}$ implies $A^c = \{\omega \in \Omega : \omega \not\in A\} \in \mathcal{F}$.
  3. $A_1, A_2, \cdots$ in $\mathcal{F}$ implies $A = \cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

- P is a function whose domain is $\mathcal{F}$ and whose range is a subset of $[0, 1]$. The function P must satisfy:

  1. $P(\emptyset) = 0$ and $P(\Omega) = 1$.
  2. **Countable additivity**: $A_1, A_2, \cdots$ **pairwise disjoint** $(j \neq k \; A_j \cap A_k = \emptyset)$

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

These axioms guarantee that we can compute probabilities by the usual rules, including approximation. Here are some consequences of the axioms:

$$A_i \in \mathcal{F}; i = 1, 2, \cdots \text{ implies } \cap_i A_i \in \mathcal{F}$$

$$A_1 \subseteq A_2 \subseteq \cdots \text{ implies } P(\cup A_i) = \lim_{n\to\infty} P(A_n)$$

$$A_1 \supset A_2 \supset \cdots \text{ implies } P(\cap A_i) = \lim_{n\to\infty} P(A_n)$$

The last two of these three assertions are sometimes described by saying that P is *continuous*. I don't like this jargon because it does not agree very well with the standard meaning of a continuous function. There is (in what I have presented so far) no well defined *topology* or *metric* or other way to make precise the notion of a sequence of sets converging to a limit.

2.0.1 Examples

It seems wise to list a few examples of these triples which arise in various more or less sophisticated probability problems.

Example 1: Three Cards Problem

I imagine I have three cards – stiff pieces of paper. One card is green on both sides. One is red on both sides. The third card is green on one side and red on the other. I shuffle up the three cards in some container and pick one out, sliding it out of its container and onto the table in such a way that you can see only the colour on the side of the card which is up on the table. Later, when I talk about conditional probability, I will be interested in probabilities connected with the side which is face down on the table but here I just want to list the elements of Ω and describe $\mathcal{F}$ and P.

I want you to imagine that the sides of the card are labelled (in your mind, not visibly on the cards) in such a way that you can see that there are six sides of the card which could end up being the one which is showing. One card, the `RR` card has red on both sides and $\omega_1 = RR1$ means the first of these two sides is showing which $\omega_2 = RR2$ denotes the outcome that the second of these two sides is showing. I use $\omega_3 = RG1$ to denote the outcome where the Red / Green card is selected and the red side is up and $\omega_4 = RG2$ to denote the outcome where the same card is drawn but the green side is up. The remaining two elementary outcomes are $\omega_5 = GG1$ and $\omega_6 = GG2$ in what I hope is quite obvious notation.

So now $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4, \omega_5, \omega_6\}$ is the sample space with six elements. There are many other possible notations for the elements of this sample space of course. I now turn to describing $\mathcal{F}$ and P.

In problems where Ω is finite or countably infinite we almost always take $\mathcal{F}$ to be the family of all possible subsets of Ω. So in this case $\mathcal{F}$ is the collection of all subsets of Ω. To make a subset of Ω we must decide for each of the six elements of Ω whether or not to put that element in the set. This makes 2 possible choices for ω_1, then for each of these 2 choices for ω_2 and so on. So there are $2^6 = 64$ subsets of Ω; all 64 are in $\mathcal{F}$. In order to be definite I will try to list the pattern:

$$\mathcal{F} = \{\emptyset, \{\omega_1\}, \ldots, \{\omega_6\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \ldots, \{\omega_5, \omega_6\}, \ldots, \Omega\}$$

My list includes 1 set with 0 elements, 6 sets with 1 element, 6 choose 2 sets with 2 elements (total of 15), 6 choose 3 with 3 elements (20 such), 6 choose 4 (=15) with 4 elements, 6 with 5 elements and Ω.

Finally I am supposed to describe P. The usual way, when Ω is finite, to assign probabilities is to give some probability, say p_i to the ith elementary outcome ω_i. In our case it is reasonable to assume that all 6 sides of the cards have the same chance of ending up visible so all

$$p_i = P(\{\omega_i\}) = \frac{1}{6}.$$

Then the probability of any subset of Ω is found by adding up the probabilities of the elementary outcomes in that set. So, for instance

$$P(\{\omega_1, \omega_3, \omega_4\}) = \frac{3}{6} = \frac{1}{2}.$$

The event "the side showing is red" is a subset of Ω, namely,

$$\{\omega_1, \omega_2, \omega_3\}.$$

The event "the side face down is red" is also subset of Ω, namely,

$$\{\omega_1, \omega_2, \omega_4\}.$$

The event "the side face down is green" is

$$\{\omega_3, \omega_5, \omega_6\}.$$

Example 2: Coin Tossing till First Head Problem

Now imagine tossing a coin until you get "heads" which I denote H. To simplify the problem I will assume that you quit tossing either when you get H OR when you have tossed the coin three times without getting H. Letting T denote tails the elements of Ω are, in obvious notation:

$$\{\omega_1, \omega_2, \omega_3, \omega_4\} \equiv \{H, TH, TTH, TTT\}$$

Again $\mathcal{F}$ is the collection of all $2^4 = 16$ subsets of Ω and we specify P by assigning probabilities to elementary outcomes. The most natural probabilities to assign are $p_1 = 1/2$, $p_2 = 1/4$ and $p_3 = p_4 = 1/8$. I will return to this assumption when I discuss independence.

Example 3: Coin Tossing till First Head Problem, infinite case

Now imagine tossing the coin until you get "heads" no matter how many tosses are required. Let ω_k be a string of k tails T followed by H. Then

$$\Omega = \{\omega_0, \omega_1, \omega_2, \cdots\}$$

which has infinitely many elements. Again $\mathcal{F}$ is the collection of all subsets of Ω; the number of such subsets is uncountably infinite so I won't make a list! We specify P by assigning probabilities to elementary outcomes. In order to add a bit to the example I will consider a biased coin. The most natural probabilities to assign are then

$$p_i = P(\{\omega_i\}) = p(1-p)^i.$$

This list of numbers adds up to 1, as it must, to ensure $P(\Omega) = 1$; you should recognize the sum of a geometric series.

Example 4: Coin Tossing forever

In order to discuss such things as the law of large numbers and many other probability problems it is useful to imagine the conceptual experiment of tossing the coin forever. In this case a single “elementary outcome”, ω is actually an infinite sequence of Hs and Ts. One ω might be

$$HTHTHTHTHTHTHT\cdots$$

where the heads and tails alternate for ever. It would be typical to say

$$\Omega = \{\omega = (\omega_1, \omega_2, \ldots);\ \text{such that each } \omega_i \in \{H, T\}\}.$$

You can think about how many elements there are in Ω by taking a typical ω and replacing each H with a 1, then each T with a 0. Then put “0.” in front and think of the result as a binary number between 0 and 1. So for instance the sequence above of alternating 0s and 1s is

$$\omega = 0.1010101010\cdots = \frac{1}{2}\left(1 + \frac{1}{4} + \left(\frac{1}{4}\right)^2 + \cdots\right)$$

which is just 2/3 by summing a geometric series.

The summary is that there are as many elements in Ω as there are numbers between 0 and 1 – an uncountably infinite number. It turns out that this is the situation where we just can’t cope, logically, with having $\mathcal{F}$ be the collection of *all* subsets of Ω. If you want to know which subsets go into $\mathcal{F}$ you need to find out about *Borel* sets.

In fact we take $\mathcal{F}$ to be “the smallest σ-field” which contains all sets of the form

$$B_i \equiv \{\omega \in \Omega : \omega_i = H\}$$

which is the subset of Ω obtained by keeping only outcomes whose ith toss is H. There is a bit of mathematical effort to prove the existence of any such “smallest” σ-field; it is the intersection of all σ-fields which contain the given special sets. Much greater effort is needed to understand the structure of this σ-field but I want to emphasize that if you can give a truly clear and explicit description of a subset of Ω that subset will be a Borel set – a member of $\mathcal{F}$.

Finally we have to say something about how to compute probabilities. Let’s start with an intuitive presentation using the idea that we might be talking about independent tosses of a fair coin; I will define independence precisely later but for now I just want you to use what you already know about independent events. Let

$$C = B_1 \cap B_2^c \cap B_3 \cap B_4^c \cap B_5 \cap B_6^c \cdots .$$

The only point in C is the sequence of alternating heads and tails I wrote down up above. So what is the probability of C. Certainly

$$P(C) \geq P(B_1 \cap B_2^c \cap B_3 \cap B_4^c \cap B_5 \cap B_6^c \cdots B_{2n}^c)$$

for any n. For independent tosses of a fair coin we compute the probability of this intersection by just multiplying 1/2 by itself $2n$ times to get 2^{-n}. But if $P(C) \leq 2^{-n}$ for all n then $P(C) = 0$. In the same way we can check that $P(\{\omega\}) = 0$ for every elementary outcome ω!

This just means we *cannot* compute probabilities of an event by adding up probabilities of elementary outcomes in the event – that always gives 0. Instead we use the idea of independence and the *assumption* that the various B_i are independent and have probability 1/2 to compute any probability we want; sometimes this is *hard*.

2.1 Random Variables

:

Definition: A **Vector valued random variable** is a function function $X : \Omega \mapsto R^p$ such that, writing $X = (X_1, \ldots, X_p)$,

$$P(X_1 \le x_1, \ldots, X_p \le x_p)$$

is defined for any constants $(x_1, \ldots, x_p)$. Formally the notation

$$X_1 \le x_1, \ldots, X_p \le x_p$$

describes a subset of Ω or **event**:

$$\{\omega \in \Omega : X_1(\omega) \le x_1, \ldots, X_p(\omega) \le x_p\} \ .$$

Remember X is a function on Ω so X_1 is also a function on Ω; that is why we can stick in the argument ω of the function.

ASIDE: 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 $X(\omega)$.

There is a subtle mathematical point being made here. Not every function from Ω to R^p is a random variable or random vector. The problem is that the set

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

might not be in $\mathcal{F}$! For our fourth example this is a potential mathematical (but not practical) problem.

2.1.1 Borel sets

In this subsection I give a small presentation of the notion of Borel sets in R^p. The material is not really part of this course.

Definition: The **Borel** σ-field in R^p is the smallest σ-field in R^p containing every open ball.

Definition: For clarity the open ball of radius $r > 0$ centred at $x \in R^p$ is

$$\{y \in R^p : ||y - x|| < r\}$$

where

$$||u|| = \sqrt{\sum_1^p u_i^2}$$

for a vector $u \in R^p$. The quantity $||u||$ is called the Euclidean norm of u; it is also the usual notion of length of a vector.

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 $\{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$. This is equivalent to

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

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 $\mathcal{F}$. We also write

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$$

and call this set the *inverse image* of A under X. So the distribution of X is

$$P_X(A) = P(X^{-1}(A))$$

which is defined for all Borel sets $A \in R^p$.
Remark: The definition of a random variable depends only on the functions and the σ-fields involved and NOT on the probability P.

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

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

I will not always use the subscript X to indicate which random vector is being discussed. When there is no real possibility of confusion I will just write F.

Here are some properties of F for $p = 1$:

1. $0 \le F(x) \le 1$.

2. $x > y \Rightarrow F(x) \ge F(y)$ (monotone non-decreasing).

3. $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$.

4. $\lim_{x \searrow y} F(x) = F(y)$ (right continuous).

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

6. $F(x) - F(x-) = P(X = x)$.

7. $F_X(t) = F_Y(t)$ 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.

Proof: The values of F are probabilities so they are between 0 and 1. If F is the cdf of X and $y < x$ then

$$\{X \le y\} \subseteq \{X \le x\}$$

so

$$F(y) = P(X \le y) \le P(X \le x) = F(x).$$

Since F is monotone the assertions about limits may be checked by considering a sequence x_n. For instance, to prove the first half of the third assertion we take x_n to be any sequence decreasing to $-\infty$ – such as $x_n = -n$, say. If

$$A_n = \{X \le x_n\}$$

then

$$A_1 \supseteq A_2 \supseteq \cdots$$

and

$$\cap_{n=1}^\infty A_n = \emptyset$$

so by the "continuity" of P

$$0 = P(\emptyset) = \lim_{n\to\infty} P(A_n) = \lim_{n\to\infty} F(x_n).$$

The argument at ∞ uses unions in place of intersections and a sequence x_n increasing to ∞.

Assertion 4 considers a sequence x_n decreasing to y and then with the A_i as above we find

$$\cap_{n=1}^\infty A_n = \{X \le y\}$$

so that right continuity of F comes from the continuity of P. Assertion 5 does the parallel thing with unions and shows $F(y-) = P(X < y)$.

Assertion 6 comes from the fact that

$$\{X < x\} \cup \{X = x\} = \{X \le x\}.$$

The union is disjoint so

$$F(y-) + P(X = x) = F(y).$$

The final point, property 7, is much more sophisticated – much harder to prove. If you want to read about it you can look at the appendix on Monotone Class arguments if I ever get it done. •

For $p = 1$ any function F with properties 1, 2, 3 and 4 is the cumulative distribution function of some random variable X. For $p > 1$ the situation is a bit more complicated. Consider the case $p = 2$ and two points (u_1, u_2) and (v_1, v_2). If $v_1 \ge u_1$ and $v_2 \ge u_2$ then the event $X_1 \le u_1, X_2 \le u_2$ is a subset of the event $X_1 \le v_1, X_2 \le v_2$. This means that

$$F(u_1, u_2) = P(X_1 \le u_1, X_2 \le u_2) \le P(X_1 \le v_1, X_2 \le v_2) = F(v_1, v_2).$$

In this sense F is monotone non-decreasing. But even if F is continuous, monotone non-decreasing and satisfies properties 1 and 3 above we cannot be sure it is a cdf. Think about the rectangle

$$R \equiv \{(x_1, x_2) : u_1 < x_1 \le v_1, u_2 < x_2 \le v_2\}$$

The probability that X lands in this rectangle must be at least 0 but in terms of F you should be able to check that

$$\begin{aligned} P(X \in R) &= P(u_1 < X_1 \le v_1, u_2 < X_2 \le v_2) \\ &= F(v_1, v_2) - F(u_1, v_2) - F(v_1, u_2) + F(u_1, u_2). \end{aligned}$$

So this combination of values of F at the four corners of the rectangle must be non-negative. For a thorough discussion of the properties of multivariate cumulative distributions see some reference which **I must add**.

2.2 Discrete versus Continuous Distributions

Definition: The distribution of a random variable X is called **discrete** (we also say X is 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

$$f_X(x) = P(X = x)\,.$$

Definition: The distribution of a random variable X is called **absolutely continuous** (again we also say X is absolutely continuous) if there is a function f such that

$$P(X \in A) = \int_A f(x)dx \tag{2.1}$$

for any (Borel) set A. This is a p dimensional integral in general. Equivalently

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

Definition: Any f satisfying (2.1) is a **density** of X.

There are a few important warnings and observations here:

- Many statisticians use the word *continuous* instead of the phrase *absolutely continuous* for this property.

- Others use the word *continuous* to mean only that F is a continuous function.

- If X is absolutely continuous then for most (*almost all*) x the function F is differentiable at x and
$$F'(x) = f(x)\,.$$
- Absolute continuity is the property which is needed for a function to be equal to the integral of its derivative. If the function is continuously differentiable, for instance, then it is continuous. If F is continuously differentiable except at a finite number of points where it is continuous then F is absolutely continuous.

Example: The Uniform[0,1] distribution. We say that X is Uniform[0,1] if

$$F(x) = \begin{cases} 0 & x \le 0 \\ x & 0 < x < 1 \\ 1 & x \ge 1\,. \end{cases}$$

which is equivalent to

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

Example: The standard exponential distribution. We say that X is exponential with mean 1 (sometimes written Exp(1)) if

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

or equivalently

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

Remark: I am not going to create notes on all the well known distributions. I expect you will know something about all the famous distributions (including the uniform and exponential distributions I just mentioned).

2.3 Independence, Conditioning and Bayes' Theorem

2.4 Independence, conditional distributions and modelling

When analyzing data statisticians need to specify a statistical model for the data. That is, we regard the data as random variables and specify possible joint distributions for the data. Sometimes the modelling proceeds by modelling the joint density of the data explicitly.

More commonly, however, modelling amounts to a specification in terms of marginal and conditional distributions.

We begin by describing independence. Our description is formal, mathematical and precise. It should be said however that the definitions work two ways. Often we will assume that events or random variables are independent. We will argue that such an assumption is justified by a lack of causal connection between the events – in such a case knowledge of whether or not one event happens should not affect the probability the other happens. This is more subtle than it sounds, though, as we will see when we discuss Bayesian ideas.

Definition: Events A and B are independent if

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

(Notation: we often shorten the notation for intersections by omitting the intersection sign. ThusAB is the event that both A and B happen, which is also written $A \cap B$.)

Definition: A sequence of 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 $1 \le i_1 < \cdots < i_r \le p$.

Example: If we have $p = 3$ independent events then the following equations hold:

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

All these equations are needed for independence! If you have 4 events there are 11 equations; for general p there are $2^p - p - 1$.

Example: Here is a small example to illustrate the fact that all these equations are really needed. In the example there are three events any two of which are independent but where it is not true that all three are independent. Toss a fair coin twice and define the following events.

$$\begin{aligned}
A_1 &= \{\text{first toss is a Head}\} \\
A_2 &= \{\text{second toss is a Head}\} \\
A_3 &= \{\text{first toss and second toss different}\}
\end{aligned}$$

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: We say that two 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: We say that a set of random variables $X_1, \dots, X_p$ are **independent** if, for any $A_1, \dots, A_p$, we have

$$P(X_1 \in A_1, \cdots, X_p \in A_p) = \prod_{i=1}^{p} P(X_i \in A_i).$$

Theorem 1 *1. If $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are independent then for all x, y*

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

2. *If $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ are independent with joint density $f_{X,Y}(x,y)$ then X and Y have densities f_X and f_Y, and (for almost all, in the sense of Lebesgue measure) x and y we have*

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

3. *If X and Y independent with marginal densities f_X and f_Y then (X,Y) has a joint density given by*

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

4. *If $F_{X,Y}(x,y) = F_X(x)F_Y(y)$ for* **all** *x, y then X and Y are independent.*

5. *If (X,Y) has joint density $f(x,y)$ and there exist $g(x)$ and $h(y)$ st $f(x,y) = g(x)h(y)$ for (almost)* **all** *(x,y) then X and Y are independent with densities given by*

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

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

6. *If the pair (X,Y) is discrete with joint probability mass function $f(x,y)$ and there exist functions $g(x)$ and $h(y)$ such that $f(x,y) = g(x)h(y)$ for* **all** *(x,y) then X and Y are independent with probability mass functions given by*

$$f_X(x) = g(x) \Big/ \sum_u g(u)$$

and

$$f_Y(y) = h(y) \Big/ \sum_u h(u)\,.$$

Proof: Some of these assertions are quite technical – primarily those involving densities. My class notes provide only the direct proofs. Here I give more detailed proofs but note that they are based on ideas which are not really part of the course most years.

1. Since X and Y are independent so are the events $X \le x$ and $Y \le y$; hence

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

2. It is notationally simpler to suppose X and Y real valued. General dimensions are not really much harder, however. In assignment 2 I ask you to show that existence of the joint density $f_{X,Y}$ implies the existence of marginal densities f_X and f_Y. Since X, Y have a joint density, we have, 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 $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$

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

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

3. 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 $C = A \times B$ (the Cartesian product of A and B)

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

To prove that g is $f_{X,Y}$ 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. The collection of sets C for which identity holds has closure properties which guarantee that this collection includes the Borel sets. Here are some details.

Definition: A collection $\mathcal{M}$ of subsets of some set E is called a *monotone class* if, whenever $A_1, A_2, \ldots$ all belong to $\mathcal{M}$ and either

$$A_1 \subseteq A_2 \subseteq \cdots$$

or

$$A_1 \supseteq A_2 \supseteq \cdots$$

then, in the first case,

$$\cup_{i=1}^\infty A_i \in \mathcal{M}$$

and, in the second case,

$$\cap_{i=1}^\infty A_i \in \mathcal{M}.$$

Definition: A collection $\mathcal{F}$ of subsets of some set E is called a *field* if:

$$\begin{aligned} \emptyset &\in \mathcal{F} \\ A \in \mathcal{F} &\implies A^c \in \mathcal{F} \\ A_1, \ldots, A_p \in \mathcal{F} &\implies \cup_{i=1}^p A_i \in \mathcal{F}. \end{aligned}$$

This definition is simply the definition of a σ field but with the weaker requirement of closure under finite rather than countable unions.

Lemma 1 *The smallest monotone class containing a field $\mathcal{F}$ is the smallest σ-field containing $\mathcal{F}$.*

Proof: The power set of E (the collection of all subsets of E) is both a σ-field and a monotone class containing $\mathcal{F}$. By "smallest" σ-field containing $\mathcal{F}$ we mean the intersection of all σ-fields containing $\mathcal{F}$; the previous sentence says this is not an empty intersection. The meaning of "smallest" monotone class is analogous. Let $\mathcal{H}$ denote the smallest σ-field and $\mathcal{M}$ the smallest monotone class containing $\mathcal{F}$.

Any σ field containing $\mathcal{F}$ is a monotone class so the smallest monotone class containing $\mathcal{F}$ is a subset of the smallest σ-field containing $\mathcal{F}$. That is, $\mathcal{H} \supseteq \mathcal{M}$. It remains to prove the other direction. Let $\mathcal{G}$ be the collection of all sets $A \in \mathcal{M}$ such that $A^c \in \mathcal{M}$. If $A \in calF$ then $A^c \in \mathcal{F}$ so $\mathcal{G}$ includes $\mathcal{F}$. If $A_1 \subseteq A_2 \subseteq \cdots$ are all sets in $\mathcal{G} \subseteq \mathcal{M}$ then $A \equiv \cup_n A_n \in \mathcal{M}$. On the other hand

$$A_1^c \supseteq A_2^c \supseteq \cdots$$

are all sets in $\mathcal{M}$. Since $\mathcal{M}$ is a monotone class we must have

$$\cap_n A_n^c \in \mathcal{M}$$

but $\cap_n A_n^c = A^c$ so $A^c \in \mathcal{M}$. That is, $\mathcal{G}$ is closed under monotone increasing unions (one of the two properties of a monotone class.

Similarly if

$$A_1 \supseteq A_2 \supseteq \cdots$$

are all sets in $\mathcal{G}$ then $A \equiv \cap_n A_n \in \mathcal{M}$ and

$$A_1^c \subseteq A_2^c \subseteq \cdots$$

are all sets in $\mathcal{M}$. Since $\mathcal{M}$ is a monotone class we must have

$$\cup_n A_n^c \in \mathcal{M}.$$

But $\cup_n A_n^c = A^c$ so $A^c \in \mathcal{M}$. Again we see that $\mathcal{G}$ is closed under monotone decreasing unions. Thus $\mathcal{G}$ is a monotone class containing $\mathcal{F}$. Since it was defined by taking only sets from $\mathcal{M}$ we must have $\mathcal{G} = \mathcal{M}$. That is:

$$A \in \mathcal{M} \implies A^c \in \mathcal{M}.$$

Next I am going to show that $\mathcal{M}$ is closed under countable unions, that is, if $A_1, A_2, \ldots$ are all in $\mathcal{M}$ then so is their union. (Notice that this union might not be a monotone union.) If I can establish this assertion then I will have proved that $\mathcal{M}$ is a σ-field containing $\mathcal{F}$ so $\mathcal{M} \supseteq \mathcal{H}$. This would finish the proof that $\mathcal{M} = \mathcal{H}$.

First fix a $B \in \mathcal{F}$ and let now $\mathcal{G}$ be the collection of all $A \in \mathcal{M}$ such that $A \cup B \in \mathcal{M}$. Just as in the previous part of the argument prove that this new $\mathcal{G}$ is a monotone class containing $\mathcal{F}$. This shows $\mathcal{G} = \mathcal{M}$ and that for every $A \in \mathcal{M}$ and every $B \in \mathcal{F}$ we have $A \cup B \in \mathcal{M}$. Now let $\mathcal{G}$ be the collection of all $B \in \mathcal{M}$ such that for all $A \in \mathcal{M}$ we have $A \cup B \in \mathcal{M}$. Again $\mathcal{G}$ contains $\mathcal{F}$. Check that this third $\mathcal{G}$ is a monotone class and deduce that for every $A \in \mathcal{M}$ and every $B \in \mathcal{M}$ we have $A \cup B \in \mathcal{M}$. In other words: $\mathcal{M}$ is closed under finite unions (by induction on the number of sets in the union).

We have now proved that $\mathcal{M}$ is a field and a monotone class. If $A_1, A_2, \ldots$ are all in $\mathcal{M}$ define $B_n = \cup_{i=1}^n A_i$. Then

(a) $B_1 \subseteq B_2 \subseteq \cdots$.

(b) Each $B_i \in \mathcal{M}$.

(c) $A \equiv \cup_n A_n = \cup_n B_n$

Since $\mathcal{M}$ is a monotone class this last union must be in $\mathcal{M}$. That is $\cup_n A_n \in \mathcal{M}$. This proves $\mathcal{M}$ is a σ-field. •

4. Another monotone class argument.

5.
$$\begin{aligned} P(X \in A, Y \in B) &= \int_A \int_B g(x)h(y)dydx \\ &= \int_A g(x)dx \int_B h(y)dy\,. \end{aligned}$$

Take $B = \mathbb{R}^1$ to see that
$$P(X \in A) = c_1 \int_A g(x)dx$$
where $c_1 = \int h(y)dy$. So $c_1 g$ 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.

6. The discrete case is easier.

Our next theorem asserts something students think is nearly obvious. It is proved by another monotone class argument but the proof is less important than the meaning. The idea is that if U, V, W, X, Y and Z are independent then, for instance U/V, $W+X$ and Ye^Z are independent.

Theorem 2 *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. Similarly $X_1, \ldots, X_{q_1}$, $X_{q_1+1}, \ldots, X_{q_2}$ and so on are independent (provided $q_1 < q_2 < \cdots$).*

Example: Suppose X and Y are independent standard exponential random variables. That is, X and Y have joint density
$$f_{X,Y}(x,y) = e^{-x}1(x>0)e^{-y}1y > 0.$$

Let
$$U = \min\{X,Y\} \text{ and } W = \max\{X,Y\}$$

I will find the joint cdf and joint density of U and W. Begin by considering the event $\{U \le u, W \le w\}$. If $u \le 0$ or $w \le 0$ then the probability is 0 so now assume $u > 0$ and $w > 0$. We then have

$$\begin{aligned} \{U \le u, W \le w\} &= \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w\} \\ &= \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w, X < Y\} \\ &\quad \cup \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w, X > Y\} \\ &\quad \cup \{\min\{X,Y\} \le u, \max\{X,Y\} - \min\{X,Y\} \le w, X = Y\} \end{aligned}$$

The first of these three events is
$$\{X \le u, X < Y \le X + w\}$$
while the second is
$$\{Y \le u, Y < X \le Y + w\}.$$

The third event is a subset of $\{X = Y\}$ which has probability 0. Thus

$$F_{U,W}(u, w) = P(X \leq u, X < Y \leq X + w) + P(Y \leq u, Y < X \leq Y + w).$$

Since X and Y are independent and have the same distribution the two probabilities on the right hand side are equal and we compute only the first. To do so we integrate the joint density of the random variables over the set

$$\{(x, y) : 0 < x \leq u, x < y < x + w\}.$$

The second restriction makes it natural to integrate in the y direction first then in the x direction second. We get

$$P(X \leq u, X < Y \leq X + w) = \int_0^u \int_x^{x+w} e^{-x}e^{-y}\, dy\, dx.$$

The inside integral is just

$$e^{-x}\left(e^{-x} - e^{-(x+w)}\right) = e^{-2x}\left(1 - e^{-w}\right)$$

so

$$P(X \leq u, X < Y \leq X + w) = \left(1 - e^{-w}\right) \int_0^u e^{-2x}\, dx = \left(1 - e^{-w}\right)\left(1 - e^{-2u}\right)/2.$$

Assembling the results we get

$$F_{U,W}(u, w) = \begin{cases} (1 - e^{-w})(1 - e^{-2u}) & u, w > 0 \\ 0 & \text{otherwise.} \end{cases}$$

This function can be rewritten using indicators

$$F_{U,W}(u, w) = \left(1 - e^{-w}\right) 1(w > 0) \left(1 - e^{-2u}\right) 1(u > 0).$$

This evidently factors as the product $F_U(u)F_W(w)$ where

$$\begin{aligned} F_U(u) &= \left(1 - e^{-2u}\right) 1(u > 0) \\ F_W(w) &= \left(1 - e^{-w}\right) 1(w > 0). \end{aligned}$$

Thus we find $U \perp\!\!\!\perp W$ and that U has an exponential distribution with mean 1/2 while W has an exponential distribution with mean 1.

2.5 Conditional probability

The interpretation of probability as long run relative frequency motivates the following definitions of conditional probability. Suppose we have an experiment in which two events A and B are defined and suppose that $P(B) > 0$. Imagine an infinite sequence of independent repetitions of the experiment. Amongst the first n repetitions there must be close to $nP(B)$ occasions where event B occurs in the sense that the ratio number of occurrences divided by n gets close to (B). That is

$$\frac{\#\ B\text{s in first } n \text{ trials}}{n} \to P(B).$$

Also

$$\frac{\#\text{ times both } A \text{ and } B \text{ occur in first } n \text{ trials}}{n} \to P(AB).$$

So if we just pick out of the first n trials those trials where B occur and then see what fraction of these *also* have A occurring we get

$$\frac{\#\text{ times both } A \text{ and } B \text{ occur in first } n \text{ trials}}{\#\ B\text{s in first } n \text{ trials}} \to \frac{P(AB}{P(B)}.$$

This leads to our basic definition.

Definition: We define the conditional probability of an event A given an event B with $P(B) > 0$ by

$$P(A|B) = P(AB)/P(B).$$

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 an absolutely continuous random variable X we have $P(X = x) = 0$ for all x. So what is $P(A|X = x)$ or $f_{Y|X}(y|x)$ since we may not divide by 0? As is usual in mathematics we define the ratio 0/0 by taking a suitable limit:

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

If, e.g., X, Y have joint density $f_{X,Y}$ then with $A = \{Y \le y\}$ we have

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

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

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

We now define the conditional cumulative distribution function of Y given $X = x$ by

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

If we differentiate this formula by y we get the undergraduate definition of the conditional density of Y given $X = x$, namely,

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

in words we find "conditional = joint/marginal".

Example: The 3 cards problem revisited. This is the problem where we have 3 cards – red on both sides, green on both sides and red on one / green on the other. We draw a card and see the colour on the side which is face up. Suppose we see Red. What is the chance the side face down is Red?

Students sometimes think the answer is 1/2. They say: either I am looking at the all red card or the red/green card. These are equally likely so this conditional probability is 1/2. This is wrong – the two cards are not equally likely given that the side facing up is Red.

To see this clearly we should go back to the basics. Let A be the event that we see a red side. In terms of the elementary outcomes in the example at the start of Chapter 2 we have

$$A = \{\omega_1, \omega_2, \omega_3\}.$$

Let B be the event that the side face down is red. Then

$$B = \{\omega_1, \omega_2, \omega_4\}.$$

We then have

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{2/6}{3/6} = \frac{2}{3}.$$

It is also possible to do this more intuitively but to do so you have to be careful. You are conditioning on the event that you are looking at 1 of the 3 red sides – all equally likely. Of these three sides two have the property that the other side is red. That makes the conditional probability 2/3.

2.5.1 Bayes Theorem

The definition of conditional probability shows that if $P(A) > 0$ and $P(B) > 0$ then we have

$$P(AB) = P(A|B)P(B) = P(B|A)P(A).$$

The crucial point about this observation is that one formula conditions on B and the other on A. Bayes theorem just rewrites this formula to emphasize the change in order of conditioning:

Theorem 3 *If A and B are two events with $P(A) > 0$ and $P(B) > 0$ then*

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

It seems to me to be useful to relate this to some reasoning ideas. If a certain statement P implies a statement Q then Q is always true whenever P is true. Of course if Q is not true then neither is P. That is, the statement "not Q" implies the statement "not P". In terms of probabilities the analogy is that if $P(B|A) = 1$ then $P(A^c|B^c) = 1$ (assuming that $P(B^c) \neq 0$). This follows from

$$\begin{aligned} P(A^c|B^c) &= \frac{P(A^cB^c)}{P(B^c)} \\ &= \frac{1 - P(A \cup B)}{P(B^c)} \\ &= \frac{1 - P(A) - P(B) + P(B|A)P(A)}{1 - P(B)} \\ &= \frac{1 - P(A) - P(B) + P(A)}{1 - P(B)} \\ &= \frac{1 - P(B)}{1 - P(B)} = 1. \end{aligned}$$

It is NOT a theorem of logic that if P implies Q then Q implies P. But there is a sense in which if P usually happens and usually when P happens so does Q then Q usually happens and when Q happens usually P does too. Let's look at the formula with statements P and Q replaced by events A and B. Imagine that P is "A happens" and Q is "B happens".

Then

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

so if both terms on the left are nearly 1 ("usually happens") then both terms on the right must be nearly 1 (because if either were small the product would be too small to equal the thing on the left which is nearly 1).

The idea underlying Bayes' Theorem can be translated into the language of conditional densities:

$$f_{X|Y} = \frac{f_{Y|X} f_X}{f_Y}$$

Nowadays Bayesians like to write

$$(x|y) = (y|x)(x)/(y)$$

with the parentheses indicating densities and the letters indicating variables. This notation uses the letter in the argument of a function to indicate which function is being discussed and is at least a bit dangerous since

$$(1|2) = (2|1)(1)/(2)$$

doesn't really tell you which variables are under discussion even though it a special case of the formula above with $x = 1$ and $y = 2$.

More general formulas arise like

$$P(ABCD) = P(A|BCD)P(B|CD)P(C|D)P(D)$$

This formula can be rewritten in many orders to get a variety of equivalent expressions which, divided by some of the terms involved give theorems like that of Bayes. Also, if $A_1, \ldots, A_k$ are *mutually exclusive and exhaustive* then

$$P(A_1|B) = \frac{P(B|A_1)P(A_1)}{\sum_i P(B|A_j)P(A_j)}$$

Bayes theorem is often written in this form. Of course the denominator is just $P(B)$. I remark that *mutually exclusive* means pairwise disjoint and *exhaustive* means

$$\cup_1^k A_i = \Omega.$$

The density formula is really analogous to this more general looking version of Bayes' theorem since integrals are limits of sums and

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{f_{Y|X}(y|x)f_X(x)}{\int_u f_{XY}(u,y)du}.$$

Chapter 3

Expectation and Moments

I begin by reviewing the usual undergraduate definitions of expected value. For absolutely continuous random variables X we usually say:

Definition: If X has density f then

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

For discrete random variables we say:

Definition: If X has discrete density f then

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

There is something of a problem with these two definitions. They seem to define, for instance, $\mathrm{E}(X^2)$, in two different ways. If X has density f_X then we would have

$$\mathrm{E}(X^2) = \int x^2 f_X(x)\,dx.$$

But we could also define $Y = X^2$ and try to figure out a density f_Y for Y. Then we would have

$$\mathrm{E}(Y) = \int y f_Y(y) dy.$$

Are these two formulas the same? The answer is yes.

Fact: If $Y = g(X)$ for some one-to-one smooth function g (by which I mean say g is continuously differentiable) then

$$\begin{aligned}\mathrm{E}(Y) &= \int y f_Y(y)\,dy = \int g(x) f_Y(g(x))g'(x)\,dx \\ &= \mathrm{E}\{g(X)\}\end{aligned}$$

by change of variables formula for integration so we must have

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

For the moment I won't prove this but let me consider the case where, for instance $Y = e^{2X}$. Then $g(x) = e^{2x}$ and $g'(x) = 2e^{2x}$. Moreover

$$\begin{aligned} f_X(x) &= \frac{d}{dx}F_X(x) \\ &= \frac{d}{dx}P(X \le x) \\ &= \frac{d}{dx}P(e^{2X} \le e^{2x}) \\ &= \frac{d}{dx}P(Y \le e^{2x}) \\ &= \frac{d}{dx}F_Y(e^{2x}) \\ &= f_Y(e^{2x})\frac{d}{dx}e^{2x} \end{aligned}$$

as advertised.

3.0.2 General Definition of E

There are random variables which are neither absolutely continuous nor discrete. I now give a definition of expected value which covers such cases and includes both discrete and continuous random variables.

Definition: We say that 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.

Definition: For a simple random variable X we define

$$\mathrm{E}(X) = \sum a_i P(A_i)\,.$$

I remark that logically it might be possible to write X in two ways, say

$$\sum_{i=1}^n a_i 1(\omega \in A_i) = \sum_{i=1}^m b_i 1(\omega \in B_i)$$

some constants $a_1, \ldots, a_n$, $b_1, \ldots, b_m$ and events $A_1, \ldots, A_n$ and $B_1, \ldots, B_m$. I claim that if this happens then we must have

$$\sum_{i=1}^n a_i P(A_i) = \sum_{i=1}^m b_i P(B_i).$$

I won't prove the claim!

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

Definition: If $X \geq 0$ then

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

This notation hides the fact that for positive, simple, random variables X we appear to have given 2 definitions for $\mathrm{E}(X)$. It is possible to prove they are the same.

Finally we extend the definition to general random variables:

Definition: A random variable X is **integrable** if

$$\mathrm{E}(|X|) < \infty\,.$$

In this case we define

$$\mathrm{E}(X) = \mathrm{E}\{\max(X, 0)\} - \mathrm{E}\{\max(-X, 0)\}\,.$$

Again it might seem we have another definition for simple random variable or for non-negative random variables but it is possible to prove all the definitions agree.

Fact: : E is a linear, monotone, positive operator. This means:

1. **Linear**: $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$ provided X and Y are integrable.
2. **Positive**: $P(X \geq 0) = 1$ implies $\mathrm{E}(X) \geq 0$.
3. **Monotone**: $P(X \geq Y) = 1$ and X, Y integrable implies $\mathrm{E}(X) \geq \mathrm{E}(Y)$.

Jargon: An *operator* is a function whose domain is itself a set of functions. That makes E an operator because random variables are functions. Sometimes we call operators whose range is in real or complex numbers a *functional*.

3.0.3 Convergence Theorems

There are some important theorems about interchanging limits with integrals and our definition of E is really the definition of an integral. In fact you will often see a variety of notations:

$$\begin{aligned}\mathrm{E}(g(X)) &= \int g(x)F(dx)\\ &= \int g(x)dF(x)\\ &= \int g dF\end{aligned}$$

Sometimes the integral notations make it easier to see how a calculation works out. The notation $dF(x)$ has the advantage that if F has a density $f = F'$ we can write

$$dF(x) = f(x)dx.$$

In calculus courses there is quite a bit of attention paid to such questions as when

$$\frac{d}{dy}\int g(x,y)dx = \int \frac{\partial}{\partial y} g(x,y)dx.$$

The issue is that the definition of a derivative involves a limit. The left hand side is

$$\lim_{h\to 0}\int \frac{g(x,y+h)-g(x,y)}{h}dx$$

while the right hand side is

$$\int \lim_{h\to 0}\frac{g(x,y+h)-g(x,y)}{h}dx$$

and the issue is whether or not you can pull limits in and out of integrals. You often can; the next two theorems give precise conditions for this to work.

Theorem 4 (Monotone Convergence) *If* $0 \le X_1 \le X_2 \le \cdots$ *and* $X = \lim X_n$ *(the limit* X *automatically exists) then*

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

Remark: In the hypotheses we need $P(X_{n+1} \ge X_n) = 1$ and $P(X_1 \ge 0) = 1$.

Theorem 5 (Dominated Convergence) *If* $|X_n| \le Y_n$ *and* $\exists$ *a random variable* X *such that* $X_n \to X$ *(technical details of this convergence come later in the course) and a random variable* Y *such that* $Y_n \to Y$ *with* $\lim_{n\to\infty} \mathrm{E}(Y_n) = \mathrm{E}(Y) < \infty$ *then*

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

Remark: The dominated convergence theorem is often used with all Y_n the same random variable Y. In this case the hypothesis that $\lim_{n\to\infty} \mathrm{E}(Y_n) = \mathrm{E}(Y) < \infty$ is just the hypothesis that $\mathrm{E}(Y) < \infty$.

Remark: These theorems are used in *approximation.* We compute the limit of the expected values of a sequence of random variables X_n and then approximate $\mathrm{E}(X_{225})$ (or whatever n we actually have instead of 225) by $\mathrm{E}(X)$.

3.0.4 Connection to ordinary integrals

Theorem 6 *With this definition of E:*

1. *if X has density $f(x)$ (even in R^p say) and $Y = g(X)$ then*

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

(This could be a multiple integral.)

2. *If X has probability mass function f then*

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

3. *The first conclusion works, e.g., even if X has a density but Y doesn't.*

3.0.5 Moments

- **Definition**: The r^{th} moment (about the origin) of a real random variable X is $\mu_r' = \mathrm{E}(X^r)$ (provided it exists).

- We generally use μ for $\mathrm{E}(X)$.

- **Definition**: The r^{th} central moment is

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

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

- **Definition**: For an R^p valued random vector X

$$\mu_X = \mathrm{E}(X)$$

is the vector whose i^{th} entry is $\mathrm{E}(X_i)$ (provided all entries exist).

- **Definition**: The $(p \times p)$ variance covariance matrix of X is

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

which exists provided each component X_i has a finite second moment.

3.0.6 Moments and independence

Theorem 7 *If $X_1,\ldots,X_p$ are independent and each X_i is integrable then $X = X_1 \cdots X_p$ is integrable and*

$$\mathrm{E}(X_1 \cdots X_p) = \mathrm{E}(X_1) \cdots \mathrm{E}(X_p)\,.$$

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

$$\begin{aligned}
\mathrm{E}(X_1 \cdots X_p) &= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} \mathrm{E}(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p})) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\
&= \sum_{j_1 \ldots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\
&= \sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1}) \cdots \sum_{j_p} x_{pj_p} P(X_p = x_{pj_p}) \\
&= \prod \mathrm{E}(X_i)\,.
\end{aligned}$$

Non-negative Case: Now consider non-negative random variables X_i, Let X_{in} be X_i rounded down to the nearest multiple of 2^{-n} to a maximum of n. That is: if

$$\frac{k}{2^n} \le X_i < \frac{k+1}{2^n}$$

then $X_{in} = k/2^n$ for $k = 0, \ldots, n2^n$. For $X_i > n$ put$X_{in} = n$. Now apply the case we have just done:

$$\mathrm{E}(\prod X_{in}) = \prod \mathrm{E}(X_{in})\,.$$

Monotone convergence applies to both sides to prove the result for non-negative X_i.

General case: now consider general X_i and write each X_i as the difference of positive and negative parts:

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

Write out $\prod_i |X_i|$ as a sum of products and apply the positive case to see that if all the X_i are integrable then so is $\prod_i X_i$.

3.0.7 Conditional Expectations

- Abstract definition of conditional expectation is:

- **Definition**: $\mathrm{E}(Y|X)$ is any function of X such that
$$\mathrm{E}\left[R(X)\mathrm{E}(Y|X)\right] = \mathrm{E}\left[R(X)Y\right]$$
for any bounded function $R(X)$.

- **Definition**: $\mathrm{E}(Y|X=x)$ is a function $g(x)$ such that
$$g(X) = \mathrm{E}(Y|X)$$

- **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:

$$\begin{aligned}
\mathrm{E}(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\
&= \int R(x) \int y f(y|x) dy f_X(x) dx \\
&= \int \int R(x) y f_X(x) f(y|x) dy dx \\
&= \int \int R(x) y f_{X,Y}(x,y) dy dx \\
&= \mathrm{E}(R(X)Y)
\end{aligned}$$

Interpretation of conditional expectation

- **Intuition**: Think of $\mathrm{E}(Y|X)$ as average Y holding X fixed.

- Behaves like ordinary expected value but functions of X only are like constants:
$$\mathrm{E}(\sum A_i(X)Y_i|X) = \sum A_i(X)\mathrm{E}(Y_i|X)$$

- Statement called Adam's law by Jerzy Neyman – he used to say it comes before all the others:
$$\mathrm{E}[\mathrm{E}(Y|X)] = \mathrm{E}(Y)$$
which is just the definition of $\mathrm{E}(Y|X)$ with $R(X) \equiv 1$.

- 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:
$$\mathrm{Var(Y)} = \mathrm{Var}(\mathrm{E}(Y|X)) + \mathrm{E}[\mathrm{Var}(Y|X)]$$

- The conditional variance means
$$\mathrm{Var}(Y|X) = \mathrm{E}[(Y - \mathrm{E}(Y|X))^2|X].$$

3.0.8 Moments

Moment is an old word from physics used in such terms as moments of inertia. There is actually a good analogy between the physics use of the term and our use. If you made a block of wood shaped like the density of a random variable X and you tried to balance the block (it will be thin, long, flat on the bottom and curved into the shape of the density on the top) on a pencil the pencil would have to be located under the mean of the density. The *moment of force* about this pencil would be 0. Warning: go elsewhere to learn physics.

Definition: The r^{th} moment (about the origin) of a real random variable X is $\mu_r' = \mathrm{E}(X^r)$ (provided it exists – that is, provided X^r is integrable).

Notation: We generally use μ for $\mathrm{E}(X)$.

Definition: The r^{th} central moment is

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

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

Definition: For an R^p valued random vector X

$$\mu_X = \mathrm{E}(X)$$

is the vector whose i^{th} entry is $\mathrm{E}(X_i)$ (provided all entries exist). Similarly for matrices we take expected values entry by entry.

Definition: The $(p \times p)$ variance covariance matrix of X is

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

which exists provided each component X_i has a finite second moment.

The ijth entry in $(X-\mu)(X-\mu)^t$ is $(X_i-\mu_i)(X_j-\mu_j)$. As a result this matrix has diagonal entries which are the usual variances of the individual X_i and off diagonal entries which are covariances.

3.0.9 Moments and independence

Theorem 8 *If $X_1,\ldots,X_p$ are independent and each X_i is integrable then $X = X_1\cdots X_p$ is integrable and*

$$\mathrm{E}(X_1\cdots X_p) = \mathrm{E}(X_1)\cdots\mathrm{E}(X_p)\,.$$

Proof: First 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

$$\begin{aligned}
E(X_1 \cdots X_p) &= \sum_{j_1 \dots j_p} x_{1j_1} \cdots x_{pj_p} E(1(X_1 = x_{1j_1}) \cdots 1(X_p = x_{pj_p})) \\
&= \sum_{j_1 \dots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1} \cdots X_p = x_{pj_p}) \\
&= \sum_{j_1 \dots j_p} x_{1j_1} \cdots x_{pj_p} P(X_1 = x_{1j_1}) \cdots P(X_p = x_{pj_p}) \\
&= \sum_{j_1} x_{1j_1} P(X_1 = x_{1j_1}) \cdots \sum_{j_p} x_{pj_p} P(X_p = x_{pj_p}) \\
&= \prod E(X_i)\,.
\end{aligned}$$

Now we consider the case of general $X_i \geq 0$. Let X_{in} be X_i rounded down to nearest multiple of 2^{-n} (to maximum of n). That is, if

$$\frac{k}{2^n} \leq X_i < \frac{k+1}{2^n}$$

then we define $X_{in} = k/2^n$ for $k = 0, \dots, n2^n$ and for $X_i > n$ we put $X_{in} = n$.

Now we apply the case we have just done:

$$E(\prod X_{in}) = \prod E(X_{in})\,.$$

Finally we apply the monotone convergence theorem to both sides.

It remains to consider X_i which might not be positive. Use the previous case to prove that

$$|\prod X_i| = \prod |X_i|$$

is integrable. Then expend the product of positive minus negative parts,

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

Next check that all of the 2^p terms you get, after expanding out, are integrable and apply the previous case. The details are algebraically messy and not very informative in my view. An alternative theory is that I am too lazy to write them out.

3.1 Conditional Expectations

I am going to give here the abstract "definition" of a conditional expectation. The definition is indirect – it is a thing which has a certain property. That means that I ought to prove there is a thing with that property and that the thing with the property is unique. As usual – I won't be doing that here.

The abstract definition of conditional expectation is:

Definition: $\mathrm{E}(Y|X)$ is any function of X such that

$$\mathrm{E}\left[R(X)\mathrm{E}(Y|X)\right] = \mathrm{E}\left[R(X)Y\right]$$

for any bounded function $R(X)$.

Definition: $\mathrm{E}(Y|X = x)$ is a function $g(x)$ such that

$$g(X) = E(Y|X)$$

that is, such that $g(X)$ satisfies the previous definition.

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:

$$\begin{aligned}
E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\
&= \int R(x) \int y f(y|x) dy f_X(x)dx \\
&= \int \int R(x) y f_X(x) f(y|x) dy dx \\
&= \int \int R(x) y f_{X,Y}(x, y) dy dx \\
&= E(R(X)Y)
\end{aligned}$$

3.1.1 Interpretation and properties of conditional expectation

- **Intuition**: Think of $E(Y|X)$ as average Y holding X fixed.
- Behaves like ordinary expected value but functions of X only are like constants:

$$E(\sum A_i(X)Y_i|X) = \sum A_i(X)E(Y_i|X)$$

- Statement 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 $R(X) \equiv 1$.

- 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

$$\mathrm{Var}(Y|X) = E[(Y - E(Y|X))^2|X].$$

3.2 Generating Functions

3.2.1 Moment Generating Functions

There are many uses of generating functions in mathematics. We often study the properties of a sequence a_n of numbers by creating the function

$$\sum_{n=0}^{\infty} a_n s^n$$

In statistics the most commonly used generating functions are the probability generating function (for discrete variables), the moment generating function, the characteristic function and the cumulant generating function. I begin with moment generating functions:

Definition: The moment generating function of a real valued random variable X is

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

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

Definition: The moment generating function of a random vector $X \in R^p$ is

$$M_X(u) = \mathrm{E}[e^{u^t X}]$$

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

This function has a formal connection to moments obtained by taking expected values term by term; in fact if $M_X(t)$ is finite for all $|t| < \epsilon$ then it is legitimate to take expected values term by term for $|t| < \epsilon$. We get

$$\begin{aligned} M_X(t) &= \sum_{k=0}^{\infty} \mathrm{E}[(tX)^k]/k! \\ &= \sum_{k=0}^{\infty} \mu_k' t^k/k!\,. \end{aligned}$$

Sometimes we can find the power series expansion of M_X and read off the moments of X from the coefficients of $t^k/k!$.

Theorem 9 *If M is finite for all $t \in [-\epsilon, \epsilon]$ for some $\epsilon > 0$ then*

1. *Every moment of X is finite.*

2. *M is C^∞ (in fact M is analytic).*

3. *$\mu_k' = \frac{d^k}{dt^k} M_X(0)$.*

Note: A function is C^∞ if it has continuous derivatives of all orders.

Note: Analytic means the function has a convergent power series expansion in neighbourhood of each $t \in (-\epsilon, \epsilon)$.

The proof, and many other facts about moment generating functions, rely on advanced techniques in the field of complex variables. I won't be proving any of these assertions.

3.2.2 Moment Generating Functions and Sums

One of the most useful facts about moment generating functions is that the moment generating function of a sum of independent variables is the product of the individual moment generating functions.

Theorem 10 *If $X_1, \ldots, X_p$ are independent random vectors in $\mathbb{R}^p$ and $Y = \sum X_i$ then the moment generating function of Y is the product of those of the individual X_i:*

$$M_Y(u) = \mathrm{E}(e^{u^tY}) = \prod_i \mathrm{E}(e^{u^tX_i}) = \prod_i M_{X_i}(u).$$

If we could find the power series expansion of M_Y then we could find the moments of M_Y. The problem, however, is that the power series expansion of M_Y not nice function of the expansions of individual M_{X_i}. There is a related fact, namely, that the first 3 moments (meaning μ, σ^2 and μ_3) of Y are sums of those of the X_i:

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

(I have given the univariate versions of these formulas but the multivariate versions are correct as well. The first line is a vector, the second a matrix and the third an object with 3 subscripts.) However:

$$\begin{aligned}\mathrm{E}[(Y-\mathrm{E}(Y))^4] &= \sum\{\mathrm{E}[(X_i-\mathrm{E}(X_i))^4] - 3\mathrm{E}^2[(X_i-\mathrm{E}(X_i))^2]\}\\ &\quad + 3\left\{\sum \mathrm{E}[(X_i-\mathrm{E}(X_i))^2]\right\}^2\end{aligned}$$

These observations lead us to consider cumulants and the cumulant generating function. Since the logarithm of a product is a sum of logarithms we are led to consider taking logs of the moment generating function. The result will give us *cumulants* which add up properly.

Definition: the cumulant generating function of a a random vector X by

$$K_X(u) = \log(M_X(u))\,.$$

Then if $X_1, \ldots, X_n$ are independent and $Y = \sum_1^n X_i$ we have

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

Note that moment generating functions are all positive so that the cumulant generating functions are defined wherever the moment generating functions are.

Now K_Y has a power series expansion. I consider here only the univariate case.

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

Definition: the κ_r are the cumulants of Y.

Observe that

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

In other words cumulants of independent quantities add up. Now we examine the relation between cumulants and moments by relating the power series expansion of M with that of its logarithm. The cumulant generating function is

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

Call the quantity in $[\ldots]$ x and expand

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

Stick in the power series

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

Expand out 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

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

Reference: Kendall and Stuart (or a new version called *Kendall's Theory of Advanced Statistics* by Stuart and Ord) for formulas for larger orders r.

Example: The normal distribution: Suppose $X_1, \ldots, X_p$ independent, $X_i \sim N(\mu_i, \sigma_i^2)$ so that

$$\begin{aligned} M_{X_i}(t) &= \int_{-\infty}^{\infty} e^{tx} e^{-\frac{1}{2}(x-\mu_i)^2/\sigma_i^2} dx/(\sqrt{2\pi}\sigma_i) \\ &= \int_{-\infty}^{\infty} e^{t(\sigma_i z+\mu_i)} e^{-z^2/2} dz/\sqrt{2\pi} \\ &= e^{t\mu_i} \int_{-\infty}^{\infty} e^{-(z-t\sigma_i)^2/2+t^2\sigma_i^2/2} dz/\sqrt{2\pi} \\ &= e^{\sigma_i^2 t^2/2 + t\mu_i}\,. \end{aligned}$$

The cumulant generating function is then

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

The cumulants are $\kappa_1 = \mu_i$, $\kappa_2 = \sigma_i^2$ and every other cumulant is 0. 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: The χ^2 distribution: In you homework I am asking you to derive the moment and cumulant generating functions and moments of a Gamma random variable. Now suppose $Z_1, \ldots, Z_\nu$ independent $N(0,1)$ rvs. By definition the random variable $S_\nu = \sum_1^\nu Z_i^2$ has χ_ν^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 moment generating function of S_1 is

$$(1-2t)^{-1/2}\,.$$

It follows that

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

which is (from the homework) the moment generating function of a Gamma($\nu/2, 2$) random variable. So the χ_ν^2 distribution has a Gamma($\nu/2, 2$) density given by

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

Example: The Cauchy distribution: The Cauchy density is

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

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. *Every* t distribution has exactly same moment generating function. So we cannot use moment generating functions to distinguish such distributions. The problem is that these distributions do not have infinitely many finite moments. So we now develop a substitute substitute for the moment generating function which is defined for every distribution, namely, the characteristic function.

3.2.3 Aside on complex arithmetic

Complex numbers are a fantastically clever idea. The idea is to imagine that -1 has a square root and see what happens. We add $i \equiv \sqrt{-1}$ to the real numbers. Then, we insist that all the usual rules of algebra are unchanged. So, if i and any real numbers a and b are to be complex numbers then so must be $a + bi$. Now let us look at each of the arithmetic operations to see how they have to work:

- Multiplication: 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: we follow the usual rules (commutative, associative and distributive laws) to get

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

- Additive inverses:

  $$-(a+bi) = -a + (-b)i.$$

  Notice that $0 + 0i$ functions as 0 – it is an additive identity. In fact we normally just write 0.

- Multiplicative inverses:

  $$\begin{aligned}\frac{1}{a+bi} &= \frac{1}{a+bi}\frac{a-bi}{a-bi}\\ &= \frac{a-bi}{a^2 - abi + abi - b^2i^2} = \frac{a-bi}{a^2+b^2}\,.\end{aligned}$$

- Division:

  $$\frac{a+bi}{c+di} = \frac{(a+bi)}{(c+di)}\frac{(c-di)}{(c-di)} = \frac{ac - bd + (bc+ad)i}{c^2+d^2}\,.$$

  This rule for clearing the complex number from the denominator is a perfect match for the technique taught in high school and used in calculus, for dealing with fractions involving $a + b\sqrt{c}$ in the denominator.

- You should now notice that the usual rules of arithmetic don't require any more numbers than

  $$x + yi$$

  where x and y are real. So the complex numbers $\mathbb{C}$ are just all these numbers.

- **Transcendental functions**: For real x have $e^x = \sum x^k/k!$ and $e^{a+b} = e^a e^b$ so we want to insist that
$$e^{x+iy} = e^x e^{iy}\,.$$
The problem is how to compute e^{iy}?

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

- We can thus write
$$e^{x+iy} = e^x(\cos(y) + i\sin(y))$$

- Identify $x + yi$ with the corresponding point (x, y) in the plane.

- Picture the complex numbers as forming a plane.

- 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
$$x + iy = \sqrt{x^2 + y^2}\, e^{i\theta} = re^{i\theta}$$
for an angle $\theta \in [0, 2\pi)$.

- Multiplication revisited: if $x + iy = re^{i\theta}$ and $x' + iy' = r'e^{i\theta'}$ then when we multiply we get
$$(x + iy)(x' + iy') = re^{i\theta}r'e^{i\theta'} = rr'e^{i(\theta+\theta')}\,.$$

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

- **Definition**: The **modulus** of $x + iy$ is
$$|x + iy| = \sqrt{x^2 + y^2}\,.$$

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

- Some identities: $z = x + iy = re^{i\theta}$ and $z' = x' + iy' = r'e^{i\theta'}$.

- Then
$$\begin{aligned} z\overline{z} &= x^2 + y^2 = r^2 = |z|^2 \\ \frac{z'}{z} &= \frac{z'\overline{z}}{|z|^2} = rr'e^{i(\theta'-\theta)} \\ \overline{re^{i\theta}} &= re^{-i\theta}. \end{aligned}$$

3.2.4 Notes on calculus with complex variables

The rules for calculus with complex numbers are really very much like the usual rules. 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.

Fact: (This fact is not used explicitly in course). If $f : \mathbb{C} \mapsto \mathbb{C}$ is differentiable then f is analytic (has power series expansion).

3.2.5 Characteristic Functions

Definition: The characteristic function of a real random variable X is

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

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

Since

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

we find that

$$\phi_X(t) = \mathrm{E}(\cos(tX)) + i\mathrm{E}(\sin(tX)) \,.$$

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

The characteristic function is called "characteristic" because if you know it you know the distribution of the random variable involved. That is what is meant in mathematics when we say something characterizes something else.

Theorem 11 *For any two real random vectors X and Y (say p-dimensional) the following are equivalent:*

1. *X and Y have the same distribution, that is, for any (Borel) set $A \subset \mathbb{R}^p$ we have*

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

2. *$F_X(t) = F_Y(t)$ for all $t \in \mathbb{R}^p$.*

3. *$\phi_X(u) = \mathrm{E}(e^{iu^tX}) = \mathrm{E}(e^{iu^tY}) = \phi_Y(u)$ for all $u \in \mathbb{R}^p$.*

Moreover, all these are implied if there is $\epsilon > 0$ such that for all $|t| \le \epsilon$

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

3.3 Inversion Formulae

3.3.1 Inversion

The previous theorem is non-constructive characterization. That is, it says that ϕ_X determines F_X and f_X but it does not say how to find the latter from the former. This raises the question: Can get from ϕ_X to F_X or f_X by **inversion**.

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

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

You multiply this by e^{-ijt} and integrate from 0 to 2π. This produces

$$\int_0^{2\pi} e^{-ijt}\phi_X(t)\,dt = \sum_k P(X=k)\int_0^{2\pi} e^{i(k-j)t}\,dt.$$

Now for $k \neq j$ the derivative of

$$e^{i(k-j)t}$$

with respect to t is just

$$i(k-j)e^{i(k-j)t}$$

so the integral is simply

$$\left.\frac{e^{i(k-j)t}}{i(k-j)}\right|_{t=0}^{t=2\pi} = \frac{\cos(2(k-j)\pi)+i\sin(2(k-j)\pi)-\cos(0)-i\sin(0)}{i(k-j)} = \frac{1+0i-1-0i}{i(k-j)} = 0.$$

The integral with $k=j$, however, is different. It is just

$$\int_0^{2\pi} e^{i0t}dt = \int_0^{2\pi} 1dt = 2\pi.$$

So

$$\int_0^{2\pi} e^{-ijt}\phi_X(t)\,dt = 2\pi P(X=j).$$

Now suppose X has continuous bounded density f. Define

$$X_n = [nX]/n$$

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

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

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

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

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

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

(Dominated convergence: $|e^{iu}| \leq 1$.)

Range of integration converges to the whole real line.

If $k/n \to x$ left hand side converges to density $f(x)$ while 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; the book by Loève on probability is a good source for such formulas and their proofs.

All such formulas are called **Fourier inversion formulas**. The characteristic function is also called the **Fourier transform** of f or F.

3.3.2 Inversion of the Moment Generating Function and Saddlepoint Approximations

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

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

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. So: there is an inversion formula for M_X using a complex *contour integral*:

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.

The inversion formula just derived was

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

Now imagine making a change of variables to $z = it$. As t, a real variable, goes from $-\infty$ to ∞ the variable z runs up the imaginary axis. We also have $dz = i\,dt$. This leads to the following inversion formula for the moment generating function

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

(the limits of integration indicate a contour integral running up the imaginary axis.)

It is now possible to replace contour (using complex variables theory) with the line $Re(z) = c$. ($Re(Z)$ denotes the real part of z, that is, x when $z = x + iy$ with x and y real.) We must choose c so that $M(c) < \infty$. In this case we rewrite the inversion formula using the cumulant generating function $K(t) = \log(M(t))$ in the following form:

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

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 we 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 called 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 a normal density calculation; it gives

$$\sqrt{2\pi/K''(c)}\,.$$

Thus our saddlepoint approximation is

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

The tactic used here is essentially the same idea as in Laplace's approximation whose most famous example is Stirling's formula

Example: Stirling's approximation to a factorial. We may show, by induction on n and integration by parts that

$$n! = \int_0^\infty \exp(n\log(x) - x)dx\,.$$

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 $y = (x-n)/\sqrt{n}$; get approximation

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

or

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

Note: I am being quite sloppy about limits of integration; this is a fixable error but I won't be doing the fixing. A real proof must show that the integral over x not near n is negligible.

Chapter 4

Distribution Theory

The basic problem of distribution is to compute the distribution of statistics when the data come from some model. You start with assumptions about the density f or the cumulative distribution function F of some random vector $X = (X_1, \ldots, X_p)$; typically X is your data and f or F come from your model. If you don't know f you need to try to do this calculation for all the densities which are possible according to your model. So now suppose $Y = g(X_1, \ldots, X_p)$ is some function of X — usually some statistic of interest.

How can we compute the distribution or CDF or density of Y?

4.1 Univariate Techniques

Method 1: our first method is to compute the cumulative distribution function of Y by integration and differentiate to find the density f_Y.

Example: Suppose $U \sim \text{Uniform}[0,1]$ and $Y = -\log U$.

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

so that Y has a standard exponential distribution.

Example: The χ^2 density. Suppose $Z \sim N(0,1)$, that is, that Z has density

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

and let $Y = Z^2$. Then

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

Now differentiate

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

to get

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

Now we differentiate:

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

There is a similar formula for the other derivative. Thus

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

We will find **indicator** notation useful:

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

which we use to write

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

This changes our definition unimportantly at $y = 0$.
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 them anyway. Remember the fundamental theorem of calculus:

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

at any x where f is continuous.

This leads to the following summary: for $Y = g(X)$ with X and Y each real valued

$$\begin{aligned} P(Y \le y) &= P(g(X) \le y) \\ &= P(X \in g^{-1}(-\infty, y])\,. \end{aligned}$$

Take d/dy to compute the density

$$f_Y(y) = \frac{d}{dy}\int_{\{x: g(x) \le y\}} f_X(x)\,dx\,.$$

Often we can differentiate without doing the integral.
Method 2: One general case is handled by the method of change of variables. Suppose that g is one to one. I will do the case where g is increasing and differentiable.

We begin from the interpretation of density (based on the notion that the density is give by F'):

$$\begin{aligned} f_Y(y) &= \lim_{\delta y \to 0} \frac{P(y \le Y \le y + \delta y)}{\delta y} \\ &= \lim_{\delta y \to 0} \frac{F_Y(y+\delta y) - F_Y(y)}{\delta y} \end{aligned}$$

and

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

Now assume $y = g(x)$. Define δy by $y + \delta y = g(x + \delta x)$. Then

$$P(y \le Y \le g(x + \delta x)) = P(x \le X \le x + \delta x) \, .$$

We get

$$\frac{P(y \le Y \le y + \delta y))}{\delta y} = \frac{P(x \le X \le x + \delta x)/\delta x}{\{g(x+\delta x) - y\}/\delta x} \, .$$

Take the limit as $\delta x \to 0$ to get

$$f_Y(y) = f_X(x)/g'(x) \text{ or } f_Y(g(x))g'(x) = f_X(x) \, .$$

Alternative view: we can now try to look at this calculation in a slightly different way: each probability above is the integral of a density. The first is the integral of f_Y from $y = g(x)$ to $y = g(x + \delta x)$. The interval is narrow so f_Y is nearly constant over this interval and

$$P(y \le Y \le g(x + \delta x)) \approx f_Y(y)(g(x + \delta x) - g(x)) \, .$$

Since g has a derivative $g(x + \delta x) - g(x) \approx \delta x g'(x)$ so we get

$$P(y \le Y \le g(x + \delta x)) \approx f_Y(y) g'(x) \delta x \, .$$

The same idea applied to $P(x \le X \le x + \delta x)$ gives

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

so that

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

or, cancelling the δx in the limit

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

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

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

It is often more useful to express the whole formula in terms of the new variable y to get a formula for $f_Y(y)$. We do this by solving $y = g(x)$ to get x in terms of y, that is, find a formula for $x = g^{-1}(y)$ and then see that

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

This is just the change of variables formula for doing integrals.

Remark: : For g decreasing $g' < 0$ but then the interval $(g(x), g(x + \delta x))$ is really $(g(x + \delta x), g(x))$ so that $g(x) - g(x + \delta x) \approx -g'(x)\delta x$. In both cases this amounts to the formula

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

This leads to what I think is a very useful **Mnemonic**:

$$f_Y(y)dy = f_X(x)dx \, .$$

To use the mnemonic to find a formula for $f_Y(y)$ you solve that equation for $f_Y(y)$. The right hand side will have dx/dy which is the derivative of x with respect to y when you have a formula for x in terms of y. The x is $f_X(x)$ must be replaced by the equivalent formula using y to make sure your formula for $f_Y(y)$ has *only* y in it – not x.

Example: Suppose $X \sim$ 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 $Y = \log X$ or $g(x) = \log(x)$. Solve $y = \log x$ to get $x = \exp(y)$ or $g^{-1}(y) = e^y$. Then $g'(x) = 1/x$ and $1/g'(g^{-1}(y)) = 1/(1/e^y) = e^y$. Hence

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

For any y, $e^y > 0$ so the indicator is always just 1. Thus

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

Now define $\phi = \log\beta$ and $\theta = 1/\alpha$; this is called a *reparametrization.* Then

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

This is the **Extreme Value** density with **location** parameter ϕ and **scale** parameter θ. You should be warned that there are several distributions are called “Extreme Value”.

Marginalization. Sometimes we have a few variables which come from many variables and we want the joint distribution of the few. For example we might want the joint distribution of $\bar{X}$ and s^2 when we have a sample of size n from the normal distribution. We often approach this problem in two steps. The first step, which I describe later, involves padding out the list of the few variables to make as many as the number of variables you started with (so padding out the list with $n-2$ other variables in the normal case). Then the second step is called marginalization: compute the marginal density of the variables of interest by integrating away the others.

Here is the simplest multivariate problem. We begin with

$$X = (X_1, \ldots, X_p), \qquad Y = X_1$$

(or in general Y is any X_j). We know the joint density of X and want simply the density of Y. The relevant theorem is one I have already described:

Theorem 12 *If X has density $f(x_1, \ldots, x_p)$ and $q < p$ then $Y = (X_1, \ldots, X_q)$ has density*

$$f_Y(x_1, \ldots, x_q) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_p)\, dx_{q+1} \ldots dx_p\,.$$

In fact, $f_{X_1,\ldots,X_q}$ is the **marginal** density of $X_1, \ldots, X_q$ and f_X is the **joint** density of X. Really they are both just densities. "Marginal" just serves to distinguish it from the joint density of X.

Example: The function $f(x_1, x_2) = Kx_1x_2 1(x_1 > 0, x_2 > 0, x_1 + x_2 < 1)$ is a density provided

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

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 $K = 24$. The marginal density of X_1 is Beta$(2,3)$:

$$\begin{aligned} f_{X_1}(x_1) =& \int_{-\infty}^{\infty} 24x_1x_2 1(x_1 > 0, x_2 > 0, x_1 + x_2 < 1)\, dx_2 \\ =& 24 \int_0^{1-x_1} x_1x_2 1(0 < x_1 < 1) dx_2 \\ =& 12x_1(1-x_1)^2 1(0 < x_1 < 1)\,. \end{aligned}$$

A more general problem has $Y = (Y_1, \ldots, Y_q)$ with $Y_i = g_i(X_1, \ldots, X_p)$. We distinguish the cases where $q > p$, $q < p$ and $q = p$.

Case 1: $q > p$. In this case Y **won't** have a density for "smooth" transformations g. In fact Y will have a **singular** or discrete distribution. This problem is rarely of real interest. (But, e.g., the vector of all residuals in a regression problem has a singular distribution.)

Case 2: $q = p$. In this case we use a multivariate change of variables formula. (See below.)

Case 3: $q < p$. In this case we pad out Y–add on $p-q$ more variables (carefully chosen) say $Y_{q+1}, \ldots, Y_p$. We define these extra variables by finding functions $g_{q+1}, \ldots, g_p$ and setting, for $q < i \le p$, $Y_i = g_i(X_1, \ldots, X_p)$ and then let $Z = (Y_1, \ldots, Y_p)$. We need to choose g_i so that we can use the Case 2 change of variables on $g = (g_1, \ldots, g_p)$ to compute f_Z. We then hope to find f_Y by integration:

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

4.2 Multivariate Change of Variables

Suppose $Y = g(X) \in R^p$ with $X \in R^p$ having density f_X. **Assume g is a one to one ("injective") map,** i.e., $g(x_1) = g(x_2)$ if and only if $x_1 = x_2$. Find f_Y using the following steps (sometimes they are easier said than done).

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

Step 2 : Use our basic equation

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

and rewrite it in the form

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

Step 3 : Now we need an interpretation of the derivative $\frac{dx}{dy}$ when $p > 1$:

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

which is the so called **Jacobian**.

- Equivalent formula inverts the matrix:

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

- 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 bivariate normal density. The **standard bivariate normal density** is

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

Let $Y=(Y_1,Y_2)$ where $Y_1=\sqrt{X_1^2+X_2^2}$ and $0 \le Y_2 < 2\pi$ is the angle from the positive x axis to the ray from the origin to the point (X_1,X_2). I.e., Y is X in polar co-ordinates. Solve for x in terms of y to get:

$$X_1 = Y_1\cos(Y_2) \qquad X_2 = Y_1\sin(Y_2)$$

This makes

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

It follows that

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

It remains to compute the marginal densities of Y_1 and Y_2. Factor f_Y as $f_Y(y_1,y_2) = h_1(y_1)h_2(y_2)$ where

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

and

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

Then

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

so the marginal density of Y_1 is a multiple of h_1. The multiplier makes $\int f_{Y_1} = 1$ but in this case

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

so that Y_1 has the Weibull or Rayleigh law

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

Similarly

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

which is the **Uniform**$(0,2\pi)$ density.

I leave you the following exercise: show that $W = Y_1^2/2$ has a standard exponential distribution. Recall: by definition $U = Y_1^2$ has a χ^2 dist on 2 degrees of freedom. I also leave you the exercise of finding the χ_2^2 density. Notice that $Y_1 \perp\!\!\!\perp Y_2$.

4.3 The Multivariate Normal Distribution

In this section I present the basics of the multivariate normal distribution as an example to illustrate our distribution theory ideas.

Definition: A random variable $Z \in R^1$ has a standard normal distribution (we write $Z \sim N(0,1)$) if and only if Z has the density

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

Note: To see that this is a density let

$$I = \int_{-\infty}^{\infty} \exp(-u^2/2)du.$$

Then

$$\begin{aligned} I^2 &= \left\{ \int_{-\infty}^{\infty} \exp(-u^2/2)du. \right\}^2 \\ &= \left\{ \int_{-\infty}^{\infty} \exp(-u^2/2)du \right\} \left\{ \int_{-\infty}^{\infty} \exp(-v^2/2)dv \right\} \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\{-(u^2+v^2)/2\}dudv \end{aligned}$$

Now do this integral in polar co-ordinates by the substitution $u = r\cos\theta$ and $v = r\sin\theta$ for $0 < r < \infty$ and $-\pi < \theta \le \theta$. The Jacobian is r and we get

$$\begin{aligned} I^2 &= \int_0^{\infty} \int_{-\pi}^{\pi} r\exp(-r^2/2)d\theta dr \\ &= 2\pi \int_0^{\infty} r\exp(-r^2/2)dr \\ &= -2\pi \exp(-r^2/2)\Big|_{r=0}^{\infty} \\ &= 2\pi. \end{aligned}$$

Thus

$$I = \sqrt{2\pi}.$$

Definition: A random vector $Z \in R^p$ has a standard multivariate normal distribution, written $Z \sim MVN(0, I)$ if and only if $Z = (Z_1, \ldots, Z_p)^t$ with the Z_i independent and each $Z_i \sim N(0,1)$.

In this case according to our theorem 4.3

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

here, 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 $Z \sim MVN(0, I)$.

Remark: If the matrix A is singular then X does not have a density. This is the case for example for the residual vector in a linear regression problem.

Remark: If the matrix A is invertible we can derive the multivariate normal density by change of variables:

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

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

So

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

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 is

$$\frac{\exp\{-(x-\mu)^t\Sigma^{-1}(x-\mu)/2\}}{(2\pi)^{p/2}(\det\Sigma)^{1/2}}\,;$$

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

Here is a question: for which μ, Σ is this a density? The answer is that this is a density for any μ but if $x \in R^p$ then

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

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. (This square root matrix A can be found via the Cholesky decomposition, e.g.)

When A is singular X will not have a density because $\exists a$ such that $P(a^tX = a^t\mu) = 1$; in this case X is confined to a hyperplane. A hyperplane has p dimensional volume 0 so no density can exist.

It is still true that the distribution of X depends only on $\Sigma = AA^t$: if $AA^t = BB^t$ then $AZ+\mu$ and $BZ+\mu$ have the same distribution. This can be proved using the characterization properties of moment generating functions.

I now make a list of three basic properties of the MVN distribution.

1. All margins of a multivariate normal distribution are multivariate normal. That is, if

$$X = \left[\begin{array}{c} X_1 \\ X_2 \end{array} \right],$$

$$\mu = \left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array} \right]$$

and

$$\Sigma = \left[\begin{array}{cc} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{array} \right]$$

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

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

3. If $X \sim MVN_p(\mu, \Sigma)$ then $MX + \nu \sim MVN(M\mu + \nu, M\Sigma M^t)$. We say that an affine transformation of a multivariate normal vector is normal.

4.4 Samples from the Normal Distribution

The ideas of the previous sections can be used to prove the basic sampling theory results for the normal family. Here is the theorem which describes the distribution theory of the most important statistics.

Theorem 13 *Suppose $X_1, \ldots, X_n$ are independent $N(\mu, \sigma^2)$ random variables. Then*

1. *$\bar{X}$ (sample mean)and s^2 (sample variance) independent.*

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

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

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

Proof: Let $Z_i = (X_i - \mu)/\sigma$. Then $Z_1, \ldots, Z_p$ are independent $N(0, 1)$. So $Z = (Z_1, \ldots, Z_p)^t$ 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 $(n-1)s_Z^2 = \sum(Z_i - \bar{Z})^2$. It is therefore enough to prove the theorem in the case $\mu = 0$ and $\sigma = 1$.

Step 1: Define

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

(So that Y has 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:

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

Solve for Z from Y: $Z_i = n^{-1/2}Y_1 + Y_{i+1}$ for $1 \le i \le n-1$. Use the identity

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

to get $Z_n = -\sum_{i=2}^n Y_i + n^{-1/2}Y_1$. So M is invertible:

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

Now use the change of variables formula to find f_Y. Let $\mathbf{y}_2$ denote the 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

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

Note: f_Y is a function of y_1 times a ftn 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.

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

The first 2 parts of the theorem are now done. The third part is a homework exercise.

I now present a derivation of the χ^2 density; this is not part of the proof of the theorem but is another distribution theory example. Suppose $Z_1, \ldots, Z_n$ are independent $N(0,1)$. 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 k spherical co-ordinates in n dimensions. The θ values run from 0 to π except last θ from 0 to 2π.) Here are the derivative formulae:

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

and

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

Fix $n = 3$ to clarify the formulae. Use the shorthand $R = \sqrt{U}$ The matrix of partial derivatives is

$$\left[\begin{array}{ccc} \frac{\cos\theta_1}{2R} & -R\sin\theta_1 & 0 \\ \frac{\sin\theta_1\cos\theta_2}{2R} & R\cos\theta_1\cos\theta_2 & -R\sin\theta_1\sin\theta_2 \\ \frac{\sin\theta_1\sin\theta_2}{2R} & R\cos\theta_1\sin\theta_2 & R\sin\theta_1\cos\theta_2 \end{array}\right] \, .$$

We can find the determinant by adding $2U^{1/2}\cos\theta_j/\sin\theta_j$ times col 1 to col $j+1$ (no change in the determinant). The resulting matrix is lower triangular with diagonal entries given by

$$\frac{\cos\theta_1}{R}, \frac{R\cos\theta_2}{\cos\theta_1}, \frac{R\sin\theta_1}{\cos\theta_2}$$

Multiply these together to get

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

which I observe is non-negative for all U and θ_1. For general n every term in the first column contains a factor $U^{-1/2}/2$ while every other entry has a factor $U^{1/2}$.

Fact: multiplying a column in a matrix by c multiplies the determinant by c.

So: the Jacobian of the transformation is

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

for some function, 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 $d\theta_{n-1}\cdots d\theta_1$.

The answer has the form

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

for some c. We can evaluate c by making

$$\begin{aligned}\int f_U(u)du &= c\int_0^\infty u^{(n-2)/2}\exp(-u/2)du \\ &= 1.\end{aligned}$$

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

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

Conclusion: the χ^2_n density is

$$\frac{1}{2\Gamma(n/2)}\left(\frac{u}{2}\right)^{(n-2)/2}e^{-u/2}1(u>0)\,.$$

The fourth part of the theorem is a consequence of first 3 parts and the definition of the t_ν distribution.

Definition: $T \sim t_\nu$ if T has same distribution as

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

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

Though the proof of the theorem is now finished I will Derive the density of T in this definition as a further example of the techniques of distribution theory. Begin with the cumulative distribution function of T written in terms of Z and U:

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

Differentiate this cdf with respect to t by 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

$$\frac{d}{dt} P(T \le t) = \int_0^\infty \frac{f_U(u)}{\sqrt{2\pi}} \left(\frac{u}{\nu}\right)^{1/2} \exp\left(-\frac{t^2 u}{2\nu}\right) du \, .$$

Plug in

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

to get

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

Substitute $y = u(1 + t^2/\nu)/2$, to get

$$dy = (1 + t^2/\nu) du/2$$

$$(u/2)^{(\nu-1)/2} = [y/(1 + t^2/\nu)]^{(\nu-1)/2}$$

leading to

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

or

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

Chapter 5

Convergence in Distribution

In the previous chapter I showed you examples in which we worked out precisely the distribution of some statistics. Usually this is not possible. Instead we are reduced to approximation. One method, nowadays likely the default method, is Monte Carlo simulation. The method can be very effective for computing the first two digits of a probability. That generally requires about 10,000 replicates of the basic experiment. Each succeeding digit required forces you to multiply the sample size by 100. I note that in this case leading zeros after the decimal point count – so to get a decent estimate of a probability down around 10^{-4} requires more than 10^8 simulations (or some extra cleverness –see the chapter later on Monte Carlo.

In this chapter I discuss a second method – large sample, or limit, theory – in which we compute limits as $n \to \infty$ to approximate probabilities. I begin with the most famous limit of this type – the central limit theorem.

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

What is the precise meaning of statements like "X and Y have approximately the same distribution"? The desired meaning is that X and Y have nearly the same cdf. But care is needed. Here are some questions designed to try to highlight why care is needed.

Q1) If n is a large number is the $N(0, 1/n)$ distribution close to the distribution of $X \equiv 0$?

Q2) Is $N(0, 1/n)$ close to the $N(1/n, 1/n)$ distribution?

Q3) Is $N(0, 1/n)$ close to $N(1/\sqrt{n}, 1/n)$ distribution?

Q4) If $X_n \equiv 2^{-n}$ is the distribution of X_n close to that of $X \equiv 0$?

Answers depend 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 $N(0, 1)$. So: we want to know probabilities like $P(X > x)$ are nearly $P(N(0, 1) > x)$. The real difficulties arise in the case of discrete random variables or in infinite dimensions: the latter is not done in this course. For discrete variables the following discussion highlights some of the problems. See the homework for an example of the so-called local central limit theorem.

Mathematicians mean one of two things by "close": Either they can provide an upper bound on the distance between the two things or they are talking about taking a limit. In this course we take limits.

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 14 *For real random variables X_n, X the following are equivalent:*

1. *X_n converges in distribution to X.*
2. *$P(X_n \le x) \to P(X \le x)$ for each x such that $P(X = x) = 0$*
3. *The limit of the characteristic functions of X_n is the characteristic function of X:*

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

   for every real t.

These are all implied by

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

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

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

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

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

  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.

Figure 5.1: Comparison of the $N(0, 1/n)$ distribution and point mass at 0.

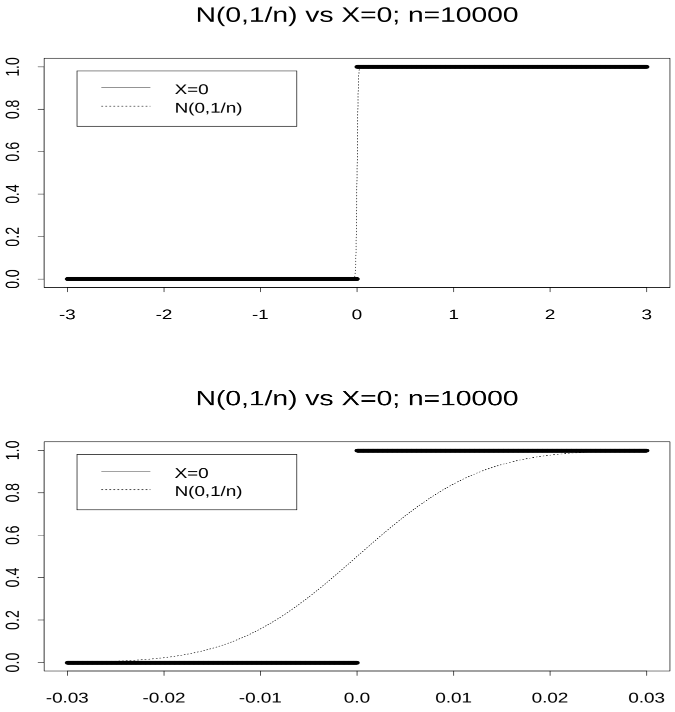


Figure 5.2: Comparison of the $N(0, 1/n)$ distribution and the $N(1/n, 1/n)$ distribution.

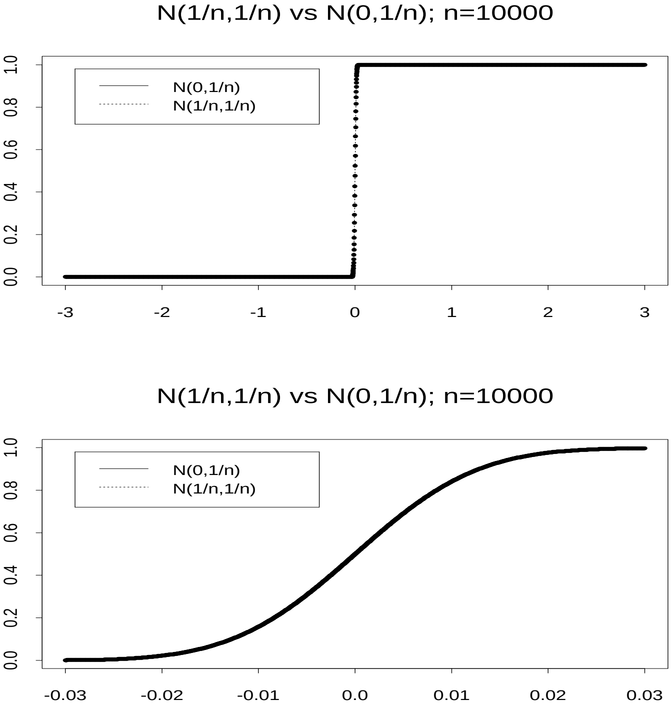
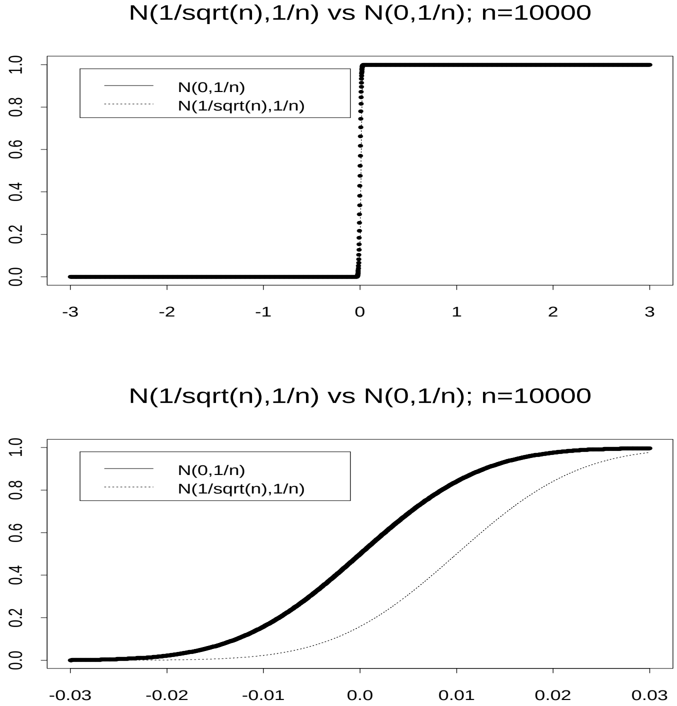


Figure 5.3: Comparison of the $N(n^{-1/2}, 1/n)$ distribution and the $N(0, 1/n)$ distribution.



- 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 $\sqrt{n}Y_n \sim N(0,1)$. 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 X_n and Y_n in the previous point 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.

Summary: to derive approximate distributions:

Show that a sequence of random variables X_n converges to some X. The limit distribution (i.e. the distribution of X) should be non-trivial, like say $N(0,1)$. Don't say: X_n is approximately $N(1/n, 1/n)$. Do say: $n^{1/2}(X_n - 1/n)$ converges to $N(0,1)$ in distribution.

Theorem 15 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,*

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

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.

5.0.1 Edgeworth expansions

It is possible to improve the normal approximation, though sometimes n has to be even larger. For the moment introduce the notation $\gamma = E(X^3)$ (remember that X is standardized to have mean 0 and standard deviation 1). Then

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

keeping one more term than I did for the central limit theorem. 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

$$\log(\phi(t)) \approx$$

$$[-t^2/2 - i\gamma t^3/6 + \cdots]$$

$$- [\cdots]^2/2 + \cdots$$

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:

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

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

3. These expansions are **asymptotic**. This means that the series indicated by $\cdots$ usually does **not** converge. For instance, when $n = 25$ it may help to take the second term but get worse if you include the third or fourth or more.

4. 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

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

2. Lyapunov CLT: for each n $X_{n1}, \ldots, X_{nn}$ independent rvs with

$$E(X_{ni}) = 0$$
$$Var(\sum_i X_{ni}) = 1$$
$$\sum E(|X_{ni}|^3) \to 0$$

then $\sum_i X_{ni}$ converges to $N(0,1)$.

3. Lindeberg CLT: 1st two conditions of Lyapunov and

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

each $\epsilon > 0$. Then $\sum_i X_{ni}$ converges in distribution to $N(0,1)$. (Lyapunov's condition implies Lindeberg's.)

4. Non-independent rvs: m-dependent CLT, martingale CLT, CLT for mixing processes.

5. Not sums: Slutsky's theorem, δ method.

Theorem 16 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 $X + c$. More generally, if $f(x, y)$ is continuous then $f(X_n, Y_n) \Rightarrow f(X, c)$.*

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

Definition: We say Y_n converges to Y in probability if

$$P(|Y_n - Y| > \epsilon) \to 0$$

for each $\epsilon > 0$.

The fact is that for Y constant convergence in distribution and in probability are the same. In general convergence in probability implies convergence in distribution. Both of these are weaker than almost sure convergence:

Definition: We say Y_n converges to Y almost surely if

$$P(\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\}) = 1\,.$$

The delta method:

Theorem 17 The δ method: *Suppose:*

- *the sequence Y_n of random variables converges to some y, a constant.*

- *there is a sequence of constants $a_n \to 0$ such that if we define $X_n = a_n(Y_n - y)$ then X_n converges in distribution to some random variable X.*

- *the function f is differentiable ftn on range of Y_n.*

Then $a_n\{f(Y_n) - f(y)\}$ converges in distribution to $f'(y)X$. (If $X_n \in R^p$ and $f : R^p \mapsto R^q$ then f' is $q \times p$ 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 $s^2 = \overline{X^2} - \bar{X}^2$.

Take $Y_n = (\overline{X^2}, \bar{X})$. 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

$$\Sigma = \left[\begin{array}{cc} \mu_4 - \sigma^4 & \mu_3 - \mu(\mu^2+\sigma^2) \\ \mu_3 - \mu(\mu^2+\sigma^2) & \sigma^2 \end{array} \right]$$

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

$$n^{1/2}(s^2-\sigma^2) \approx$$

$$n^{1/2}[1,-2\mu]\left[\begin{array}{c} \overline{X^2} - (\mu^2+\sigma^2) \\ \bar{X} - \mu \end{array} \right]$$

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

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.

5.0.2 The sample median

In this subsection I consider an example which is intended to illustrate the fact that many statistics which do not seem to be directly functions of sums can nevertheless be analyzed by thinking about sums. Later we will see examples in maximum likelihood estimation and estimating equations but here I consider the sample median.

The example has a number of irritating little points surrounding the median. First, the median of a distribution might not be unique. Second, it turns out that the sample median can be badly behaved even if the population median is unique – if the density of the distribution being studied is 0 at the population median. Third the definition of the sample median is not unique when the sample size is even. We will avoid all these complications by giving an restricting our attention to distributions with a unique median, m, and a density f which is continuous and has $f(m) > 0$.

Here is the framework. We have a sample $X_1, \ldots, X_n$ drawn from a cdf F. We assume:

1. There is a unique solution $x = m$ of the equation
$$F(x) = 1/2.$$

2. The distribution F has a density f which is continuous and has
$$f(m) > 0.$$

We will define the sample median as follows. If the sample size n is odd, say $n = 2k - 1$ then the sample median, $\hat{m}$, is the kth smallest (=kth largest) X_i. If n is even, $n = 2k$ then again we let $\hat{m}$ be the kth smallest X_i. The most important point in what follows is this:

$$\{\hat{m} \le x\} = \{\sum_i 1(X_i \le x) \ge k).$$

The random variable
$$U_n(x) = \sum_i 1(X_i \le x)$$
has a Binomial(n, p) distribution with $p = F(x)$. Thus

$$\{U_n(x) \ge k\} = \left\{ \frac{\sqrt{n}[U_n(x)/n - p]}{\sqrt{p(1-p)}} \ge \frac{\sqrt{n}(k/n - p)}{\sqrt{p(1-p)}} \right\}$$

Now put $x = m + y/\sqrt{n}$ and compute

$$\lim_{n\to\infty} \frac{\sqrt{n}(k/n - p)}{\sqrt{p(1-p)}}$$

First note that $p(1-p) \to 1/4$. Then $\sqrt{n}(k/n - 1/2) \to 0$. Next

$$\lim_{n\to\infty} \sqrt{n}(1/2 - F(x)) = f(m).$$

Assembling these pieces we find

$$\lim_{n\to\infty} \frac{\sqrt{n}(k/n - p)}{\sqrt{p(1-p)}} = -2f(m)y.$$

Finally applying the central limit theorem we find

$$\frac{\sqrt{n}[U_n(x)/n - p]}{\sqrt{p(1-p)}} \xrightarrow{d} N(0, 1).$$

This gives

$$P(\sqrt{n}(\hat{m}-m) \le y) \to 1-\Phi(-2f(m)y) = \Phi(2f(m)y)$$

Setting $u = 2f(m)y$ shows

$$\sqrt{n}(\hat{m}-m) \overset{d}{\to} N(0, 1/(4f^2(m))).$$

The important take-away point is that this is another example of how the behaviour of many statistics is determined by the behaviour of averages (because $U_n(x)/n$ is an average). I remark that similar calculations apply to other quantiles.

5.1 Monte Carlo Techniques

Modern statistics is dominated by computations made by simulation. There are many many clever simulation ideas; here we discuss only the basics. We imagine we are given random variables $X_1, \ldots, X_n$ whose joint distribution is somehow specified. We are interested in some statistic $T(X_1, \ldots, X_n)$ whose distribution we want.

Here is the basic Monte Carlo method to compute the survival function of T, that is, to compute $P(T > t)$:

1. Generate $X_1, \ldots, X_n$ from the density f.
2. Compute $T_1 = T(X_1, \ldots, X_n)$.
3. Repeat this process independently N times getting statistic values $T_1, \ldots, T_N$.
4. Estimate $p = P(T > t)$ by $\hat{p} = M/N$ where M is number of repetitions where $T_i > t$.
5. Estimate the accuracy of $\hat{p}$ using $\sqrt{\hat{p}(1-\hat{p})/N}$. In the jargon of later chapters this is the estimated standard error of $\hat{p}$.

Note: The accuracy of this computational method is inversely proportional to $\sqrt{N}$.

Next: we review some tricks to make the method more accurate.

Warning: The tricks only change the constant of proportionality — the standard error is still inversely proportional to $\sqrt{N}$.

5.1.1 Generating the Sample

Step 1 in the overall outline just presented calls for "generating" samples from the known distribution of $X_1, \ldots, X_n$. In this subsection I want to try to explain what is meant. The basic idea is to carry out an experiment which is like performing the original experiment, generating an outcome ω and calculating the value of the random variables. Instead of doing a real experiment we use a *pseudo-random number generator*, a computer program which is intended to mimic the behaviour of a real random process. This relies on a basic computing tool: pseudo uniform random numbers — variables U which have (approximately) a Uniform$[0,1]$ distribution. I will not be discussing the algorithms used for such generators. Instead we take them as a given, ignore any flaws and pretend that we have a way of generating a sequence of independent and identically distributed Uniform[0,1] variables.

5.1.2 Transformation

Other distributions are often then generated by transformation:

Example: **Exponential**: If U is Uniform[0,1] then $X = -\log U$ has an exponential distribution:

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

This generator has the following pitfall: random uniform variables generated on a computer sometimes have only 6 or 7 digits. As a consequence the tail of the generated distribution (using the transformation above) is grainy.

Here is a simplified explanation. Suppose the generated value of U is always a multiple of 10^{-6}. Then the largest possible value of X is $6\log(10)$ and the number of values larger than $3\log(10) = 6.91$ is 1000

Here is an improved algorithm

- Generate U a Uniform[0,1] variable.
- Pick a small ϵ like 10^{-3} say. If $U > \epsilon$ take $Y = -\log(U)$.
- If $U \le \epsilon$ we make use of the fact that the conditional distribution of $Y - y$ given $Y > y$ is exponential. Generate an independent new uniform variable U'. Compute $Y' = -\log(U')$. Take $Y = Y' - \log(\epsilon)$.

The resulting Y has an exponential distribution. As an exercise you should check this assertion by computing $P(Y > y)$. The new Y has 1,000,000 possible values larger than $3\log(10)$ and the largest possible values is now $9\log(10)$. As a result the distribution is much less grainy.

5.1.3 General technique: inverse probability integral transform

One standard technique which is closely connected to our exponential generator is called the inverse probability integral transformation. If Y is to have cdf F we use the following general algorithm:

- Generate $U \sim Uniform[0,1]$.
- Take $Y = F^{-1}(U)$:

$$\begin{aligned} P(Y \le y) &= P(F^{-1}(U) \le y) \\ &= P(U \le F(y)) = F(y) \end{aligned}$$

Jargon: $F^{-1}(U)$ is the inverse probability integral transform. In fact $U = F(Y)$ is called the probability integral transform of Y.

Example: Suppose X has a standard exponential distribution. Then $F(x) = 1 - e^{-x}$ and $F^{-1}(u) = -\log(1-u)$. Compare this generator to our previous method where we used U instead of $1-U$. Of course U and $1-U$ both have Uniform[0,1].

Example: **Normal**: $F = \Phi$ (this is common notation for the standard normal cumulative distribution function). There is no closed form for F^{-1}. One way to generate $N(0,1)$ is to use a numerical algorithm to compute F^{-1}.

An alternative method is the Box Müller generator:

- 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
$$Y_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)\,.$$
- As an exercise: use the change of variables technique to prove that Y_1 and Y_2 are independent $N(0,1)$ variables.

5.1.4 Acceptance Rejection

Suppose we can't calculate F^{-1} but know the density f. Find some density g and constant c such that

1. $f(x) \le cg(x)$ for each x and
2. either G^{-1} is computable or we can generate observations $W_1, W_2, \ldots$ independently from g.

Then we use the following algorithm:

1. Generate W_1.
2. Compute $p = f(W_1)/(cg(W_1)) \le 1$.
3. Generate a Uniform[0,1] random variable U_1 independent of all Ws.
4. Let $Y = W_1$ if $U_1 \le p$.
5. Otherwise get new W, U; repeat until you find $U_i \le f(W_i)/(cg(W_i))$.
6. Make Y be the last W generated.
7. This Y has density f.

5.1.5 Markov Chain Monte Carlo

Recently popular tactic, particularly for generating multivariate observations.

Theorem Suppose $W_1, W_2, \ldots$ is an (ergodic) Markov chain with stationary transitions and the stationary initial distribution of W has density f. Then starting the chain with *any* initial distribution

$$\frac{1}{n}\sum_{i=1}^{n} g(W_i) \to \int g(x)f(x)dx\,.$$

Estimate things like $\int_A f(x)dx$ by computing the fraction of the W_i which land in A.

Many versions of this technique including Gibbs Sampling and Metropolis-Hastings algorithm.

Technique invented in 1950s: Metropolis et al.

One of the authors 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

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

Variance reduction

Example: In this example we simulate to estimate the distribution of the sample mean for a sample from the Cauchy distribution. The Cauchy density is

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

The basic algorithm is

1. Generate $U_1, \ldots, U_n$ uniforms.

   The basic algorithm is

2. Define $X_i = \tan^{-1}(\pi(U_i - 1/2))$.

3. Compute $T = \bar{X}$.

4. To estimate $p = P(T > t)$ use
$$\hat{p} = \sum_{i=1}^{N} 1(T_i > t)/N$$
after generating N samples of size n.

5. This estimate is unbiased.

6. Its standard error is $\sqrt{p(1-p)/N}$.

The algorithm can be improved by using *antithetic variables*. Note first that $-X_i$ also has a Cauchy distribution. Take $S_i = -T_i$. Remember that S_i has the same distribution as T_i. 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 $\hat{p}$. Then the variance of $\tilde{p}$ is
$$(4N)^{-1}\mathrm{Var}(1(T_i > t) + 1(S_i > t))$$
$$= (4N)^{-1}\mathrm{Var}(1(|T| > t))$$
which is
$$\frac{2p(1-2p)}{4N} = \frac{p(1-2p)}{2N}$$
This variance has an extra 2 in the denominator and the numerator is also smaller – particularly for p near 1/2. So we need only half the sample size to get the same accuracy.

5.1.6 Regression estimates

Suppose $Z \sim N(0,1)$. In this example we consider ways to compute
$$\theta = E(|Z|)\,.$$
To begin with we generate N iid $N(0,1)$ variables $Z_1, \ldots, Z_N$. Compute the basic estimate $\hat{\theta} = \sum |Z_i|/N$. But we know that $E(Z_i^2) = 1$. We also know that $\hat{\theta}$ is positively correlated with $\sum Z_i^2/N$. So we try
$$\tilde{\theta} = \hat{\theta} - c(\sum Z_i^2/N - 1)$$
Notice that $E(\tilde{\theta}) = \theta$ and
$$\mathrm{Var}(\tilde{\theta}) =$$
$$\mathrm{Var}(\hat{\theta}) - 2c\mathrm{Cov}(\hat{\theta}, \sum Z_i^2/N)$$
$$+ c^2\mathrm{Var}(\sum Z_i^2/N)$$

The value of c which minimizes this is

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

We can estimate c by regressing $|Z_i|$ on Z_i^2! Notice that minimization is bound to produce a smaller variance than just using $c = 0$ which is the original estimate.

Chapter 6

Introduction to 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:

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

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

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

4. Hypothesis testing: We must choose between $\theta_0 \in \Theta_0$ and $\theta_0 \notin \Theta_0$ where $\Theta_0 \subset \Theta$.

5. 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. Some of the main schools of thought can be summarized roughly as follows:

- **Neyman Pearson**: A statistical procedure is evaluated by its long run frequency performance. Imagine repeating the data collection exercise many times, independently. Quality of procedure measured by its average performance when true distribution of X values is P_{θ_0}.

  For instance, estimates are studied by computing their sampling properties such as mean, variance, bias and mean squared error.

Definition: If $\hat\phi$ is an estimator of some parameter ϕ then the bias, variance and mean squared error are the following functions of the unknown distribution F of the data.

Bias:

$$\text{bias}_{\hat\phi}(F) = \text{E}_F(\hat\phi) - \phi(F).$$

Variance:

$$\text{bias}_{\hat\phi}(F) = \text{Var}_F(\hat\phi).$$

Mean Squared Error:

$$\text{MSE}_{\hat\phi}(F) = \text{E}\left[\left\{\hat\phi - \phi(F)\right\}^2\right].$$

Several features of these definitions deserve discussion. First, each distribution F in the model $\mathcal{F}$ must have some value for the parameter ϕ. We denote this value $\phi(F)$ in the definitions above. In parametric models the distribution F is indexed by the parameter θ and we write $\phi(\theta)$ instead of $\phi(F)$. Second, the subscripts F on E and Var remind us that while the model has many possible distributions when we come to compute probabilities and moments we have to use some particular distribution. Third, notice that the subscript F, indicating which distribution goes into computing the means and variances is the same as the one going into ϕ. Fourth, you need to know the following decomposition of MSE:

$$\text{MSE} = \text{bias}^2 + \text{Variance}.$$

Finally, the idea is that good estimators have small biases, small variances and small mean squared errors. They are being judged on the basis of their long-run or average or expected performance NOT on the basis of how well they will work with today's data. This is the Neyman-Pearson approach to inference – ask the question "how well does my statistical procedure work on average?"

Confidence sets or intervals are also to be judged on the basis of their average performance. A confidence set is a random subset $C(X)$ of Θ or Φ (where Φ is the set of possible values of some parameter ϕ). The set has *level* β if

$$P_F(\phi(F) \in C(X)) \ge \beta$$

for all $F \in \mathcal{F}$. It is absolutely crucial to note that the only thing random in this formula is the set $C(X)$, NOT, $\phi(F)$. That means that the probability describes the average behaviour of the procedure used to compute the set $C(X)$ NOT the behaviour on today's data set.

Several details should be mentioned. First if we replace $\ge \beta$ by $\equiv \beta$ then the set is *exact*. Second the random set $C(X)$ is usually just a random interval $[L(X), U(X)]$ – all the values of ϕ between these two random limits. Third in practice the desired property is more stringent that we can achieve. Generally we can only replace $\ge \beta$ with the assertion that the probability is approximately β or approximately some number $\geq \beta$.

Example: You all know that for samples of size n from the $N(\mu, \sigma^2)$ distribution the interval

$$\bar{X} \pm t_{n-1,\alpha/2} s/\sqrt{n} \text{ or } L = \bar{X} - t_{n-1,\alpha/2} s/\sqrt{n} \text{ to } U = barX + t_{n-1,\alpha/2} s/\sqrt{n}$$

is an exact level $1-\alpha$ confidence interval for μ. (As usual $t_{\nu,\alpha}$ is the upper α critical point for a Student's t distribution on ν degrees of freedom.

There are more features to discuss in a confidence interval beyond its coverage probability $P_F(\phi \in C(X))$. For instance the probability it does not include a given wrong value of ϕ should be high. The set should be as small as possible since that corresponds to a precise estimate of ϕ.

Hypothesis tests are judged on the basis of error rates. For problems when a hypothesis is true we ask how often we conclude the hypothesis is true. The probability we incorrectly conclude the hypothesis is wrong is an error rate. Note particularly that we just ask what fraction of data sets the procedure works for, NOT, whether or not it appears likely to work with today's data.

- **Bayes**: Treat θ as random just like X. Compute conditional law of unknown quantities given known quantities. In particular ask how a procedure will work on the data we actually got – no averaging over data we might have got.

  For point estimation the Bayesian would study the distribution of the estimation error $\hat{\phi}(X) - \phi(F)$ *given* the data X. Now only F is random – X is known and treated as a fixed deterministic object. The Bayesian then chooses $\hat{\phi}(X)$ to make the estimation error as small as possible – as measured by some feature of its distribution give X; this distribution is called a *posterior* distribution since it applies *after* the data are observed.

  For confidence sets the Bayesian, too, would work out a set $C(X)$ of values of ϕ which s/he considers likely to contain the true value but now the Bayesian wants

  $$P(\phi \in C(X)|X)$$

  to be large while making $C(X)$ as small as possible. Typically the Bayesian insists that

  $$P(\phi \in C(X)|X) = \beta$$

  for some given β. The Bayesian asks only about today's data X as s/he observed it and not about other data which might have been observed but was not.

  For hypothesis testing the Bayesian naturally computes the probability, given X that each hypothesis is correct.

- **Likelihood**: Try to combine previous 2 by looking only at actual data while trying to avoid treating θ as random.

  I will try, later in the course, to describe this school of inference.

We use the Neyman Pearson approach to evaluate the quality of likelihood and other methods in this course – and even to study the behaviour of Bayesian methods.

6.1 Nonparametric Inference: an introduction

6.1.1 The Empirical Distribution Function – EDF

The most common interpretation of probability is that the probability of an event is the long run relative frequency of that event when the basic experiment is repeated over and over independently. So, for instance, if X is a random variable then $P(X \leq x)$ should be the fraction of X values which turn out to be no more than x in a long sequence of trials. In general an empirical probability or expected value is just such a fraction or average computed from the data.

To make this precise, suppose we have a sample $X_1, \ldots, X_n$ of iid real valued random variables. Then we make the following definitions:

Definition: The empirical distribution function, or EDF, is

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x).$$

This is a cumulative distribution function. It is an estimate of F, the cdf of the Xs. People also speak of the empirical distribution of the sample:

$$\hat{P}(A) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \in A)$$

This is the probability distribution whose cdf is $\hat{F}_n$.

Now we consider the qualities of $\hat{F}_n$ as an estimate, the standard error of the estimate, the estimated standard error, confidence intervals, simultaneous confidence intervals and so on. To begin with we describe the best known summaries of the quality of an estimator: bias, variance, mean squared error and root mean squared error.

6.1.2 Bias, variance, MSE and RMSE

There are many ways to judge the quality of estimates of a parameter ϕ; all of them focus on the distribution of the estimation error $\hat{\phi} - \phi$. This distribution is to be computed when ϕ is the *true* value of the parameter. For our non-parametric iid sampling model the estimation error we are interested in is

$$\hat{F}_(x) - F(x)$$

where F is the true distribution function of the Xs.

The simplest summary of the size of a variable is the *root mean squared error*:

$$RMSE = \sqrt{\mathrm{E}_\theta\left[(\hat{\phi} - \phi)^2\right]}$$

In this definition the subscript θ on E is important; it specifies the true value of θ and the value of ϕ in the error must match the value of θ. For example if we were studying the $N(\mu, 1)$ model and estimating $\phi = \mu^2$ then the θ in the subscript would be μ and the ϕ in the error would be μ^2 and the two values of μ would be required to be the same.

The RMSE is measured in the same units as ϕ. That is if the parameter ϕ is a certain number of dollars then the RMSE is also some number of dollars. This makes the RMSE more scientifically meaningful than the more commonly discussed (by statisticians) mean squared error or MSE. The latter has, however, no square root and many formulas involving the MSE therefore look simpler. The weakness of MSE is one that it shares with the variance. For instance one might survey household incomes and get a mean of say $40,000 with a standard deviation of $50,000 because income distributions are very skewed to the right. The variance of household income would then be 2,500,000,000 squared dollars – ludicrously hard to interpret.

Having given that warning, however, it is time to define the MSE:

Definition: The mean squared error (MSE) of any estimate is

$$\begin{aligned} MSE &= \mathrm{E}_\theta\left[(\hat\phi-\phi)^2\right] \\ &= \mathrm{E}_\theta\left[(\hat\phi-\mathrm{E}_\theta(\hat\phi)+\mathrm{E}_\theta(\hat\phi)-\phi)^2\right] \\ &= \mathrm{E}_\theta\left[(\hat\phi-\mathrm{E}_\theta(\hat\phi))^2\right]+\left\{\mathrm{E}_\theta(\hat\phi)-\phi\right\}^2 \end{aligned}$$

In making this calculation there was a cross product term which you should check is 0. The two terms in this formula have names: the first is the variance of $\hat\phi$ while the second is the square of the bias.

Definition: The **bias** of an estimator $\hat\phi$ is

$$\mathrm{bias}_{\hat\phi}(\theta) = \mathrm{E}_\theta(\hat\phi) - \phi$$

Notice that it depends on θ. The ϕ on the right hand side also depends on the parameter θ.

Thus our decomposition above says

$$MSE = \mathrm{Variance} + (\mathrm{bias})^2.$$

In practice we often find there is a trade-off; if we try to make the variance small we often increase the bias. Statisticians often speak of a "variance-bias trade-off".

We now apply these ideas to the EDF. The EDF is an *unbiased* estimate of F. That is,

$$\begin{aligned} \mathrm{E}[\hat F_n(x)] &= \frac{1}{n}\sum_{i1=}^{n}\mathrm{E}[1(X_i \le x)] \\ &= \frac{1}{n}\sum_{i=1}^{n}F(x) = F(x) \end{aligned}$$

so the bias of $\hat{F}_n(x)$ is 0. The mean squared error is then

$$\text{MSE} = \text{Var}(\hat{F}_n(x)) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}[1(X_i \le x)] = \frac{1}{n}F(x)[1-F(x)].$$

This is very much the most common situation: the MSE is proportional to $1/n$ in large samples. So the RMSE is proportional to $1/\sqrt{n}$.

The EDF is a sample average and all sample averages are unbiased estimates of their expected values. There are many estimates in use – most estimates really – which are biased. Here is an example. Again we consider a sample $X_1, \ldots, X_n$. The sample mean is

$$\bar{X} = \frac{1}{n}X_i.$$

The sample second moment is

$$\bar{X^2} = \frac{1}{n}X_i^2.$$

These two estimates are unbiased estimates of $\text{E}(X)$ and $\text{E}(X^2)$. We might combine them to get a natural estimate of σ^2 if we remember that

$$\sigma^2 = \text{Var}(X) = \text{E}(X^2) - (\text{E}(X))^2.$$

It would then be natural to use $\bar{X}^2$ to estimate μ^2 which would lead to the following estimate of σ^2:

$$\hat{\sigma}^2 = \bar{X^2} - \bar{X}^2.$$

This estimate is biased, however, because it is a non-linear function of $\bar{X}$. In fact we find

$$\text{E}\left[(\bar{X})^2\right] = \text{Var}(\bar{X}) + \left[\text{E}(\bar{X})\right]^2 = \sigma^2/n + \mu^2.$$

So the bias of $\hat{\sigma}^2$ is

$$\text{E}\left[\bar{X^2}\right] - \text{E}\left[(\bar{X})^2\right] - \sigma^2 = \mu_2' - \mu^2 - \sigma^2/n - \sigma^2 = -\sigma^2/n.$$

In this case and many others the bias is proportional to $1/n$. The variance is proportional to $1/n$. The squared bias is proportional to $1/n^2$. So in large samples the variance is more important than the bias!

Remark: The biased estimate $\hat{\sigma}^2$ is traditionally changed to the usual sample variance $s^2 = n\hat{\sigma}^2/(n-1)$ to remove the bias.
WARNING: the MSE of s^2 is larger than that of $\hat{\sigma}^2$.

6.1.3 Standard Errors and Interval Estimation

Traditionally theoretical statistics courses spend a considerable amount of time on finding good estimators of parameters. The theory is elegant and sophisticated but point estimation itself is a silly exercise which we will not pursue here. The problem is that a bare estimate is of very little value indeed. Instead assessment of the likely size of the error of our estimate is essential. A confidence interval is one way to provide that assessment.

The most common kind of confidence interval is approximate:

$$\text{estimate} \pm 2 \text{ estimated } \textbf{standard error}$$

This is an interval of values $L(X) <$ parameter $< U(X)$ where U and L are random because they depend on the data.

What is the justification for the two SE interval above? In order to explain we introduce some notation.

Notation: Suppose $\hat{\phi}$ is the estimate of ϕ. Then $\hat{\sigma}_{\hat{\phi}}$ denotes the estimated standard error.

We often use the central limit theorem, the delta method, and Slutsky's theorem to prove

$$\lim_{n\to\infty} P_F\left(\frac{\hat{\phi}-\phi}{\hat{\sigma}_{\hat{\phi}}} \le x\right) = \Phi(x)$$

where Φ is the standard normal cdf:

$$\Phi(x) = \int_{-\infty}^{x} \frac{e^{-u^2/2}}{\sqrt{2\pi}} du.$$

Example: We illustrate the ideas by giving first what we will call *pointwise* confidence limits for $F(x)$. Define, as usual, the notation for the upper α critical point z_α by the requirement $\Phi(z_\alpha) = 1-\alpha$. Then we approximate

$$P_F\left(-z_{\alpha/2} \le \frac{\hat{\phi}-\phi}{\hat{\sigma}_{\hat{\phi}}} lez_{\alpha/2}\right) \approx 1-\alpha.$$

Then we solve the inequalities inside the probability to get the usual interval.

Now we apply this to $\phi = F(x)$ for one fixed x. Our estimate is $\hat{\phi} \equiv \hat{F}_n(x)$. The random variable $n\hat{\phi}$ has a Binomial distribution. So $\mathrm{Var}(\hat{F}_n(x)) = F(x)(1-F(x))/n$. The standard error is

$$\sigma_{\hat{\phi}} \equiv \sigma_{\hat{F}_n(x)} \equiv \mathrm{SE} \equiv \frac{\sqrt{F(x)[1-F(x)]}}{\sqrt{n}}.$$

According to the central limit theorem

$$\frac{\hat{F}_n(x)-F(x)}{\sigma_{\hat{F}_n(x)}} \xrightarrow{d} N(0,1)$$

(In the homework I ask you to turn this into a confidence interval.)

It is easier to solve the inequality

$$\left|\frac{\hat{F}_n(x)-F(x)}{\mathrm{SE}}\right| \le z_{\alpha/2}$$

if the term SE does not contain the unknown quantity $F(x)$. In the example above it did but we will modify the SE term by estimating the standard error. The method we follow uses a so-called *plug-in* procedure.

In our example we will estimate $\sqrt{F(x)[1-F(x)]/n}$ by replacing $F(x)$ by $\hat{F}_n(x)$:

$$\hat{\sigma}_{F_n(x)} = \sqrt{\frac{\hat{F}_n(x)[1-\hat{F}_n(x)}{n}}.$$

This is an example of a general strategy in which we start with an estimator, a confidence interval or a test statistic whose formula depends on some other parameter; we plug-in an estimate of that other parameter to the formula and then use the resulting object in our inference procedure. Sometimes the method changes the behaviour of our procedure and sometimes, at least in large samples, it doesn't.

In our example Slutsky's theorem shows

$$\frac{\hat{F}_n(x)-F(x)}{\hat{\sigma}_{F_n(x)}} \overset{d}{\rightarrow} N(0,1).$$

So there was no change in the limit *law* (which is common alternative jargon for the word *distribution*).

We now have two pointwise 95% confidence intervals:

$$\hat{F}_n(x) \pm z_{0.025}\sqrt{\hat{F}_n(x)[1-\hat{F}_n(x)]/n}$$

or

$$\{F(x) : \left|\frac{\sqrt{n}(\hat{F}_n(x)-F(x))}{\sqrt{F(x)[1-F(x)]}}\right| \le z_{0.025}\}$$

When we use these intervals they depend on x. Moreover, we usually look at a plot of the results against x. This leads to a problem. If we pick out an x for which the confidence interval is surprising or interesting to us we may well be picking one of the x values for which the confidence interval misses its target. After all, 1 out of every 20 confidence intervals with 95% coverage probabilities misses its target.

This suggests that what we really want is

$$P_F(L(X,x) \le F(x) \le U(X,x) \text{ for all } x) \ge 1-\alpha.$$

In that case the confidence intervals are called *simultaneous*. Thee are at least two possible methods: one is exact (meaning that the coverage probability of a 95% confidence interval is at least 95% for *every* choice of F, but conservative (meaning that the coverage is often quite a bit larger that 95% so that the interval is unnecessarily wide); the other method is approximate and less conservative. Here are some incomplete details.

The exact, conservative, procedure is base on the Dvoretsky-Kiefer-Wolfowitz inequality:

$$P_F(\exists x : |\hat{F}_n(x)-F(x)| > \sqrt{\frac{-\log(\alpha/2)}{2n}}) \le \alpha$$

The use of this inequality to generate confidence intervals is quite uncommon – the homework problems ask you to compare it to the next interval and to criticize its properties.

The approximate procedure is based on large sample limit theory. The following assertion is a famous piece of probability theory:

$$\lim_{n\to\infty} P_F(\exists x : |\sqrt{n}|\hat{F}_n(x) - F(x)| > y) = P(\exists t \in [0,1] : |B_0(t)| > y)$$

where B_0 is a *Brownian Bridge*. A Brownian Bridge is a stochastic process; in particular it is a Gaussian process, with mean $\mathrm{E}(B_0(x)) \equiv 0$ and covariance function

$$\mathrm{Cov}(B_0(x), B_0(y)) = \min\{x, y\} - xy.$$

I won't be describing precisely what that all means. You might consult some book or other which I will eventually cite I hope. It is possible, however, to choose y so that the probability on the right hand side above is α

6.1.4 Statistical Functionals

Not all parameters are created equal. Some of them have a meaning for all or at least most distribution functions or densities while others really only have a meaning inside some quite specific model. For instance, in the Weibull model density

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

there are two parameters: shape α and scale β. These parameters have no meaning in other densities; that is if the real density is normal we cannot say what α and β are. But every distribution has a median and other quantiles:

$$p^{\text{th}}\text{-quantile} = \inf\{x : F(x) \ge p\}.$$

Too, if r is a bounded function then every distribution has a value for the parameter

$$\phi \equiv \mathrm{E}_F(r(X)) \equiv \int r(x)dF(x).$$

Similarly, most distributions have a mean, variance and so on.

Definition: A function from set of all cdfs to real line is called a *statistical functional.*

Example: : The quantity $T(F) \equiv \mathrm{E}_F(X^2) - [\mathrm{E}_F(X)]^2$ is a statistical functional, namely, the variance of F. It is not quite defined for all F but it is defined for most.

The statistical functional

$$T(F) = \int r(x)dF(x)$$

is linear. The sample variance is not a linear functional.

Statistical functionals are often estimated using plug-in estimates so that

$$T(\hat{F}) = \int r(x) d\hat{F}_n(x) = \frac{1}{n}\sum_1^n r(X_i).$$

This estimate is unbiased and has variance

$$\sigma^2_{T(\hat{F})} = n^{-1}\left[\int r^2(x)dF(x) - \left\{\int r(x)dF(x)\right\}^2\right].$$

This variance can in turn be estimated using a plug-in estimate:

$$\hat{\sigma}^2_{T(\hat{F})} = n^{-1}\left[\int r^2(x)d\hat{F}_n(x) - \left\{\int r(x)d\hat{F}_n(x)\right\}^2\right].$$

And of course from that estimated variance we get an estimated standard error.

6.1.5 Bootstrap standard errors

When $r(x) = x$ we have $T(T) = \mu_F$ (the mean). The standard error of this estimate is $\sigma/\sqrt{n}$. The plug-in estimate of the standard error replaces σ with the sample standard deviation (but with n not $n-1$ as the divisor).

Now consider a general functional $T(F)$. The plug-in estimate of this is $T(\hat{F}_n)$. The plug-in estimate of the standard error of this estimate is

$$\sqrt{\text{Var}_{\hat{F}_n}(T(\hat{F}_n))}.$$

which is hard to read and seems hard to calculate in general. The solution is to simulate, particularly to estimate the standard error.

6.1.6 Basic Monte Carlo

To compute a probability or expected value we can simulate.

Example: To compute $P(|X| > 2)$ for some random variable X we use software to generate some number, say M, of replicates: $X_1^*, \ldots, X_M^*$ all having same distribution as X. Then we estimate the desired probability using the sample fraction. Here is some R code:

```
x=rnorm(1000000)
y =rep(0,1000000)
y[abs(x) >2] =1
sum(y)
```

This produced 45348 when I tried it which gives me the estimate $\hat{p} = 0.045348$. Using pnorm I find the correct answer is 0.04550026. So using a million samples gave 2 correct significant digits and an error of 2 in the third digit. Using $M = 10000$ has traditionally been more common, though I think that is changing. Using 10000, I got $\hat{p} = 0.0484$. In fact, the SE of $\hat{p}$ is $\sqrt{p(1-p)}/100 = 0.0021$. So error of up to 4 in second significant digit is reasonably likely.

6.1.7 The bootstrap

In the previous section we were drawing samples from some specific distribution – the normal distribution in the example. In bootstrapping the random variable X is replaced by the whole data set and we simulate by drawing samples from the distribution $\hat{F}_n$.

The idea is to generate new data sets (I will use a superscript $*$ as in X^* to indicate these are newly generated data sets) from the distribution F of X. Bu we don't know F so we use $\hat{F}_n$.

Example: Suppose we are interested in confidence intervals for the mean of a distribution. We will get them by simulating the distribution of the t pivot:

$$t = \frac{\sqrt{n}(\bar{X} - \mu)}{s}.$$

We have data $X_1, \ldots, X_n$ and as usual for statisticians we don't know μ or the cumulative distribution function F of the Xs. So we replace these by quantities computed from $\hat{F}_n$. Call $\mu^* = \int x d\hat{F}_n(x) = \bar{X}$. Then draw $X^*_{1,1}, \ldots, X^*_{1,n}$ an iid sample from the cdf $\hat{F}$. Repeat this sampling process M times computing t from the * values each time. Here is R code:

```
x=runif(5)
mustar = mean(x)
tv=rep(0,M)
tstarv=rep(0,M)
for( i in 1:M){
  xn=runif(5)
  tv[i]=sqrt(5)*mean(xn-0.5)/sqrt(var(xn))
  xstar=sample(x,5,replace=TRUE)
  tstarv[i]=sqrt(5)*mean(xstar-mustar)/sqrt(var(xstar))
}
```

This loop does two simulations. First, the variables `xn` and `tv` implement *parametric bootstrapping*: they simulate the t-pivot from a parametric model, namely, the Uniform[0,1] model. On the other hand `xstar` is bootstrap sample from the population `x` and `tstarv` is the t-pivot computed from `xstar`.

When I ran the code the first time I got the original data set

$$\texttt{x} = (0.7432447, 0.8355277, 0.8502119, 0.3499080, 0.8229354)$$

So `mustar` =0.7203655. Now let us look at side-by-side histograms of `tv` and `tstarv`:

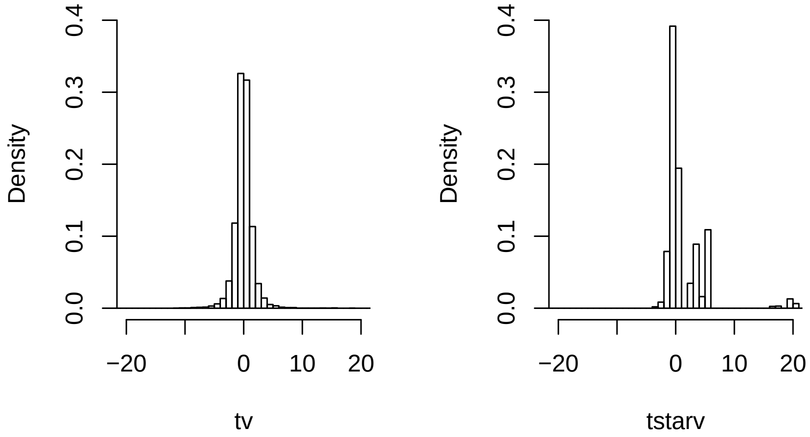
Confidence intervals: based on t-statistic: $T = \sqrt{n}(\bar{X} - \mu)/s$.

Use the bootstrap distribution to estimate $P(|T| > t)$.

Adjust t to make this 0.05. Call result c. Solve $|T| < c$ to get interval

$$\bar{X} \pm cs/\sqrt{n}.$$

Figure 6.1: Histograms of simulations. The histogram on the left is of t pivots for 1,000,000 samples of size 5 drawn from the uniform distribution. The one on the right is for t pivots computed for 1,000,000 samples of size 5 drawn from the population with just 5 elements as specified in the text



Get $c = 22.04$, $\bar{x} = 0.720$, $s = 0.211$; interval is -1.36 to 2.802. Pretty lousy interval. Is this because it is a bad idea? Repeat but simulate $\bar{X}^* - \mu^*$. Learn

$$P(\bar{X}^* - \mu^* < -0.192) = 0.025 = P(\bar{X}^* - \mu^* > 0.119)$$

Solve inequalities to get (much better) interval

$$0.720 - 0.119 < \mu < 0.720 + 0.192$$

Of course the interval missed the true value!

6.1.8 Monte Carlo Study

So how well do these methods work? We can do either a theoretical analysis or a simulation study. To describe the possible theoretical analysis let C_n be resulting interval. Usually we assume the number of bootstrap repetitions is so large that we can ignore that simulation error. Now we use theory (more sophisticated than in this course) to compute

$$\lim_{n\to\infty} P_F(\mu(F) \in C_n)$$

We say the method is *asymptotically valid* (or calibrated or accurate) if this limit is $1 - \alpha$.

The other way to assess this point is via simulation analysis: we generate many data sets of size 5 from say the Uniform[0,1] distribution. Then we carry out the bootstrap method for each data set and compute the interval C_n. Finally we count up the number of simulated uniform data sets with $0.5 \in C_n$ to get an *empirical* coverage probability. Since we will be using the method *without* knowing the true distribution we repeat the process with samples from (many) other distributions. We try to select enough distributions to give us a pretty good idea of the overall behaviour.

Remark: : Some statisticians never do anything except the simulation part. I think this is somewhat perilous – you are hard pressed to guarantee that your simulation covered all the realistic possibilities. But then, I do theory for a living.

Here is some `R` code which carries out a bit of that Monte Carlo study:

```
tstarint = function(x,M=10000){
  n = length(x)
  must=mean(x)
  se=sqrt(var(x)/n)
  xn=matrix(sample(x,n*M,replace=T),nrow=M)
  one = rep(1,n)/n
  dev= xn%*%one - must
  tst=dev/sqrt(diag(var(t(xn)))/n)
  c1=quantile(dev,c(0.025,0.975))
  c2=quantile(abs(tst),0.95)
  c(must-c1[2],must-c1[1], must -c2*se,must+c2*se)
}

lims=matrix(0,1000,4)
count=lims
for(i in 1:1000){
  x=runif(5)
  lims[i,]=tstarint(x)
}
count[,1][lims[,1]<0.5]=1
count[,2][lims[,2]>0.5]=1
count[,3][lims[,3]<0.5]=1
count[,4][lims[,4]>0.5]=1
sum(count[,1]*count[,2])
sum(count[,3]*count[,4])
```

The results for the study I did are these. For samples of size 5 from the Uniform[0,1] distribution the empirical coverage probability for the true mean of 1/2, using the bootstrap distribution of the error $\bar{X} - \mu$ is 80.4% in 1000 Monte Carlo trials using 10,000 bootstrap samples in each trial. This compares to coverage of 97.2% under the same conditions using the t pivot $\sqrt{n})(\bar{X} - \mu)/s$. (Strictly speaking t is only an approximate pivot.) For samples of size 25 I got 92.1% and 94.8%. I also tried exponential data. For $n = 5$ I got ? and ? while for $n = 25$ I got 92.1% and 94.1%.

Remark: : It is possible to put standard errors on these Monte Carlo estimates and to assess the inaccuracy induced by using only 10,000 bootstrap samples instead of infinitely many. Ignoring the latter you should be able to add standard errors to each of the percentages given. They are all roughly 0.007 or 0.7 percentage points.

Chapter 7

Estimation

7.1 Likelihood Methods of Inference

Imagine we toss a coin 6 times and get Heads twice. Let p be the probability of getting H on an individual toss and suppose the tosses are independent. Then the probability of getting exactly 2 heads is

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

This function of p is called the **likelihood** function.

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

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

Key Point: we 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 density. Notice also that the coin tossing example uses the discrete density for f.

We use likelihood for most inference problems:

1. 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.

2. 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.

3. 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.

4. Hypothesis testing: decide whether or not $\theta_0 \in \Theta_0$ where $\Theta_0 \subset \Theta$. We base our decision on the likelihood ratio
$$\frac{\sup\{L(\theta); \theta \in \Theta \setminus \Theta_0\}}{\sup\{L(\theta); \theta \in \Theta_0\}}.$$

7.1.1 Maximum Likelihood Estimation

To find the MLE we maximize L. This is a typical function maximization problem: Set the gradient of L equal to 0 and check to see that the root you find is a maximum, not a minimum or a saddle point.

Now let's examine some likelihood plots in examples:

Example: **Cauchy Data**

Suppose we have an iid sample $X_1, \ldots, X_n$ from the Cauchy(θ) density given by

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

The likelihood function is

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

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 you differentiate L, and set the result equal to 0. Notice that L is product of n terms; its derivative is

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

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

Definition: The **Log Likelihood** function is

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

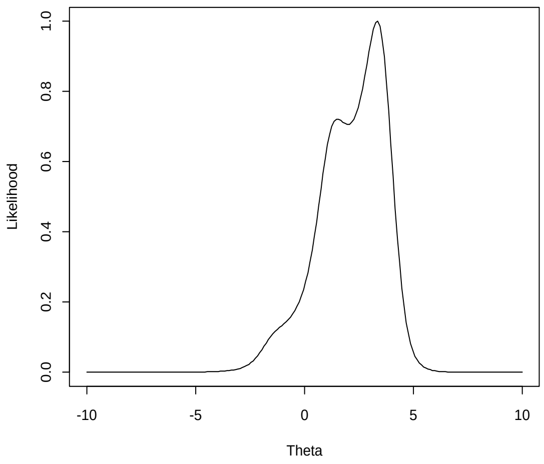
For the Cauchy problem we have

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

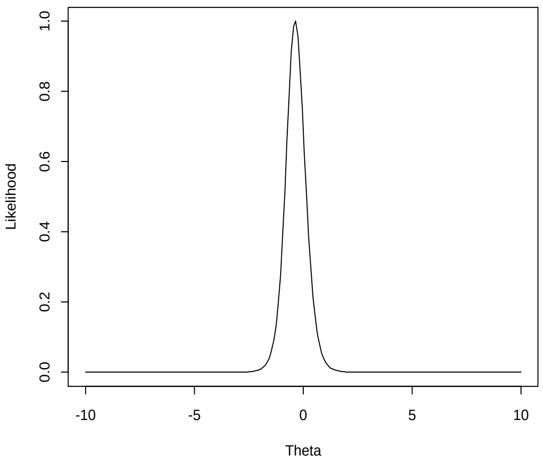
Notice the following points:

- Plots of ℓ for $n=25$ quite smooth, rather parabolic.

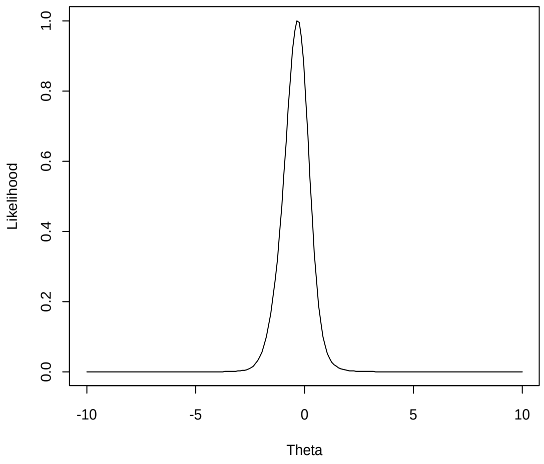
Likelihood Function: Cauchy, n=5



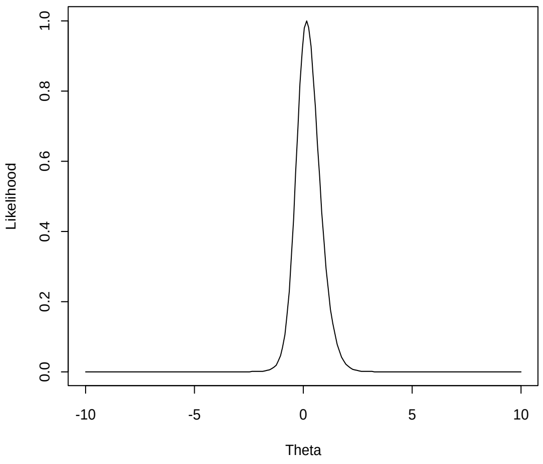
Likelihood Function: Cauchy, n=5



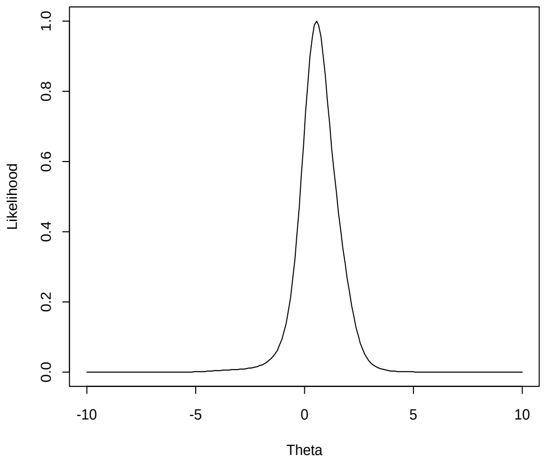
Likelihood Function: Cauchy, n=5



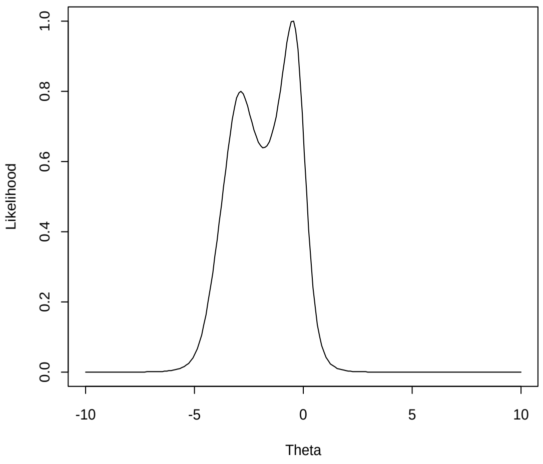
Likelihood Function: Cauchy, n=5

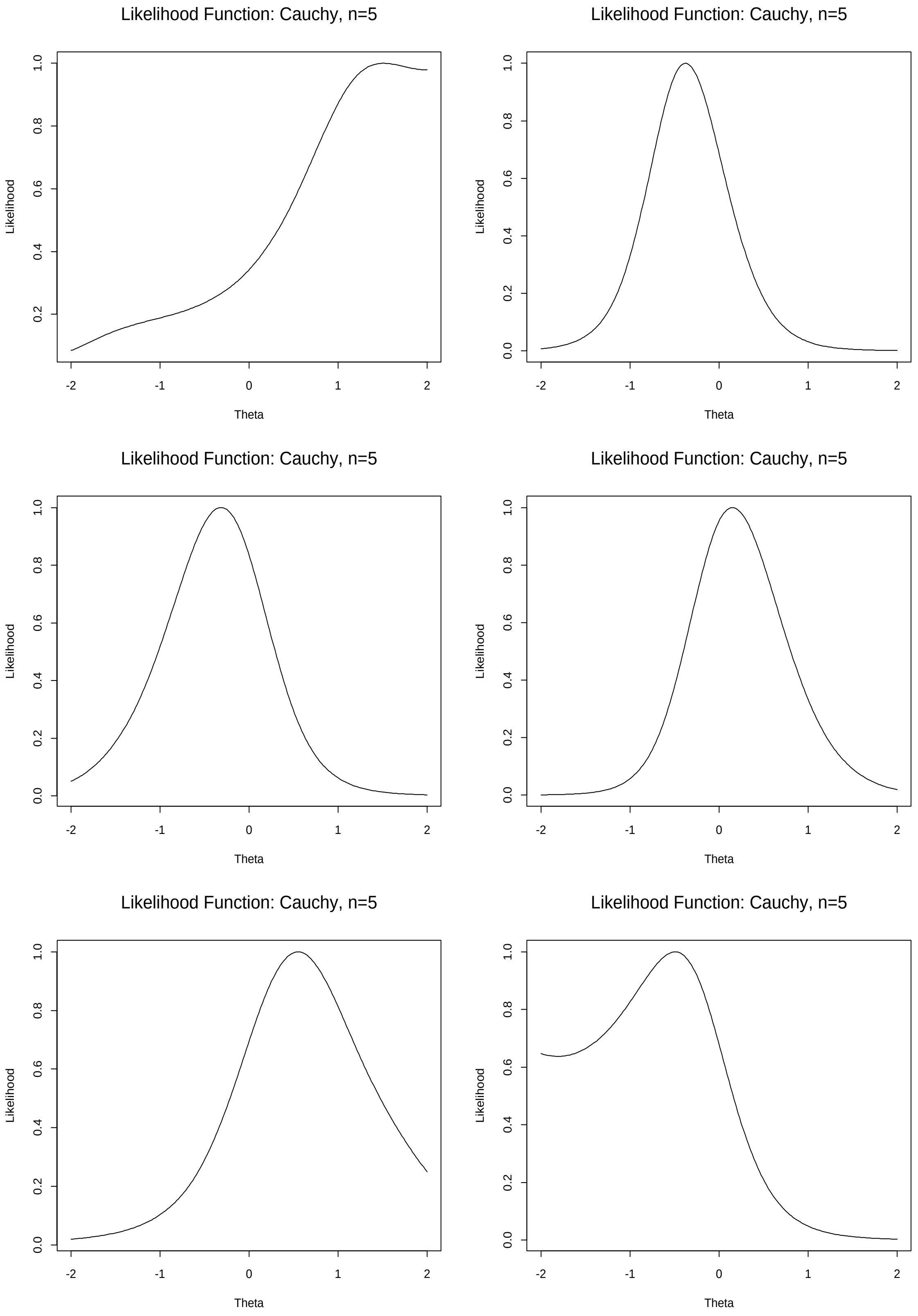


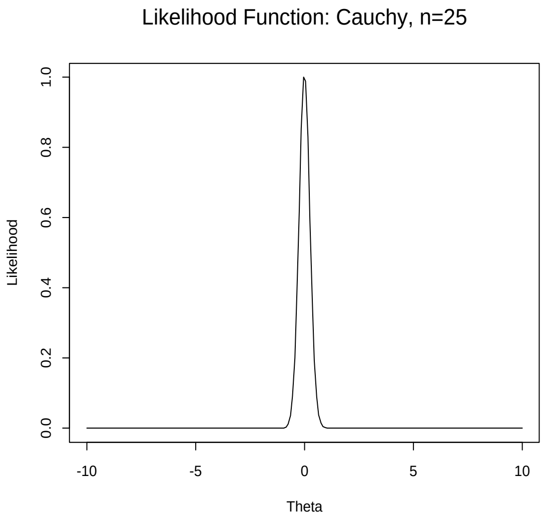
Likelihood Function: Cauchy, n=5

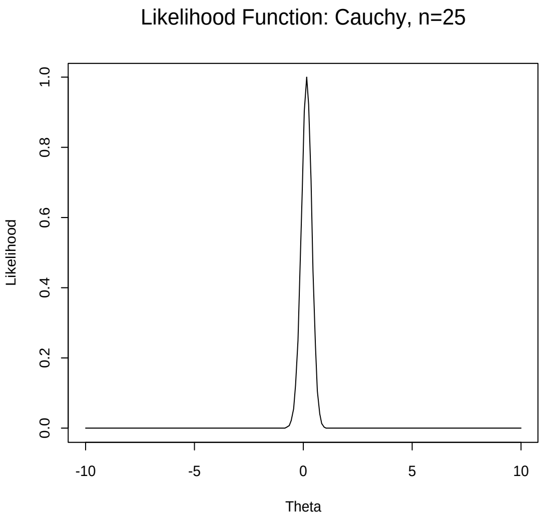


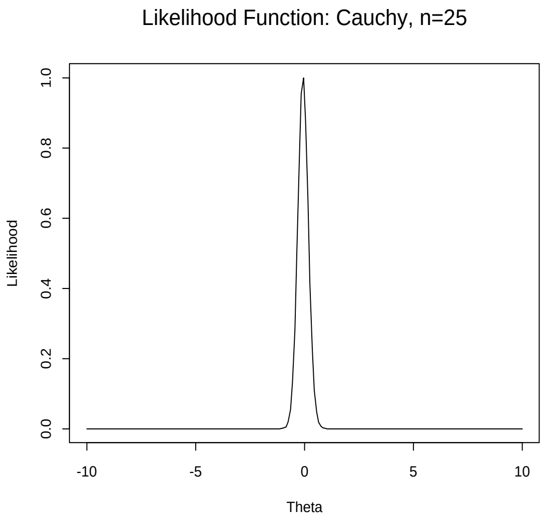
Likelihood Function: Cauchy, n=5

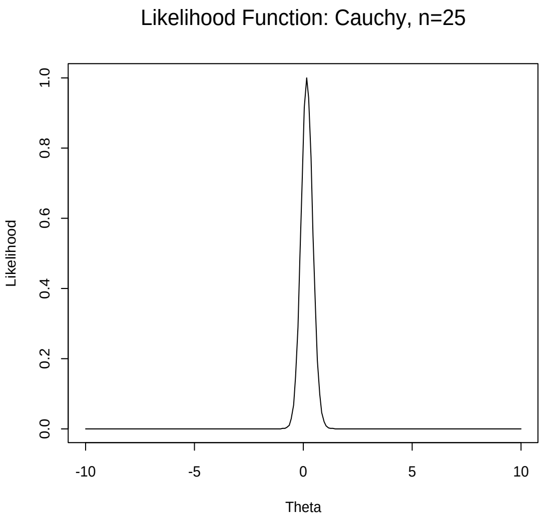


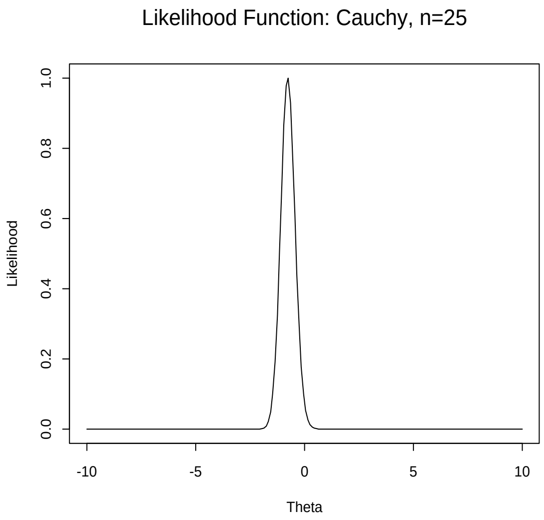

Likelihood Function: Cauchy, n=5
Likelihood Function: Cauchy, n=5
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
Theta
-2
-1
0
1
2
Likelihood Function: Cauchy, n=5
Likelihood Function: Cauchy, n=5
Likelihood Function: Cauchy, n=5
Likelihood Function: Cauchy, n=5

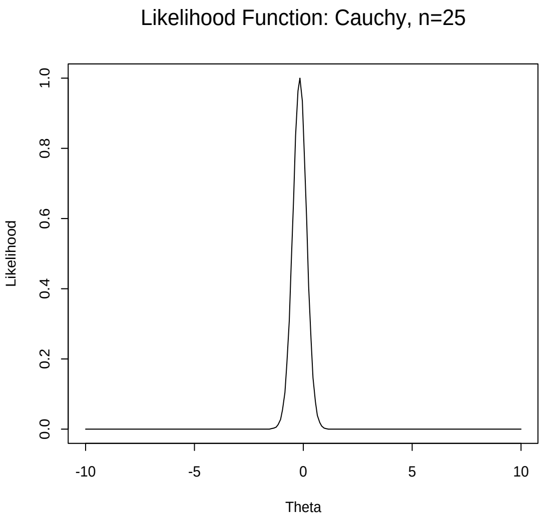

Likelihood Function: Cauchy, n=25
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
-10
-5
0
5
10
Theta

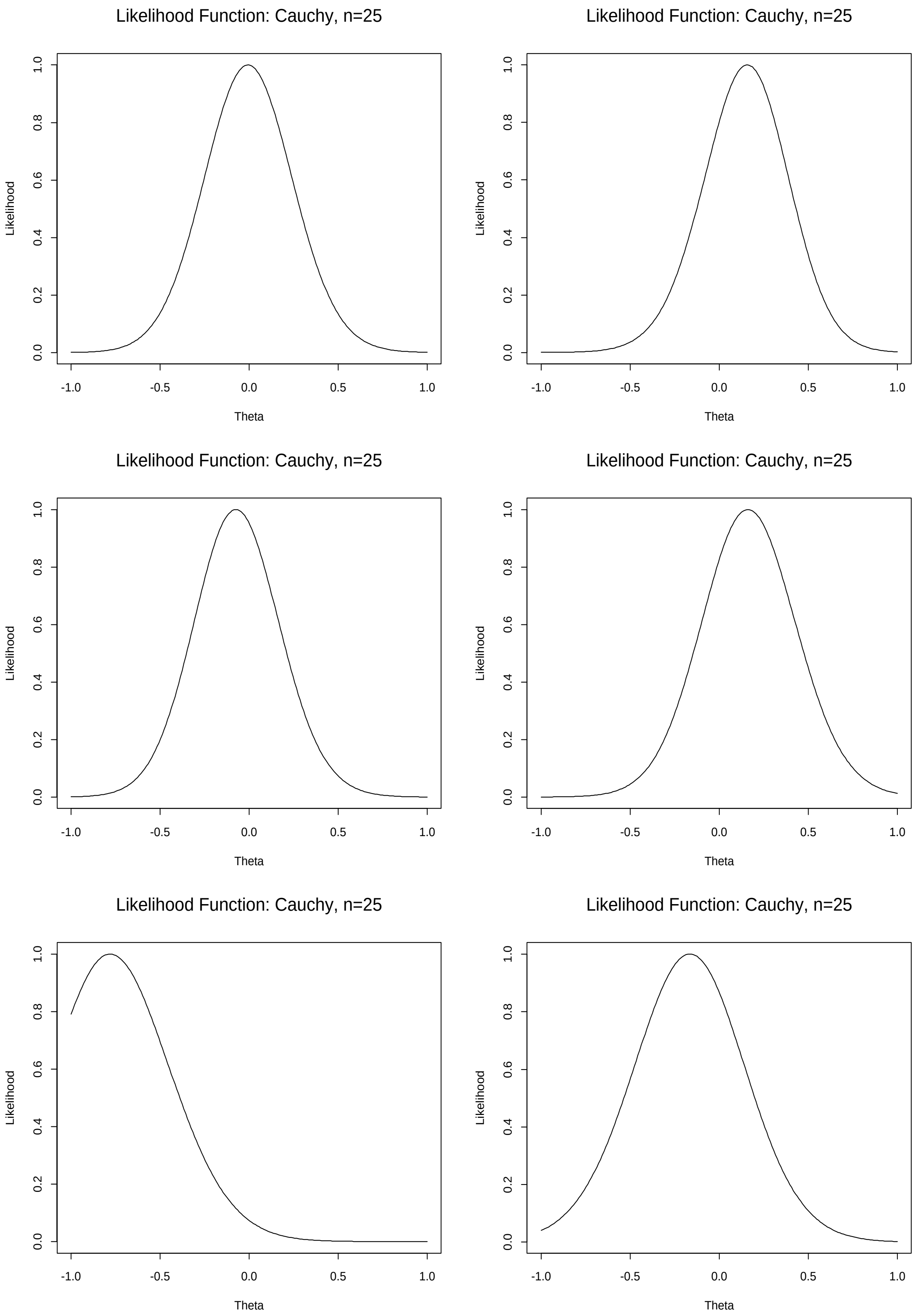

Likelihood Function: Cauchy, n=25
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
-10
-5
0
5
10
Theta

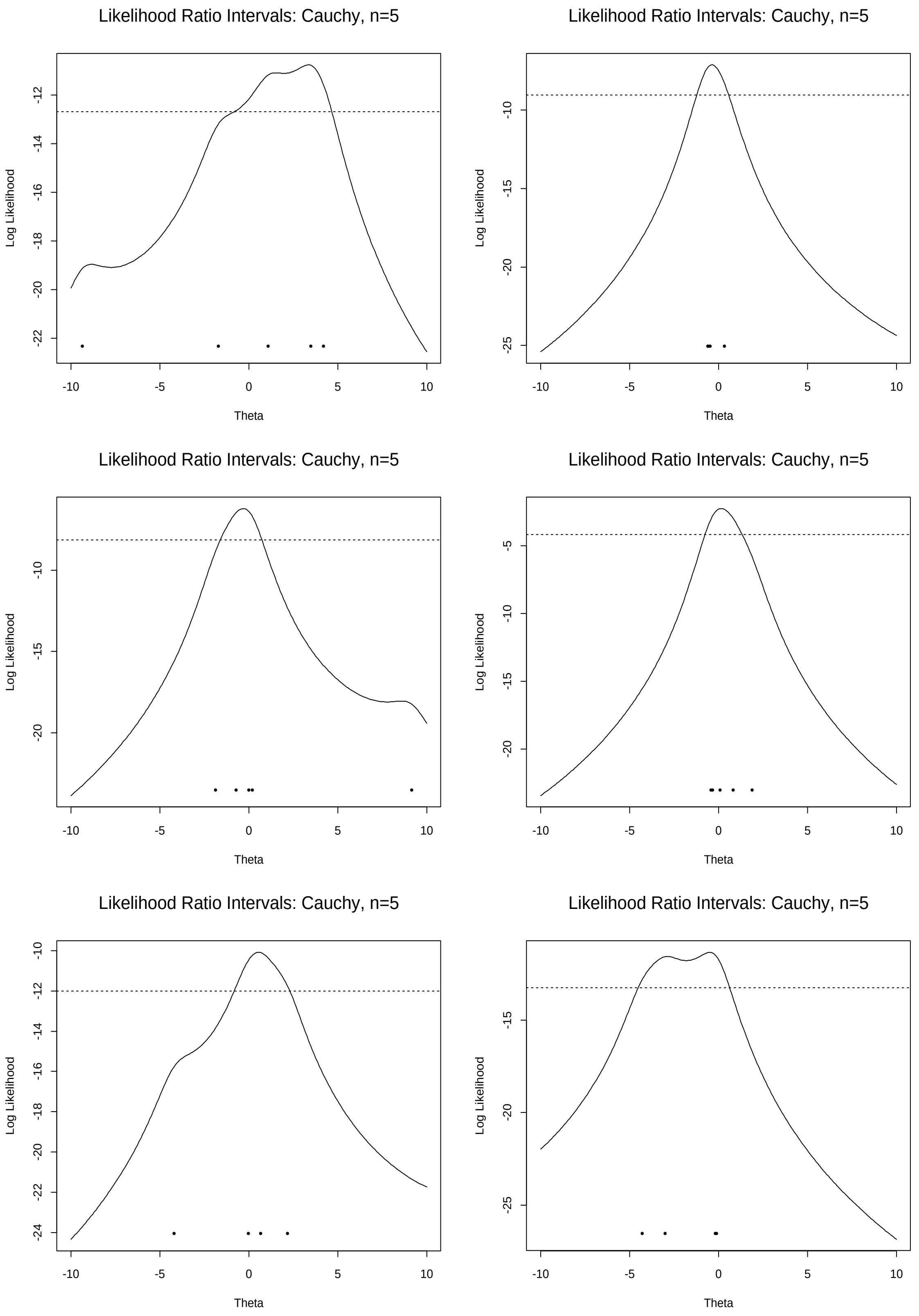

Likelihood Function: Cauchy, n=25
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
-10
-5
0
5
10
Theta

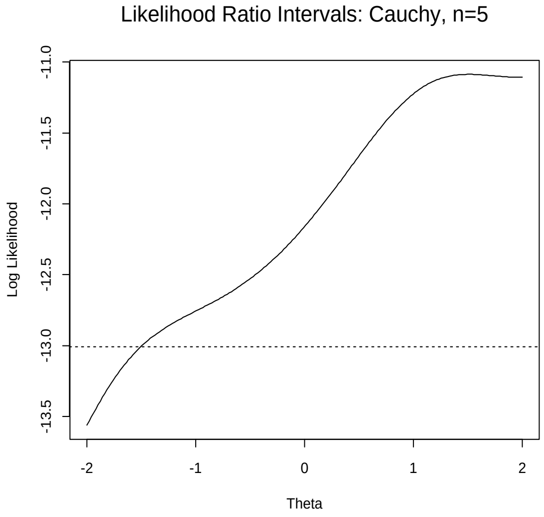

Likelihood Function: Cauchy, n=25
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
-10
-5
0
5
10
Theta

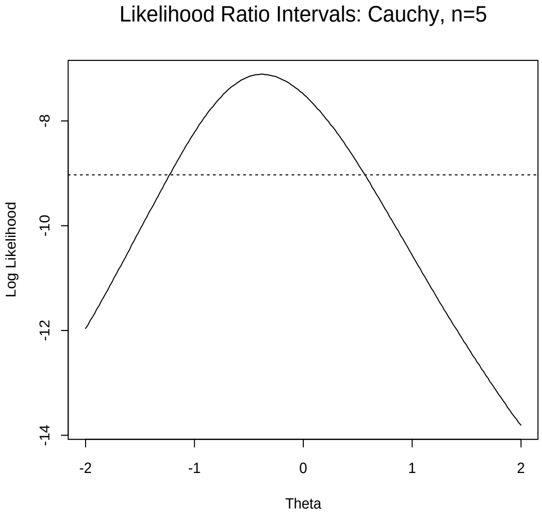

Likelihood Function: Cauchy, n=25
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
-10
-5
0
5
10
Theta

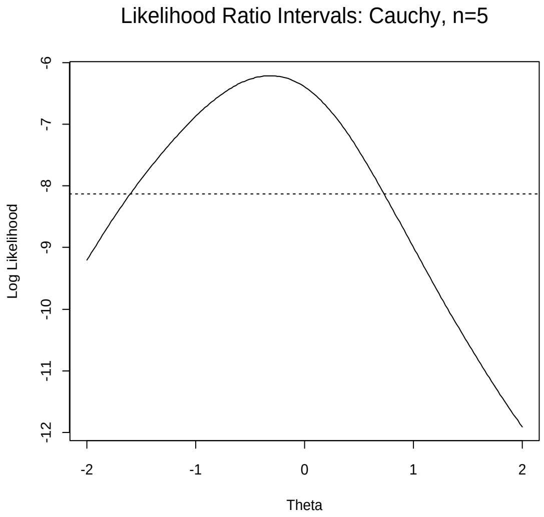

Likelihood Function: Cauchy, n=25
Likelihood
0.0
0.2
0.4
0.6
0.8
1.0
-10
-5
0
5
10
Theta

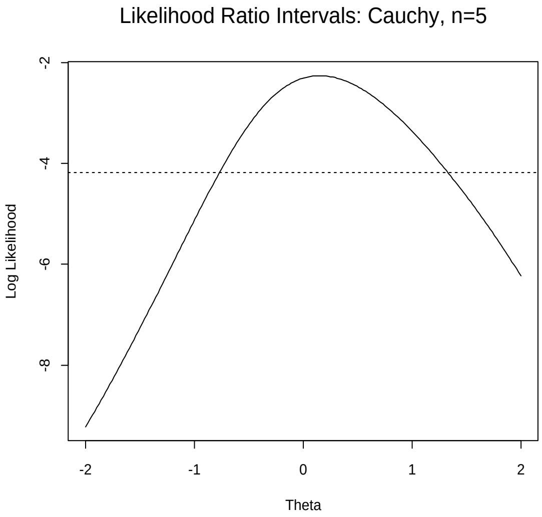

Likelihood Function: Cauchy, n=25
Likelihood
Theta
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0

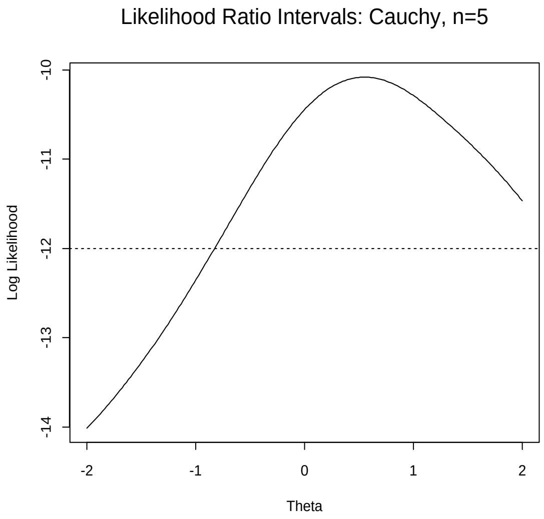

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-12
-14
-16
-18
-20
-22
-10
-5
0
5
10
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-10
-15
-20
-25
-10
-5
0
5
10
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-10
-15
-20
-10
-5
0
5
10
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-5
-10
-15
-20
-10
-5
0
5
10
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-10
-12
-14
-16
-18
-20
-22
-24
-10
-5
0
5
10
Theta
Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-15
-20
-25
-10
-5
0
5
10
Theta

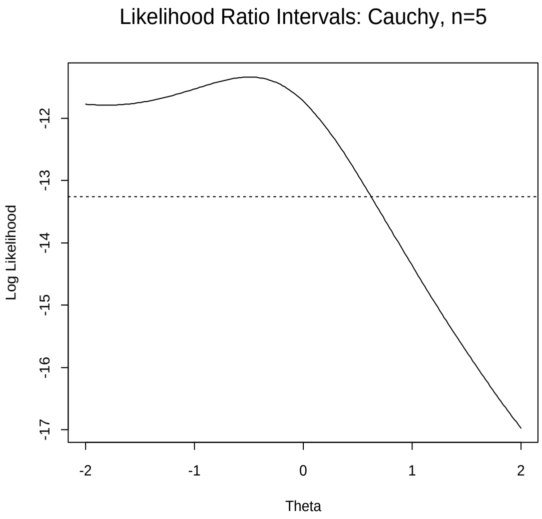

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-11.0
-11.5
-12.0
-12.5
-13.0
-13.5
-2
-1
0
1
2
Theta

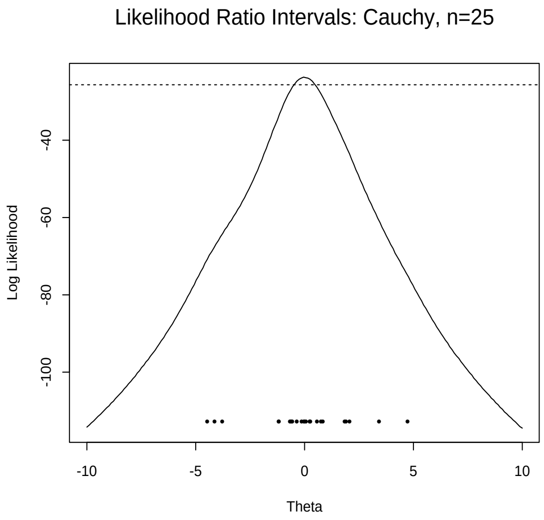

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-8
-10
-12
-14
-2
-1
0
1
2
Theta


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-6
-7
-8
-9
-10
-11
-12
-2
-1
0
1
2
Theta

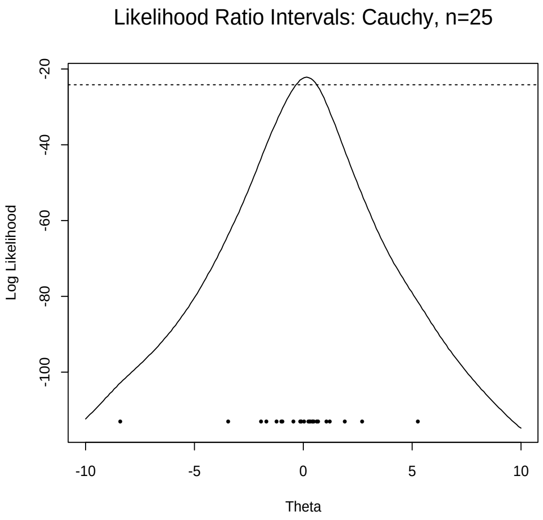

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-2
-4
-6
-8
-2
-1
0
1
2
Theta


Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-10
-11
-12
-13
-14
-2
-1
0
1
2
Theta

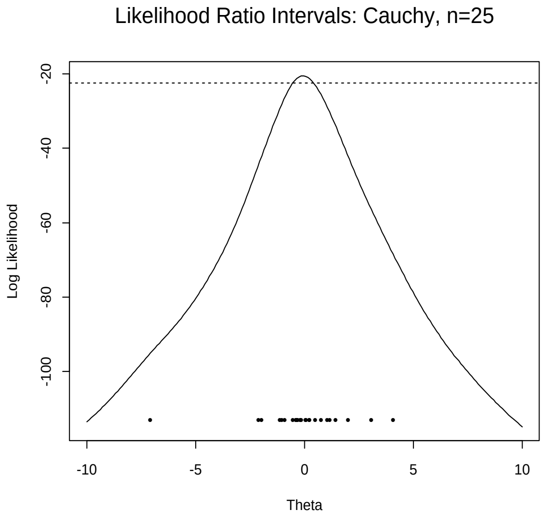

Likelihood Ratio Intervals: Cauchy, n=5
Log Likelihood
-12
-13
-14
-15
-16
-17
-2
-1
0
1
2
Theta



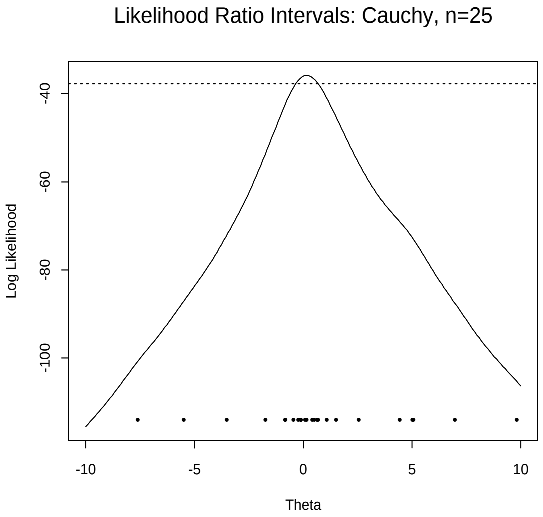
Likelihood Ratio Intervals: Cauchy, n=25



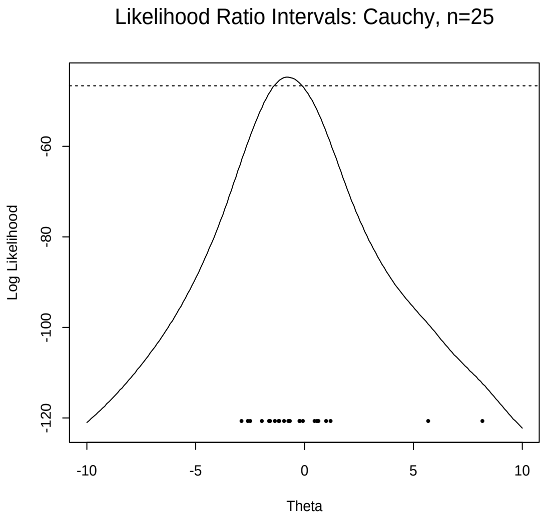
Likelihood Ratio Intervals: Cauchy, n=25



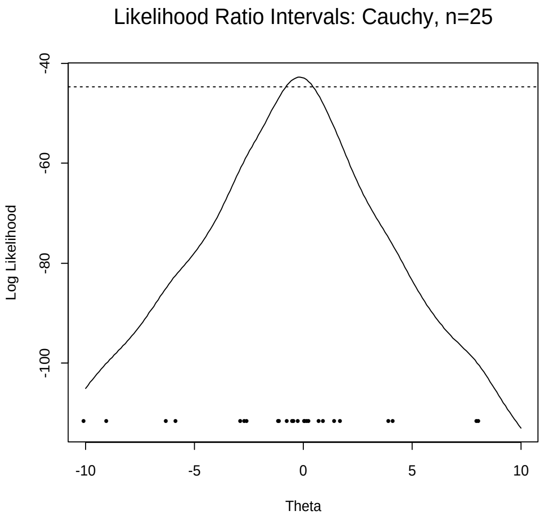
Likelihood Ratio Intervals: Cauchy, n=25



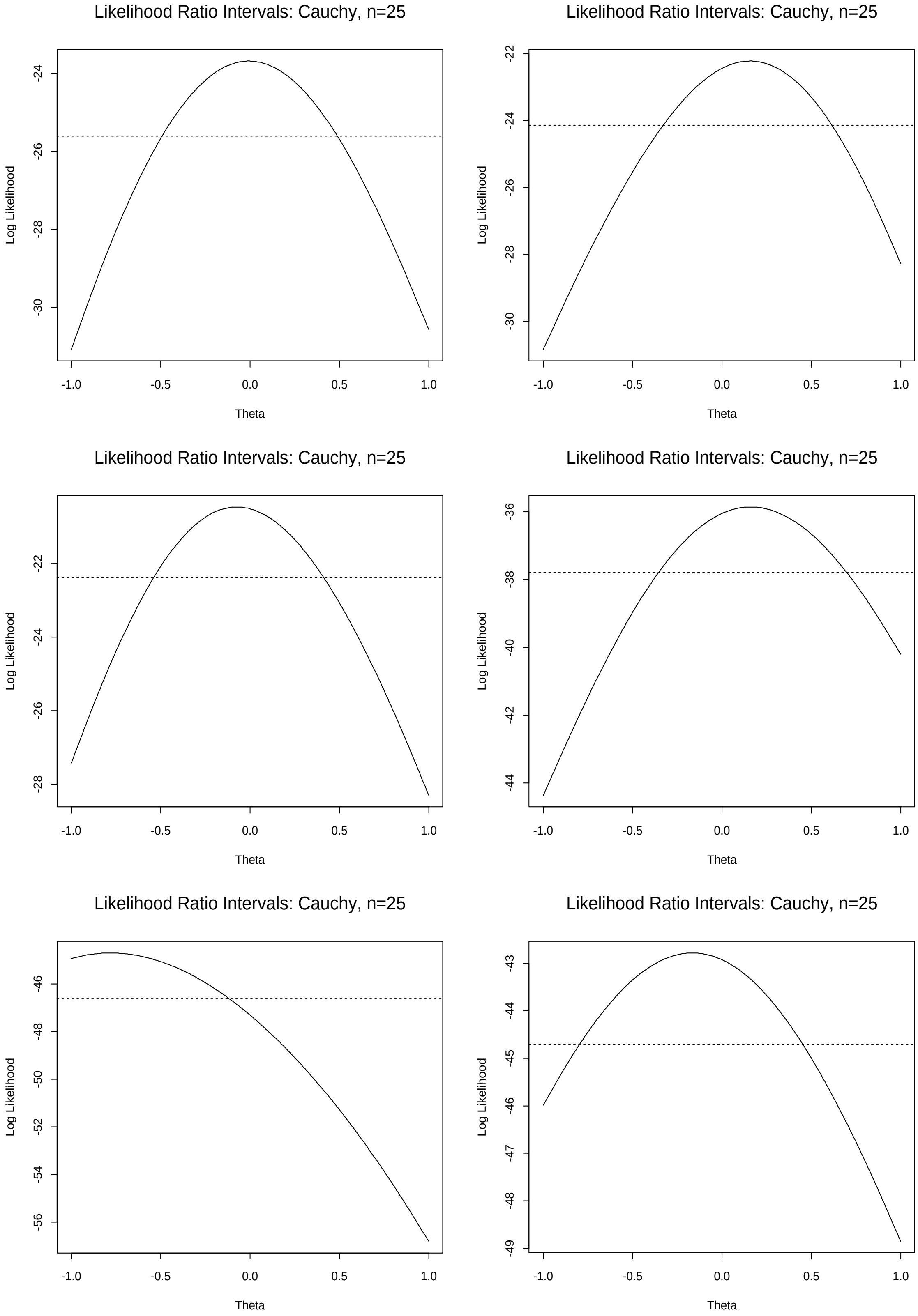
Likelihood Ratio Intervals: Cauchy, n=25



Likelihood Ratio Intervals: Cauchy, n=25



Likelihood Ratio Intervals: Cauchy, n=25


Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
-24
-26
-28
-30
-1.0
-0.5
0.0
0.5
1.0
Theta
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
-22
-24
-26
-28
-30
-1.0
-0.5
0.0
0.5
1.0
Theta
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
-22
-24
-26
-28
-1.0
-0.5
0.0
0.5
1.0
Theta
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
-36
-38
-40
-42
-44
-1.0
-0.5
0.0
0.5
1.0
Theta
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
-46
-48
-50
-52
-54
-56
-1.0
-0.5
0.0
0.5
1.0
Theta
Likelihood Ratio Intervals: Cauchy, n=25
Log Likelihood
-43
-44
-45
-46
-47
-48
-49
-1.0
-0.5
0.0
0.5
1.0
Theta

- For $n = 5$ many local maxima and minima of ℓ.

The likelihood tends to 0 as $|\theta| \to \infty$ so the maximum of ℓ occurs 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 $\hat\theta$ is usually a root of the **Likelihood Equations**

$$U(\theta) = 0$$

In our Cauchy example we find

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

[Examine plots of score functions.]

Notice: there are often multiple roots of likelihood equations.

Example: $X \sim \text{Binomial}(n, \theta)$

$$\begin{aligned} L(\theta) &= \binom{n}{X} \theta^X (1-\theta)^{n-X} \\ \ell(\theta) &= \log\binom{n}{X} + X \log(\theta) + (n - X)\log(1-\theta) \\ U(\theta) &= \frac{X}{\theta} - \frac{n-X}{1-\theta} \end{aligned}$$

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

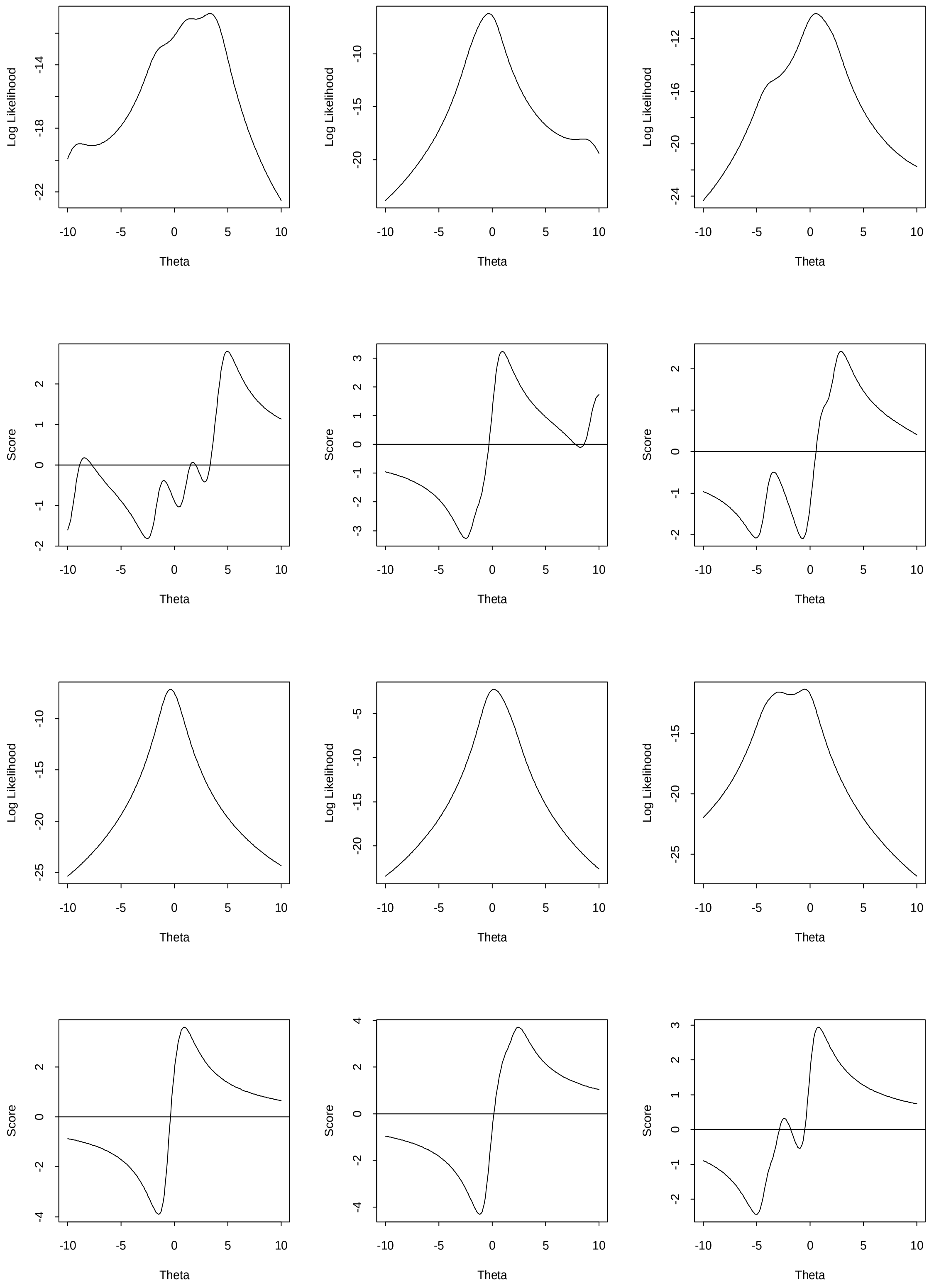
$$U(\theta) = \frac{n}{\theta} > 0$$

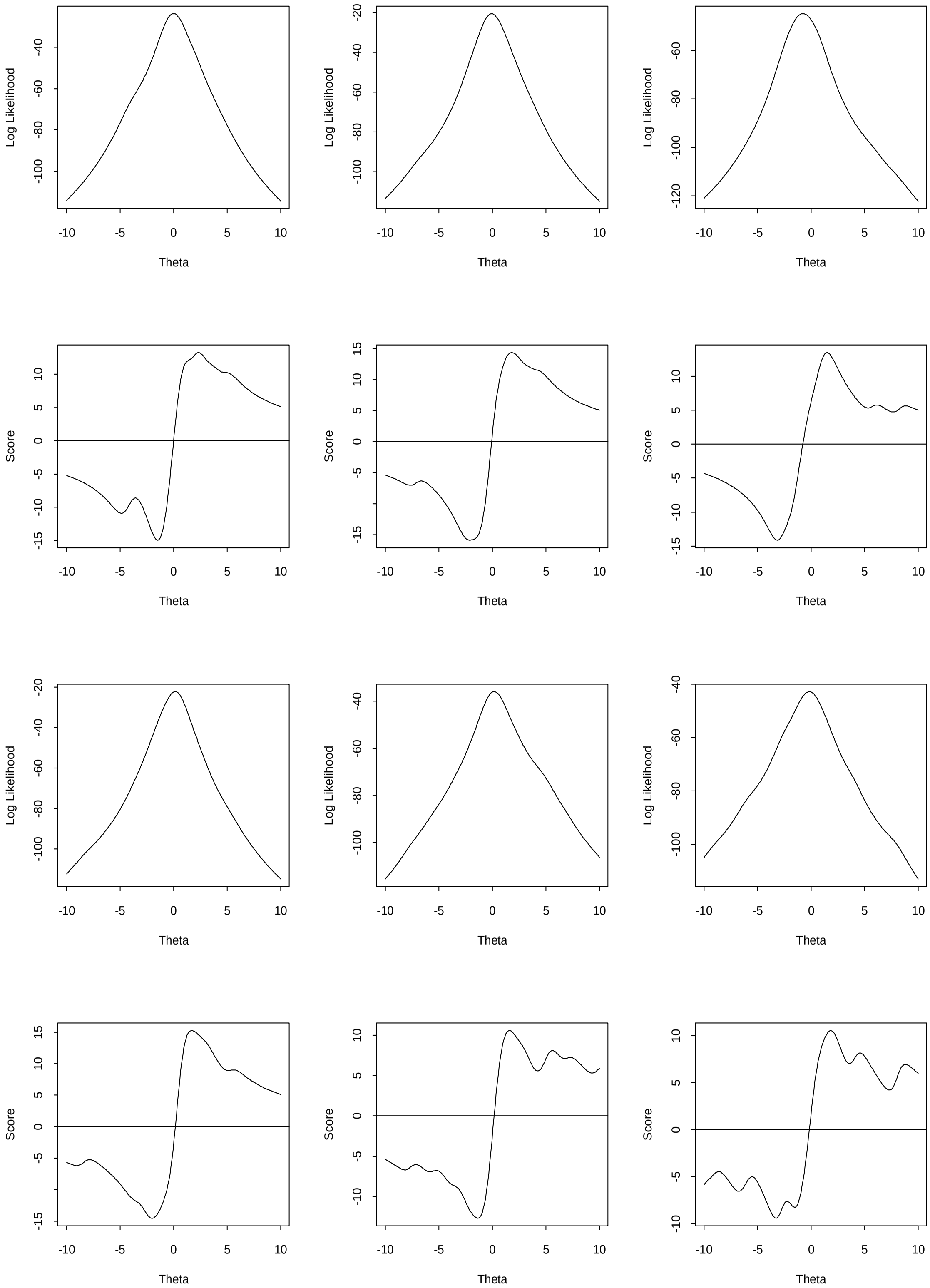
for all θ so that the likelihood is an increasing function of θ which is maximized at $\hat\theta = 1 = X/n$. Similarly when $X = 0$ the maximum is at $\hat\theta = 0 = X/n$.

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

$$\begin{aligned} L(\mu, \sigma) &= \frac{e^{-\sum (X_i - \mu)^2/(2\sigma^2)}}{(2\pi)^{n/2}\sigma^n} \\ \ell(\mu, \sigma) &= -\frac{n}{2}\log(2\pi) - \frac{\sum (X_i - \mu)^2}{2\sigma^2} - n\log(\sigma) \end{aligned}$$


Log Likelihood
-14
-18
-22
Theta
-10
-5
0
5
10
Log Likelihood
-10
-15
-20
Theta
Log Likelihood
-12
-16
-20
-24
Theta
Score
2
1
0
-1
-2
Theta
Score
3
2
1
0
-1
-2
-3
Theta
Score
2
1
0
-1
-2
Theta
Log Likelihood
-10
-15
-20
-25
Theta
Log Likelihood
-5
-10
-15
-20
Theta
Log Likelihood
-15
-20
-25
Theta
Score
2
0
-2
-4
Theta
Score
4
2
0
-2
-4
Theta
Score
3
2
1
0
-1
-2
Theta


Log Likelihood
-100
-80
-60
-40
Theta
-10
-5
0
5
10
Log Likelihood
-100
-80
-60
-40
-20
Theta
Log Likelihood
-120
-100
-80
-60
Theta
Score
-15
-10
-5
0
5
10
Theta
Score
-15
-5
0
5
10
15
Theta
Score
-15
-10
-5
0
5
10
Theta
Log Likelihood
-100
-80
-60
-40
-20
Theta
Log Likelihood
-100
-80
-60
-40
Theta
Log Likelihood
-100
-80
-60
-40
Theta
Score
-15
-5
0
5
10
15
Theta
Score
-10
-5
0
5
10
Theta
Score
-10
-5
0
5
10
Theta

and that U is

$$\left[\begin{array}{c} \frac{\sum(X_i-\mu)}{\sigma^2} \\ \frac{\sum(X_i-\mu)^2}{\sigma^3} - \frac{n}{\sigma} \end{array}\right]$$

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

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

Check this is maximum by computing one more derivative. Matrix H of second derivatives of ℓ is

$$\left[\begin{array}{cc} \frac{-n}{\sigma^2} & \frac{-2\sum(X_i-\mu)}{\sigma^3} \\ \frac{-2\sum(X_i-\mu)}{\sigma^3} & \frac{-3\sum(X_i-\mu)^2}{\sigma^4} + \frac{n}{\sigma^2} \end{array}\right]$$

Plugging in the mle gives

$$H(\hat\theta) = \left[\begin{array}{cc} \frac{-n}{\hat\sigma^2} & 0 \\ 0 & \frac{-2n}{\hat\sigma^2} \end{array}\right]$$

which is negative definite. Both its eigenvalues are negative. So $\hat\theta$ must be a local maximum.

[Examine contour and perspective plots of ℓ.]

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

For $X_1,\ldots,X_n$ iid log likelihood is

$$\ell(\theta) = \sum \log(f(X_i,\theta)) .$$

The score function is

$$U(\theta) = \sum \frac{\partial \log f}{\partial\theta}(X_i,\theta) .$$

The MLE $\hat\theta$ maximizes ℓ. If the maximum occurs in the interior of the parameter space and the log likelihood is continuously differentiable then $\hat\theta$ solves the likelihood equations

$$U(\theta) = 0 .$$

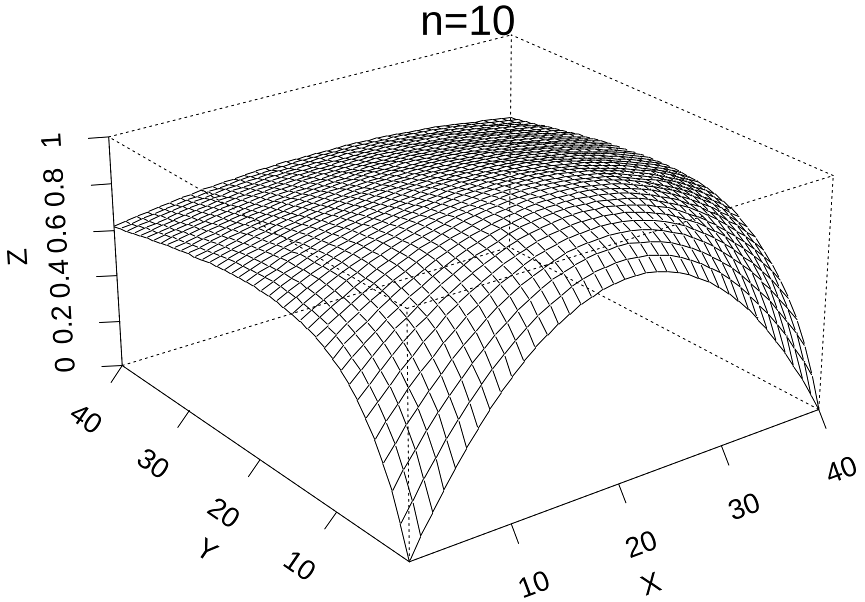
Some examples concerning existence of roots:

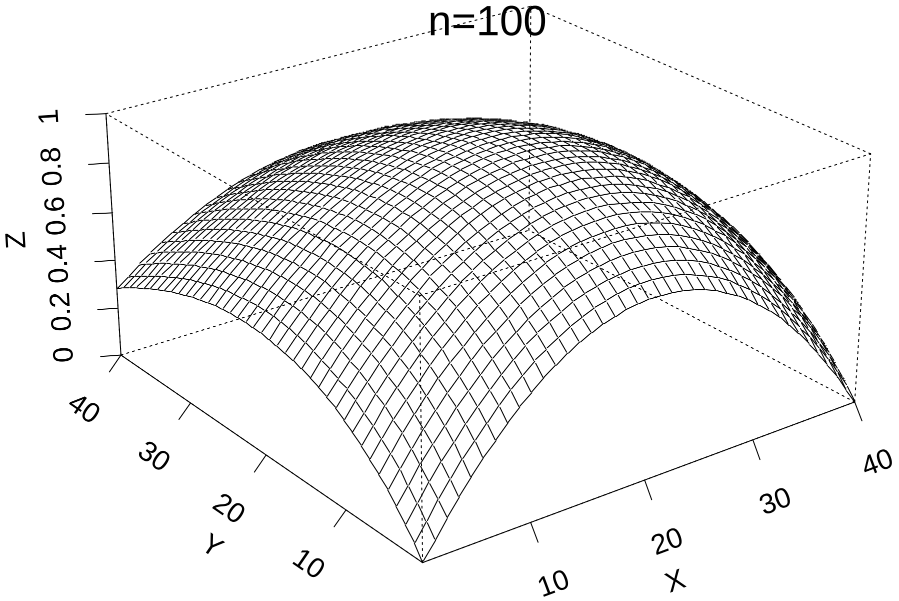
Solving $U(\theta) = 0$: Examples

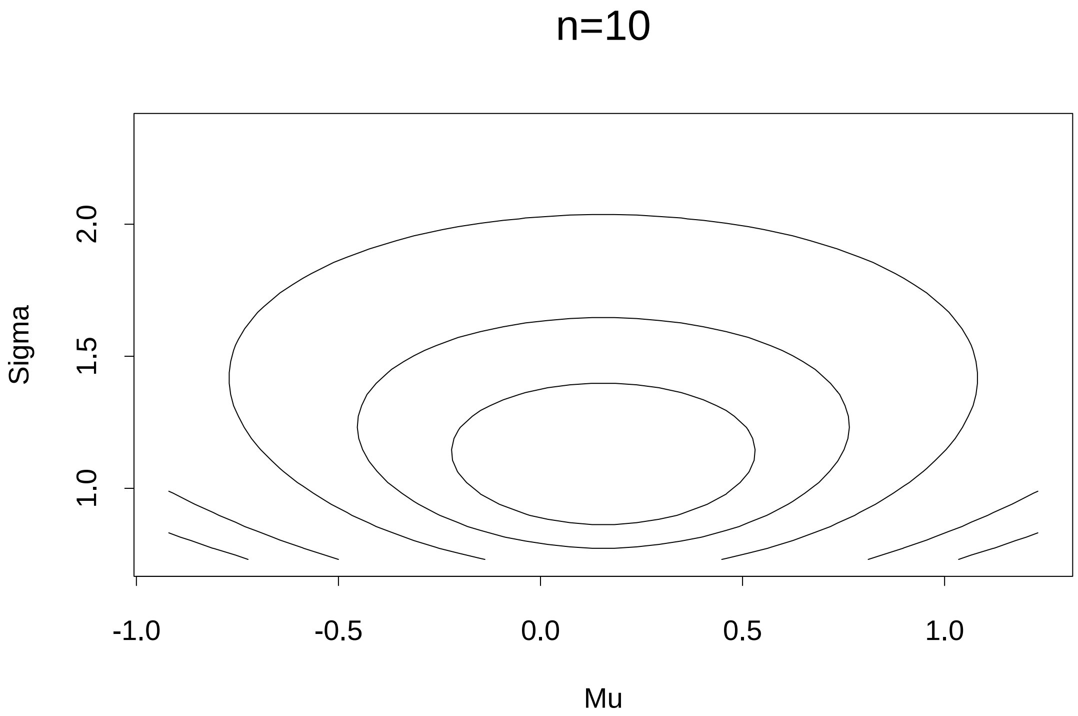
Example: $\mathbf{N}(\mu,\sigma^2)$ In this case the unique root of the likelihood equations is a global maximum.

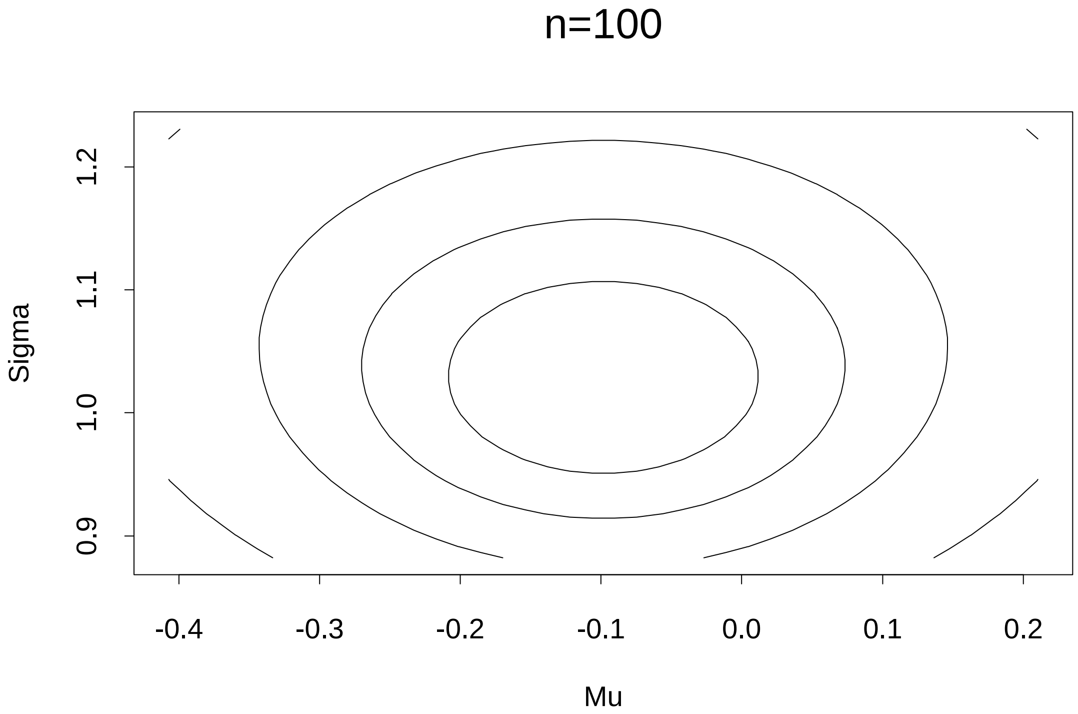
Remark: Suppose we called $\tau = \sigma^2$ the parameter. The score function still has two components: the first component is the same as before but the second component is

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


n=10
Z
0
0.2
0.4
0.6
0.8
1
Y
10
20
30
40
X
10
20
30
40


n=100
Z
0
0.2
0.4
0.6
0.8
1
Y
10
20
30
40
X
10
20
30
40


n=10
Sigma
2.0
1.5
1.0
-1.0
-0.5
0.0
0.5
1.0
Mu


n=100
Sigma
1.2
1.1
1.0
0.9
-0.4
-0.3
-0.2
-0.1
0.0
0.1
0.2
Mu

Setting the new likelihood equations equal to 0 still gives

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

This is an example of a general **invariance** (or more properly **equivariance**) principal: If $\phi = g(\theta)$ is some reparametrization of a model (a one to one relabelling of the parameter values) then $\hat{\phi} = g(\hat{\theta})$. This idea does not apply to estimators derived from other principles of estimation.]

Example: **Cauchy: location** θ

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

Example: **Binomial(**n, θ**)**

If $X = 0$ or $X = n$: no root of likelihood equations; likelihood is monotone. Other values of X: unique root, a global maximum. Global maximum at $\hat{\theta} = X/n$ even if $X = 0$ or n.

Example: **The 2 parameter exponential**

The density is

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

Log-likelihood is $-\infty$ for $\alpha > \min\{X_1, \ldots, X_n\}$ and otherwise is

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

Increasing function of α till α reaches

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

which gives mle of α. Now plug in $\hat{\alpha}$ for α; get so-called profile likelihood for β:

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

Set β derivative equal to 0 to get

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

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

Example: **Three parameter Weibull**

The density in question is

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

The log-likelihood is

$$-n\log(\beta) - \sum_{i=1}^n \left(\frac{X_i-\alpha}{\beta}\right)^\gamma + (\gamma-1)\sum_{i=1}^n \log((X_i-\alpha)/\beta).$$

There are three likelihood equations:

$$\begin{aligned}
0 &= \frac{\partial\ell}{\alpha} = \frac{\gamma}{\beta}\sum_{i=1}^n \left(\frac{X_i-\alpha}{\beta}\right)^{\gamma-1} \sum_{i=1}^n \frac{\gamma-1}{X_i-\alpha} \\
0 &= \frac{\partial\ell}{\beta} = \frac{n\gamma}{\beta} + \frac{\gamma}{\beta}\sum_{i=1}^n \left(\frac{X_i-\alpha}{\beta}\right)^{\gamma} \\
0 &= \frac{\partial\ell}{\gamma} = -\sum_{i=1}^n \log((X_i-\alpha)/\beta)\left(\frac{X_i-\alpha}{\beta}\right)^{\gamma} + \sum_{i=1}^n \log((X_i-\alpha)/\beta).
\end{aligned}$$

First set the β derivative equal to 0 and find

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

where $\hat\beta(\alpha,\gamma)$ indicates that the mle of β could be found by finding the mles of the other two parameters and then plugging into the formula above. It is not possible to find explicit formulas for the estimates of the remaining two parameters; numerical methods are needed.

However, putting $\gamma < 1$ and letting $\alpha \to X_{(1)}$ will make the log-likelihood go to ∞. As a result, the MLE is not uniquely defined: 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 $\gamma > 1$ the theory we detail below applies to the local maximum and not to the global maximum of the likelihood equations. You could look at (**?**, RALMAS3par)

7.2 Large Sample Theory for Maximum Likelihood

Large Sample Theory

We can study the approximate behaviour of $\hat\theta$ by studying the function U. Notice first that U is a sum of independent random variables and remember the law of large numbers:

Theorem 18 *If $Y_1, Y_2, \ldots$ are iid with mean μ then*

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

For the strong law of large numbers we mean

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

and for the weak law of large numbers we mean that

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

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

Now suppose θ_0 is true value of θ. Then

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

where

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

Example: $N(\mu, 1)$ data:

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

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

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

Consider first the case $\mu < \mu_0$. Then the derivative of $\ell(\mu)$ is likely to be positive so that ℓ increases as μ increases. For $\mu > \mu_0$ the derivative of ℓ is probably negative and so ℓ tends to be decreasing for $\mu > 0$. Hence: ℓ is likely to be maximized close to μ_0.

We can repeat these ideas for a more general case. To do so we study the random variable

$$\log[f(X_i,\theta)/f(X_i,\theta_0)].$$

You know the inequality

$$\mathrm{E}(X)^2 \le \mathrm{E}(X^2)$$

(the difference between the two is $\mathrm{Var}(X) \ge 0$.) This inequality admits an important generalization called Jensen's inequality:

Theorem 19 *If g is a convex function ($g'' \ge 0$ roughly) then*

$$g(\mathrm{E}(X)) \le \mathrm{E}(g(X))$$

The special case above has $g(x) = x^2$. Here we use $g(x) = -\log(x)$. This function is convex because $g''(x) = x^{-2} > 0$. We get

$$-\log(\mathrm{E}_{\theta_0}[f(X_i,\theta)/f(X_i,\theta_0)] \le \mathrm{E}_{\theta_0}[-\log\{f(X_i,\theta)/f(X_i,\theta_0)\}]$$

But

$$\begin{aligned} \mathrm{E}_{\theta_0}\left[\frac{f(X_i,\theta)}{f(X_i,\theta_0)}\right] &= \int \frac{f(x,\theta)}{f(x,\theta_0)} f(x,\theta_0)dx \\ &= \int f(x,\theta)dx \\ &= 1 \end{aligned}$$

We can reassemble the inequality and this calculation to get

$$\mathrm{E}_{\theta_0}[\log\{f(X_i,\theta)/f(X_i,\theta_0)\}] \le 0$$

In fact this inequality is strict unless the θ and θ_0 densities are actually the same.

Now let $\mu(\theta) < 0$ be this expected value. Then for each θ we find

$$\frac{\ell(\theta)-\ell(\theta_0)}{n} \frac{\sum \log[f(X_i,\theta)/f(X_i,\theta_0)]}{n} \to \mu(\theta)$$

This proves that the likelihood is probably higher at θ_0 than at any other single fixed θ. This idea can often be stretched to prove that the mle is **consistent**; to do so we need to establish **uniform** convergence in θ.

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.

Here are some more precise statements of possible conclusions. Use the following notation

$$N(\epsilon) = \{\theta : |\theta - \theta_0| \le \epsilon\}\,.$$

Suppose:

1. $\hat\theta_n$ is global maximizer of ℓ.
2. $\hat\theta_{n,\delta}$ maximizes ℓ over $N(\delta) = \{|\theta-\theta_0| \le \delta\}$.
3. $$\begin{aligned} A_\epsilon &= \{|\hat\theta_n - \theta_0| \le \epsilon\} \\ B_{\delta,\epsilon} &= \{|\hat\theta_{n,\delta} - \theta_0| \le \epsilon\} \\ C_L &= \{\exists!\theta \in N(L/n^{1/2}) : U(\theta)=0, U'(\theta) < 0\} \end{aligned}$$

Theorem 20 *1. Under unspecified conditions* **I** $P(A_\epsilon) \to 1$ *for each* $\epsilon > 0$.

2. Under unspecified conditions **II** *there is a* $\delta > 0$ *such that for all* $\epsilon > 0$ *we have* $P(B_{\delta,\epsilon}) \to 1$.

3. Under unspecified conditions **III** *for all* $\delta > 0$ *there is an* L *so large and an* n_0 *so large that for all* $n \ge n_0$, $P(C_L) > 1-\delta$.

4. Under unspecified conditions **III** *there is a sequence* L_n *tending to* ∞ *so slowly that* $P(C_{L_n}) \to 1$.

The point is that the conditions get weaker as the conclusions get weaker. There are many possible conditions in the literature. See the book by Zacks (**?**, Zacks)or some precise conditions.

7.2.1 Asymptotic Normality

Study shape of log likelihood near the true value of θ. Assume $\hat\theta$ is a root of the likelihood equations close to θ_0. Taylor expansion (1 dimensional parameter θ):

$$U(\hat\theta) = 0 = U(\theta_0) + U'(\theta_0)(\hat\theta - \theta_0) + U''(\tilde\theta)(\hat\theta - \theta_0)^2/2$$

for some $\tilde\theta$ between θ_0 and $\hat\theta$.

WARNING: This form of the remainder in Taylor's theorem is not valid for multivariate θ. Derivatives of U are sums of n terms.

So each derivative should be proportional to n in size.

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.

Ignoring second derivative term we get

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

Now look at terms U and U'.

Normal case:

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

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

Derivative is

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

Next derivative U'' is 0.

Notice: 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:

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

Notice: interchanged 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.

Differentiate identity just proved:

$$\int \frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta) dx = 0$$

Take derivative of both sides wrt θ; pull derivative under integral sign:

$$\int \frac{\partial}{\partial \theta}\left[\frac{\partial \log f}{\partial \theta}(x,\theta) f(x,\theta)\right] dx = 0$$

Do the derivative and get

$$-\int \frac{\partial^2 \log(f)}{\partial \theta^2} f(x,\theta) dx$$

$$= \int \frac{\partial \log f}{\partial \theta}(x,\theta) \frac{\partial f}{\partial \theta}(x,\theta) dx$$

$$= \int \left[\frac{\partial \log f}{\partial \theta}(x,\theta)\right]^2 f(x,\theta) dx$$

Definition: The **Fisher Information** is

$$I(\theta) = -\mathrm{E}_\theta(U'(\theta)) = n\mathrm{E}_{\theta_0}(V_1)$$

We refer to $\mathcal{I}(\theta_0) = \mathrm{E}_{\theta_0}(V_1)$ 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

$$n^{-1/2}U(\theta_0) \Rightarrow N(0,\sigma^2)$$

where $\sigma^2 = \mathrm{Var}(U_i) = \mathrm{E}(V_i) = \mathcal{I}(\theta)$.

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

$$-U'(\theta_0)/n \to \mathrm{E}_{\theta_0}[V_1] = \mathcal{I}(\theta_0)$$

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: assuming $\hat\theta = \hat\theta_n$ is a consistent root of $U(\theta) = 0$.

- $n^{-1/2}U(\theta_0) \Rightarrow MVN(0, \mathcal{I})$ where

  $$\mathcal{I}_{ij} = \mathrm{E}_{\theta_0}\left\{V_{1,ij}(\theta_0)\right\}$$

  and

  $$V_{k,ij}(\theta) = -\frac{\partial^2 \log f(X_k, \theta)}{\partial\theta_i \partial\theta_j}$$

- If $\mathbf{V}_k(\theta)$ is the matrix $[V_{k,ij}]$ then

  $$\frac{\sum_{k=1}^n \mathbf{V}_k(\theta_0)}{n} \to \mathcal{I}$$

- If $\mathbf{V}(\theta) = \sum_k \mathbf{V}_k(\theta)$ then

  $$\{\mathbf{V}(\theta_0)/n\} n^{1/2}(\hat\theta - \theta_0) - n^{-1/2}U(\theta_0) \to 0$$

  in probability as $n \to \infty$.

- Also

  $$\{\mathbf{V}(\hat\theta)/n\} n^{1/2}(\hat\theta - \theta_0) - n^{-1/2}U(\theta_0) \to 0$$

  in probability as $n \to \infty$.

- $n^{1/2}(\hat\theta - \theta_0) - \{\mathcal{I}(\theta_0)\}^{-1}U(\theta_0) \to 0$ in probability as $n \to \infty$.

- $n^{1/2}(\hat\theta - \theta_0) \Rightarrow MVN(0, \mathcal{I}^{-1})$.

- 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 all these different forms? Use limit laws to test hypotheses and compute confidence intervals. Test $H_o : \theta = \theta_0$ using one of the 4 quantities as test statistic. Find confidence intervals using quantities as *pivots*. E.g.: second and fourth limits 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

$$\left| \frac{\sqrt{2n}(\hat{\sigma} - \sigma)}{\sigma} \right| \le z_{\alpha/2}$$

Similar interval can be derived from 3rd expression, though this is much more complicated.

Usual summary: mle is consistent and asymptotically normal with an asymptotic variance which is the inverse of the Fisher information.

Problems with maximum likelihood

1. Many parameters lead to poor approximations. MLEs can be far from right answer. See homework for Neyman Scott example where MLE is not consistent.

2. Multiple roots of the likelihood equations: you must choose the right root. Start with different, consistent, estimator; apply iterative scheme like Newton Raphson to likelihood equations to find MLE. Not many steps of NR generally required if 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. Remember the definitions:

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

$$\mathrm{E}(X^r)$$

Definition: (**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.

Example: The **Gamma** model: 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

$$\begin{aligned} \alpha\beta &= \overline{X} \\ \alpha(\alpha+1)\beta^2 &= \overline{X^2} \end{aligned}$$

or

$$\begin{aligned} \alpha\beta &= \overline{X} \\ \alpha\beta^2 &= \overline{X^2} - \overline{X}^2. \end{aligned}$$

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

$$\tilde{\alpha} = \overline{X}/\tilde{\beta} = (\overline{X})^2/(\overline{X^2} - \overline{X}^2) .$$

The method of moments 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.

Score function 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.

Estimating Equations

Same large sample ideas arise whenever estimates derived by solving some equation.

Example: large sample theory for **Generalized Linear Models**.

Suppose Y_i is number of cancer cases in some group of people characterized by values x_i of some covariates.

Think of x_i as containing variables like age, or a dummy for sex or average income or . . ..

Possible parametric regression model: Y_i has a Poisson distribution with mean μ_i where the mean μ_i depends somehow on x_i.

Typically assume $g(\mu_i) = \beta_0 + x_i\beta$; g is **link** function.

Often $g(\mu) = \log(\mu)$ and $x_i\beta$ is a matrix product: x_i row vector, β column vector.

"Linear regression model with Poisson errors".

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 $\log(\mathrm{E}(Y_i)) = x_i\beta$. Then we have assumed that

$$\mathrm{E}_\beta(U(\beta)) = 0$$

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

$$\mathrm{Var}(U) = \sum x_i^2 \mathrm{Var}(Y_i)\,.$$

If Y_i is Poisson with mean μ_i (and so $\mathrm{Var}(Y_i) = \mu_i$) this is

$$\mathrm{Var}(U) = \sum x_i^2 \mu_i\,.$$

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 $\sigma_U^2 = \sum x_i^2 \mathrm{Var}(Y_i)$ and so

$$\hat\beta - \beta \approx N(0, \sigma_U^2/(\sum x_i^2\mu_i)^2)$$

If $\mathrm{Var}(Y_i) = \mu_i$, as it is for the Poisson case, the asymptotic variance simplifies to $1/\sum x_i^2\mu_i$.

Other estimating equations are possible, popular. If w_i is any set of deterministic weights (possibly depending on μ_i) then 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.

Idea widely used:

Example: Generalized Estimating Equations, Zeger and Liang.

Abbreviation: GEE.

Called by econometricians Generalized Method of Moments.

An estimating equation is unbiased if

$$\mathrm{E}_\theta(U(\theta)) = 0$$

Theorem 21 *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

$$\sqrt{C(\theta_0)}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$
$$\sqrt{C(\hat{\theta})}(\hat{\theta} - \theta_0) \Rightarrow N(0,1)$$

Other ways to estimate A, B and C lead to the same conclusions. There are multivariate extensions using matrix square roots.

Chapter 8

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 $\Theta_0 \cap \Theta_1 = \emptyset$.

A rule for making the required choice can be described in two ways:

1. In terms of the set

   $$R = \{X : \text{we choose } \Theta_1 \text{ if we observe } X\}$$

   called the *rejection* or *critical* region of the test.

2. 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 $R_\phi = \{x : \phi(x) = 1\}$.

Neyman Pearson approach treats two hypotheses asymmetrically. Hypothesis H_o referred to as the *null* hypothesis (traditionally 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 might be interested 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.

8.0.2 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.

Jerzy Neyman and Egon Pearson (Egon's father Karl Pearson was also a famous statistician) invented the jargon which surrounds their philosophy of hypothesis testing. Unfortunately much of the jargon is lame:

Definition: **Type I error** is the error made when $\theta = \theta_0$ but we choose H_1, that is, $X \in R_\phi$.

Definition: **Type II error** is the error made when $\theta = \theta_1$ but we choose H_0.

Definition: The **level** of a simple versus simple test is

$$\alpha = P_{\theta_0}(\text{We make a Type I error})$$

or

$$\alpha = P_{\theta_0}(X \in R_\phi) = E_{\theta_0}(\phi(X))$$

The other error probability, denoted β, is

$$\beta = P_{\theta_1}(X \not\in R_\phi) = E_{\theta_1}(1 - \phi(X)).$$

To illustrate a general strategy I now minimize $\alpha + \beta$, the total error probability, which is given by

$$\begin{aligned} \alpha + \beta = & \quad E_{\theta_0}(\phi(X)) + E_{\theta_1}(1 - \phi(X)) \\ = & \quad \int [\phi(x) f_0(x) + (1 - \phi(x)) f_1(x)] dx \end{aligned}$$

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

$$\phi(x) f_0(x) + (1 - \phi(x)) f_1(x)$$

is between $f_0(x)$ and $f_1(x)$. 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 the inequalities to express our condition in terms of the **likelihood ratio** $f_1(x)/f_0(x)$.

Theorem 22 *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 β subject to the constraint $\alpha \le \alpha_0$. Usually this is really equivalent to the constraint $\alpha = \alpha_0$ (because if you use $\alpha < \alpha_0$ you could make the rejection region R larger and keep $\alpha \le \alpha_0$ 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 $p = p_1 = 3/4$. (It might be possible to conjure up some genetics problem in which this was vaguely realistic but I think it would be a stretch.)

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. For example, try to get $\alpha_0 = 0.05$ with $n = 5$. The possible values of α are $0, 1/32 = 0.03125, 2/32 = 0.0625$, etc. Here are all the rejection regions which are possible for $\alpha_0 = 0.05$:

Region	α	β
$R_1 = \emptyset$	0	1
$R_2 = \{x = 0\}$	0.03125	$1 - (1/4)^5$
$R_3 = \{x = 5\}$	0.03125	$1 - (3/4)^5$

So R_3 minimizes β subject to $\alpha < 0.05$.

Now raise α_0 slightly to 0.0625; the possible rejection regions are R_1, R_2, R_3 and $R_4 = R_2 \cup R_3$. The first three have the same α and β as before while R_4 has $\alpha = \alpha_0 = 0.0625$ and $\beta = 1 - (3/4)^5 - (1/4)^5$. Thus R_4 is the best rejection region!

The problem is that if all trials are failures this "optimal" R chooses $p = 3/4$ rather than $p = 1/2$. But $p = 1/2$ makes 5 failures much more likely than $p = 3/4$ so it seems clear there must be a flaw in the theory; R_4 cannot really be the optimal way of doing hypothesis testing.

The real problem is discreteness. Here is a solution to the problem: Expand the set of possible values of ϕ to $[0, 1]$. 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 actually 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 $P_{\theta_0}(X \ge 5)$ or $P_{\theta_0}(X \ge 4)$ or ....

Example: For a smaller example I consider the case of $n = 3$ so that the random variable X has 4 possible values; there are then 2^4 possible rejection regions (subsets of $\{0, 1, 2, 3\}$). Here is a table of the levels for each possible rejection region R:

R	α
$\emptyset$	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
{0,1,2}, {1,2,3}	7/8
{0,1,2,3}	1

The best level 2/8 test has rejection region $\{0, 3\}$, $\beta = 1 - [(3/4)^3 + (1/4)^3] = 36/64$. The best level 2/8 test using randomization rejects when $X = 3$ and, when $X = 2$ tosses a coin with $P(H) = 1/3$, then rejects if you get H. The level of this randomized test is $1/8 + (1/3)(3/8) = 2/8$; 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

$$\begin{aligned}\pi(\theta) &= E_\theta(\phi(X)) \\ \alpha &= E_0(\phi(X)) \quad \text{and} \\ \beta &= 1 - E_1(\phi(X))\end{aligned}$$

Note that a test using a rejection region C is equivalent to

$$\phi(x) = 1(x \in C)$$

Theorem 23 (The Neyman Pearson Lemma) *In testing f_0 against f_1 the probability β of a type II error is minimized, subject to $\alpha \le \alpha_0$ 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

$$P_0(\frac{f_1(X)}{f_0(X)} \ge \lambda) \ge \alpha_0$$

and

$$P_0(\frac{f_1(X)}{f_0(X)} \le \lambda) \ge 1 - \alpha_0$$

and where γ is any number chosen so that

$$\begin{aligned} E_0(\phi(X)) = P_0(\frac{f_1(X)}{f_0(X)} > \lambda) \\ + \quad \gamma P_0(\frac{f_1(X)}{f_0(X)} = \lambda) \\ = \alpha_0 \end{aligned}$$

The value of γ is unique if $P_0(\frac{f_1(X)}{f_0(X)} = \lambda) > 0$.

Example: Consider again the Binomial(n, p) problem with $p_0 = 1/2$ and $p_1 = 3/4$. The ratio f_1/f_0 is

$$3^x 2^{-n}.$$

If $n = 5$ this ratio is one of the numbers 1, 3, 9, 27, 81, 243 divided by 32.

Suppose we have $\alpha = 0.05$. Then λ must be one of the possible values of f_1/f_0. If we try $\lambda = 243/32$ then

$$\begin{aligned} P_0(3^X 2^{-5} \geq 243/32) &= P_0(X = 5) \\ &= 1/32 < 0.05 \end{aligned}$$

and

$$\begin{aligned} P_0(3^X 2^{-5} \geq 81/32) &= P_0(X \geq 4) \\ &= 6/32 > 0.05 \end{aligned}$$

So $\lambda = 81/32$. 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 corresponding to $\gamma = 0$ (or $\gamma = 1$). When the sample size is large you can come very close to any desired α_0 with a *non-randomized test*, that is, a test for which the function ϕ takes no values other than 0 or 1.

In our example, if $\alpha_0 = 6/32$ then we can either take λ to be 243/32 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

$$\frac{f_1(X_1,\ldots,X_n)}{f_0(X_1,\ldots,X_n)} = \exp\{\mu_1\sum X_i - n\mu_1^2/2 - \mu_0\sum X_i + n\mu_0^2/2\}$$

which simplifies to

$$\exp\{\mu_1\sum X_i - n\mu_1^2/2\}$$

Now choose λ so that

$$P_0(\exp\{\mu_1\sum X_i - n\mu_1^2/2\} > \lambda) = \alpha_0$$

Can make it equal because $f_1(X)/f_0(X)$ has a continuous distribution. Rewrite probability as

$$P_0(\sum X_i > [\log(\lambda) + n\mu_1^2/2]/\mu_1) = 1-\Phi\left(\frac{\log(\lambda)+n\mu_1^2/2}{n^{1/2}\mu_1}\right)$$

Let z_α be the upper α critical point of $N(0,1)$; then

$$z_{\alpha_0} = [\log(\lambda) + n\mu_1^2/2]/[n^{1/2}\mu_1]\,.$$

Solve this equation 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

1. ϕ^* has level $\alpha \le \alpha_o$

2. If ϕ has level $\alpha\le\alpha_0$ then for every $\theta\in\Theta_1$ we have

$$E_\theta(\phi(X)) \le E_\theta(\phi^*(X))$$

Proof of 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.

8.0.3 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

$$f(x_\lambda) \le f(x) + \lambda[g(x) - g(x_\lambda)]$$

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

$$f(x_\lambda) \le f(x) + \lambda g(x)$$

and for any x satisfying the constraint $g(x) = 0$ we have

$$f(x_\lambda) \le f(x)$$

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 $g(x_\lambda) = 0$.

8.0.4 Return to proof of NP lemma

For each $\lambda > 0$ we have seen that ϕ_λ minimizes $\lambda\alpha + \beta$ where $\phi_\lambda = 1(f_1(x)/f_0(x) \ge \lambda)$.

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 $lambda_0\alpha + \beta$. Suppose ϕ^* is some other test with level $\alpha^* \le \alpha_0$. Then

$$\lambda_0\alpha_\phi + \beta_\phi \le \lambda_0\alpha_{\phi^*} + \beta_{\phi^*}$$

We can rearrange this as

$$\beta_{\phi^*} \ge \beta_\phi + (\alpha_\phi - \alpha_{\phi^*})\lambda_0$$

Since

$$\alpha_{\phi^*} \le \alpha_0 = \alpha_\phi$$

the second term is non-negative and

$$\beta_{\phi^*} \ge \beta_\phi$$

which proves the Neyman Pearson Lemma.

Example application of NP: Again consider the Binomial(n,p) problem. In order to test $p = p_0$ versus p_1 for a $p_1 > p_0$ the NP test is of the form

$$\phi(x) = 1(X > k) + \gamma 1(X = k)$$

where we choose k so that

$$P_{p_0}(X > k) \le \alpha_0 < P_{p_0}(X \ge k)$$

and $\gamma \in [0,1)$ so that

$$\alpha_0 = P_{p_0}(X > k) + \gamma P_{p_0}(X = k)$$

This rejection region depends only on p_0 and not on p_1 so that this test is UMP for $p = p_0$ against $p > p_0$. Since this test has level α_0 even for the larger null hypothesis $p \le p_0$, it is also UMP for $p \le p_0$ against $p > p_0$.

Application of the NP lemma: In the $N(\mu,1)$ model consider $\Theta_1 = \{\mu > 0\}$ and $\Theta_0 = \{0\}$ or $\Theta_0 = \{\mu \le 0\}$. 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 $\mu \le 0$ we have

$$\begin{aligned} P_\mu(n^{1/2}\bar{X} &> z_{\alpha_0}) \\ &= P_\mu(n^{1/2}(\bar{X} - \mu) > z_{\alpha_0} - n^{1/2}\mu) \\ &= P(N(0,1) > z_{\alpha_0} - n^{1/2}\mu) \\ &\le P(N(0,1) > z_{\alpha_0}) \\ &= \alpha_0 \end{aligned}$$

(Notice the use of $\mu \le 0$. 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

$$E_0(\phi(X_1, \ldots, X_n)) \le \alpha_0$$

Fix a $\mu > 0$. According to the NP lemma

$$E_\mu(\phi(X_1, \ldots, X_n)) \le E_\mu(\phi_\mu(X_1, \ldots, X_n))$$

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 is a fairly general phenomenon: 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 24 *For a monotone likelihood ratio family the Uniformly Most Powerful level α test of $\theta \le \theta_0$ (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

$$P_{\theta_0}(T(X) > t_\alpha) + \gamma P_{\theta_0}(T(X) = t_\alpha) = \alpha_0 \,.$$

A typical family where this works is a one parameter exponential family. Usually there is no UMP test.

Example: test $\mu = \mu_0$ against the two sided alternative $\mu \neq \mu_0$ in the $N(\mu, 1)$ model. There is no UMP level α test.

If there were such a test 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.

Everybody's favourite test is the usual 2 sided z-test which rejects for large values of $|\bar{X} - \mu_0|$. This test maximizes power subject to two constraints: first, that the test have level α; second, that the power function is minimized at $\mu = \mu_0$. The second condition means that the power on alternative is larger than the power on the null.

8.1 Likelihood ratio tests

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: Consider a sample of size n from the $N(\mu, 1)$ model and test $\mu \le 0$ against $\mu > 0$. (Remember the uniformly most powerful test.) The log-likelihood is

$$-n(\bar{X} - \mu)^2/2$$

If $\bar{X} > 0$ then the global maximum in Θ_1 at $\bar{X}$. If $\bar{X} \le 0$ the global maximum in Θ_1 is at 0. Thus $\hat{\mu}_1$ which maximizes $\ell(\mu)$ subject to $\mu > 0$ is $\bar{X}$ if $\bar{X} > 0$ and 0 if $\bar{X} \le 0$. Similarly, $\hat{\mu}_0$ is $\bar{X}$ if $\bar{X} \le 0$ and 0 if $\bar{X} > 0$. Hence

$$\frac{f_{\hat{\theta}_1}(X)}{f_{\hat{\theta}_0}(X)} = \exp\{\ell(\hat{\mu}_1) - \ell(\hat{\mu}_0)\}$$

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 level α we must reject if $n^{1/2}\bar{X} > z_\alpha$. Notice that a simpler statistic with the same rejection region is the *log-likelihood ratio*

$$\lambda \equiv 2\log\left(\frac{f_{\hat{\mu}_1}(X)}{f_{\hat{\mu}_0}(X)}\right) = n\bar{X}|\bar{X}|$$

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 $\mu \ne 0$ 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 two sided t-test.

Example 3: For the $N(\mu, \sigma^2)$ problem testing $\mu = 0$ against $\mu \ne 0$ 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

$$\ell(\hat{\mu}, \hat{\sigma}^2) = -n/2 - n\log(\hat{\sigma})$$

First we maximize ℓ over the null hypothesis. Recall that

$$\ell(\mu, \sigma) = -\frac{1}{2\sigma^2}\sum(X_i - \mu)^2 - n\log(\sigma)$$

On the null $\mu = 0$ so find we $\hat{\sigma}_0$ 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. Thus the likelihood ratio test rejects for large values of $|t|$ — 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

$$\left[\begin{array}{cc} \mathcal{I}_{\phi\phi} & \mathcal{I}_{\phi\gamma} \\ \mathcal{I}_{\gamma\phi} & \mathcal{I}_{\gamma\gamma} \end{array}\right].$$

According to our large sample theory for the mle we have

$$\hat\theta \approx \theta + \mathcal{I}^{-1}U$$

and

$$\hat{\gamma}_0 \approx \gamma_0 + \mathcal{I}_{\gamma\gamma}^{-1} U_\gamma$$

If you carry out a two term Taylor expansion of both $\ell(\hat{\theta})$ and $\ell(\hat{\theta}_0)$ around θ_0 you get

$$\ell(\hat{\theta}) \approx \ell(\theta_0) + U^t \mathcal{I}^{-1} U + \frac{1}{2} U^t \mathcal{I}^{-1} V(\theta) \mathcal{I}^{-1} U$$

where V is the second derivative matrix of ℓ. Remember that $V \approx -\mathcal{I}$ and you get

$$2[\ell(\hat{\theta}) - \ell(\theta_0)] \approx U^t \mathcal{I}^{-1} U\,.$$

A similar expansion for $\hat{\theta}_0$ gives

$$2[\ell(\hat{\theta}_0) - \ell(\theta_0)] \approx U_\gamma^t \mathcal{I}_{\gamma\gamma}^{-1} U_\gamma\,.$$

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 25 *The log-likelihood ratio statistic*

$$\lambda = 2[\ell(\hat{\theta}) - \ell(\hat{\theta}_0)]$$

has, under the null hypothesis, approximately a χ_p^2 distribution.

Aside:

Theorem 26 *Suppose $X \sim MVN(0, \Sigma)$ 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_ν^2 distribution with df $\nu = trace(M\Sigma)$.*

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 $AA^t = \Sigma$ condition in the theorem is

$$AQQA^t = AQA^t$$

Since Σ is non-singular so is A. Multiply by A^{-1} on the left and by $(A^t)^{-1}$ on the right to get the identity $QQ = Q$.

The matrix Q is symmetric so $Q = P \Lambda P^t$ where Λ is a diagonal matrix containing the eigenvalues of Q and P is orthogonal matrix whose columns are the corresponding orthonormal eigenvectors. So rewrite

$$Z^t Q Z = (P^t Z)^t \Lambda (PZ)\,.$$

Notice that $W = P^t Z$ is $MVN(0, P^t P = I)$; i.e. 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

$$J = \left[\begin{array}{cc} 0 & 0 \\ 0 & \mathcal{I}_{\gamma\gamma}^{-1} \end{array} \right]$$

It is easy to check that $M\Sigma$ becomes

$$\left[\begin{array}{cc} I & 0 \\ -\mathcal{I}_{\gamma\phi}\mathcal{I}_{\phi\phi} & 0 \end{array} \right]$$

where I is a $p \times p$ identity matrix. It follows that $\Sigma M \Sigma M \Sigma = \Sigma M \Sigma$ and $trace(M\Sigma) = p$.

8.2 Optimal Unbiased Tests

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. If π is differentiable then this will imply

$$\pi'(\mu_0) = 0$$

Example: Consider data $X = (X_1, \ldots, X_n)$, a sample from the $N(\mu, 1)$ distribution. If ϕ is any test function then

$$\pi'(\mu) = \frac{\partial}{\partial\mu} \int \phi(x) f(x, \mu) dx$$

Differentiate under the integral and use

$$\frac{\partial f(x,\mu)}{\partial\mu} = \sum (x_i - \mu) f(x, \mu)$$

to get the condition

$$\int \phi(x) \bar{x} f(x, \mu_0) dx = \mu_0 \alpha_0$$

We now minimize $\beta(\mu)$ subject to the two constraints

$$E_{\mu_0}(\phi(X)) = \alpha_0$$

and

$$E_{\mu_0}(\bar{X}\phi(X)) = \mu_0 \alpha_0.$$

As in the proof of the Neyman-Pearson Lemma we use Lagrange multipliers. With two constraints we have two multipliers. So 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 in the proof of the Neyman-Pearson Lemma 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,

$$\lambda_1 \alpha + \lambda_2 E_{\mu_0}[(\bar{X} - \mu_0)\phi(X)] + \beta$$

is minimized by a rejection region of the form

$$\{\bar{X} > K_U\} \cup \{\bar{X} < K_L\}$$

Now we use the constraints of level α and unbiasedness to find K_U and K_L. That is, to satisfy the constraints we adjust K_U and K_L to get level α and $\pi'(\mu_0) = 0$. The second of these conditions shows that the rejection region is symmetric about μ_0. So the test rejects for

$$\sqrt{n}|\bar{X} - \mu_0| > z_{\alpha/2}.$$

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.

8.2.1 Lagrange Multipliers

In this little subsection I want to review the scope of the Lagrange multipliers tactic. Suppose we want to maximize some function $f(x)$ subject to some constraint like $h_1(x) = 0, \ldots, h_p(x) = 0$ over all x belonging to some set A. Fix multipliers $\lambda_1, \ldots, \lambda_p$ and suppose x^* maximizes

$$G_\lambda(x) = f(x) - \sum_1^p \lambda_i h_i(x)$$

Suppose that x^* satisfies the constraints, that is, $h_i(x^*) = 0$ for each $1 \le i \le p$. Then x^* maximizes $f(x)$ subject to the constraints.

Proof: If we have any other x which satisfies the constraints then

$$\begin{aligned} f(x) &= f(x) - \sum_i \lambda_i h_i(x) \\ &= G_\lambda(x) \\ &\le G_\lambda(x^*) \\ &= f(x^*) - \sum_i \lambda_i h_i(x^*) \\ &= f(x^*) \end{aligned}$$

In the Neyman Pearson lemma and the result in the previous sub-section the constraint that the test have level α is an inequality constraint and the result above assumes equality constraints. If some constraint, say constraint i is an inequality $h_i(x) \le 0$, and if the special point x^* actually satisfies the equality $h_i(x^*) = 0$ and if the value of λ_i is non-negative then the chain of inequalities in the proof above remains correct except that the very first line becomes

$$f(x) \le f(x) - \sum_i \lambda_i h_i(x).$$

The conclusion remains the same.

In the Neyman Pearson Lemma case the role of x is played by the test function ϕ and the function f is the power corresponding to the test ϕ. The optimal level α test described in the lemma satisfies the equality constraint and the resulting λ is non-negative so this more general argument is relevant.

8.2.2 Uniformly Most Powerful Unbiased Tests

Definition: A test ϕ^* is a Uniformly Most Powerful Unbiased (UMPU) level α_0 test if

1. ϕ^* has level $\alpha \le \alpha_0$.

2. ϕ^* is unbiased.

3. If ϕ has level $\alpha \le \alpha_0$ and ϕ is unbiased then for every $\theta \in \Theta_1$ we have

$$E_\theta(\phi(X)) \le E_\theta(\phi^*(X))$$

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 $\mu \ne \mu_0$.

8.2.3 Nuisance Parameters

In this section I will show that the usual t-test is UMPU. The conclusion applies to both the one-sided and two-sided tests.

Suppose $X_1, \ldots, X_n$ iid $N(\mu, \sigma^2)$. Consider the problem of testing $\mu = \mu_0$ or $\mu \le \mu_0$ against $\mu > \mu_0$. The parameter space is two dimensional; the boundary between the null and alternative is

$$\{(\mu, \sigma); \mu = \mu_0, \sigma > 0\}$$

If a test has $\pi(\mu, \sigma) \le \alpha$ for all $\mu \le \mu_0$ and $\pi(\mu, \sigma) \ge \alpha$ for all $\mu > \mu_0$ then $\pi(\mu_0, \sigma) = \alpha$ for all σ because the power function of any test must be continuous. (This assertion uses the dominated convergence theorem; the power function is an integral and the assertion of continuity amounts to taking a limit inside the integral.)

Think of $\{(\mu, \sigma); \mu = \mu_0\}$ as the parameter space for a model (it is a *submodel* of our original model). For this parameter space

$$S = \sum (X_i - \mu_0)^2$$

is complete and sufficient. Remember that the definitions of both completeness and sufficiency depend on the parameter space.

Now 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 guarantees that

$$g(S) \equiv \alpha.$$

Now let us fix a single value of σ and a value $\mu_1 > \mu_0$. 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

$$f_{\bar{X}|s}(t|s) = h(s, t) \exp\{\theta_1 t + c^*(\theta_1, s)\}$$

Note the disappearance of θ_2. Also note that the null is $\theta_1 = 0$. This permits application of the NP lemma to the conditional family to prove that UMP unbiased test has form

$$\phi(\bar{X}, S) = 1(\bar{X} > K(S))$$

where $K(S)$ is chosen to make the *conditional* level exactly 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.

8.2.4 Summary commentary on optimal tests

- A good test has $\pi(\theta)$ large on the alternative and small on the null.
- For one sided one parameter families with monotone likelihood ratio a UMP test exists.
- For two sided or multiparameter families the best to be hoped for is UMP Unbiased or Invariant or Similar. I have not described "Invariant" or "Similar"; if you want to see them consult the bible of testing, (**?**).
- Good tests are found as follows:

  1. Use the NP lemma to determine a good rejection region for a simple alternative.
  2. Try to express that region in terms of a statistic whose definition does not depend on the specific alternative.
  3. If this fails impose an additional criterion such as unbiasedness. Then mimic the NP lemma and again try to simplify the rejection region.

Chapter 9

Confidence Sets

Definition: 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

$$P_\theta(\phi(\theta) \in C) \ge \beta$$

Confidence sets are very closely connected with hypothesis tests:

From confidence sets to tests

Suppose C is a level $\beta = 1 - \alpha$ confidence set for ϕ. Then we may convert C to a family of hypothesis tests. To test $\phi = \phi_0$: reject if $\phi_0 \not\in C$. This test has level α.

From tests to confidence sets

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}. Define $C = \{\phi_0 : \phi = \phi_0 \text{ is not rejected}\}$; this set C is a level $1 - \alpha$ confidence set for ϕ.

Example: The usual t test gives rise in this way to the usual t confidence intervals

$$\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}.$$

Conversely μ_0 is in the usual confidence interval if and only if the t-statistic for testing $\mu = \mu_0$ is smaller than the corresponding t critical value.

Confidence sets from Pivots

Definition: A **pivot** (or pivotal quantity) is a function $g(\theta, X)$ whose distribution is the same for all θ. (The θ in pivot is same θ as being used to calculate the distribution of $g(\theta, X)$.

We can use pivots 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 data X solve the relation

$$g(\theta, X) \in A$$

to get

$$\theta \in C(X, A)\,.$$

Then $C(X, A)$ is a level β confidence set.

Example: In the $N(\mu, \sigma^2)$ model the quantity $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$ is a pivot. It leads to confidence intervals for σ as follows. Given $\beta = 1-\alpha$ consider the two points

$$\chi^2_{n-1,1-\alpha/2} \text{ and } \chi^2_{n-1,\alpha/2}.$$

Then

$$P(\chi^2_{n-1,1-\alpha/2} \le (n-1)s^2/\sigma^2 \le \chi^2_{n-1,\alpha/2}) = \beta$$

for all μ, σ. Now solve this relation to get a set of values for σ:

$$P(\frac{(n-1)^{1/2}s}{\chi_{n-1,\alpha/2}} \le \sigma \le \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}) = \beta;$$

thus the interval

$$\left[\frac{(n-1)^{1/2}s}{\chi_{n-1,\alpha/2}}, \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha/2}}\right]$$

is a level $\beta = 1-\alpha$ confidence interval.

In the same model we also have

$$P(\chi^2_{n-1,1-\alpha} \le (n-1)s^2/\sigma^2) = \beta$$

which can be solved to get

$$P(\sigma \le \frac{(n-1)^{1/2}s}{\chi_{n-1,1-\alpha}}) = \beta$$

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

$$(n-1)^{1/2}s/\chi_{n-1,\alpha_1} \text{ to } (n-1)^{1/2}s/\chi_{n-1,1-\alpha_2}$$

has level $\beta = 1-\alpha_1-\alpha_2$. For fixed β it is possible to minimize the length of the resulting interval numerically — this procedure is rarely used. See the homework for an example.

Chapter 10

Decision Theory and Bayesian Methods

Decision Theory and Bayesian Methods

Example: Decide between 4 modes of transportation to work:

- B = Ride my bike.
- C = Take the car.
- T = Use public transit.
- H = Stay home.

Costs depend on weather: R = Rain or S = Sun.

10.0.5 Ingredients of Decision Problem in the no data case

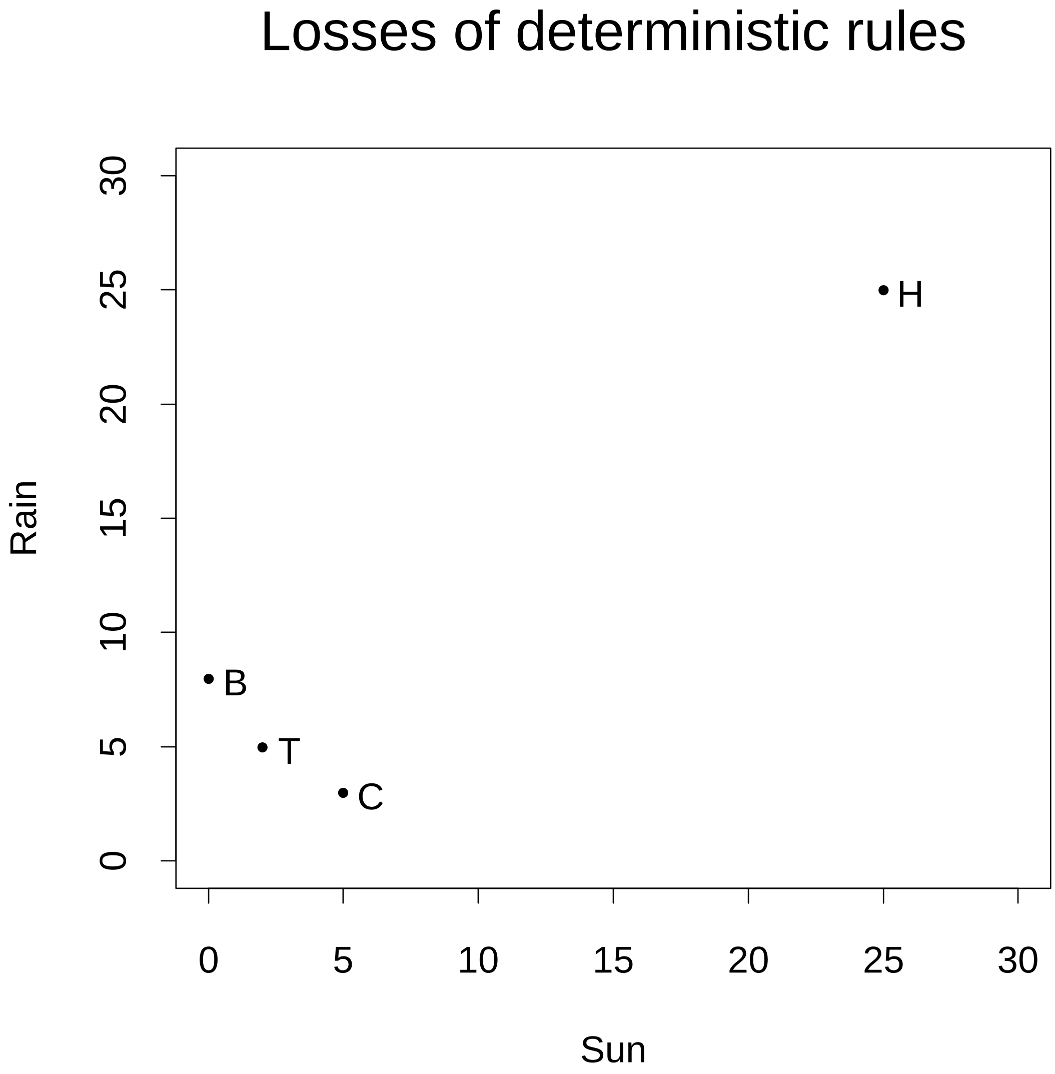
- Decision space $D = \{B, C, T, H\}$ of possible actions.
- Parameter space $\Theta = \{R, S\}$ of possible "states of nature".
- Loss function $L = L(d, \theta)$ loss incurred if do d and θ is true state of nature.

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 sup and inf 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.

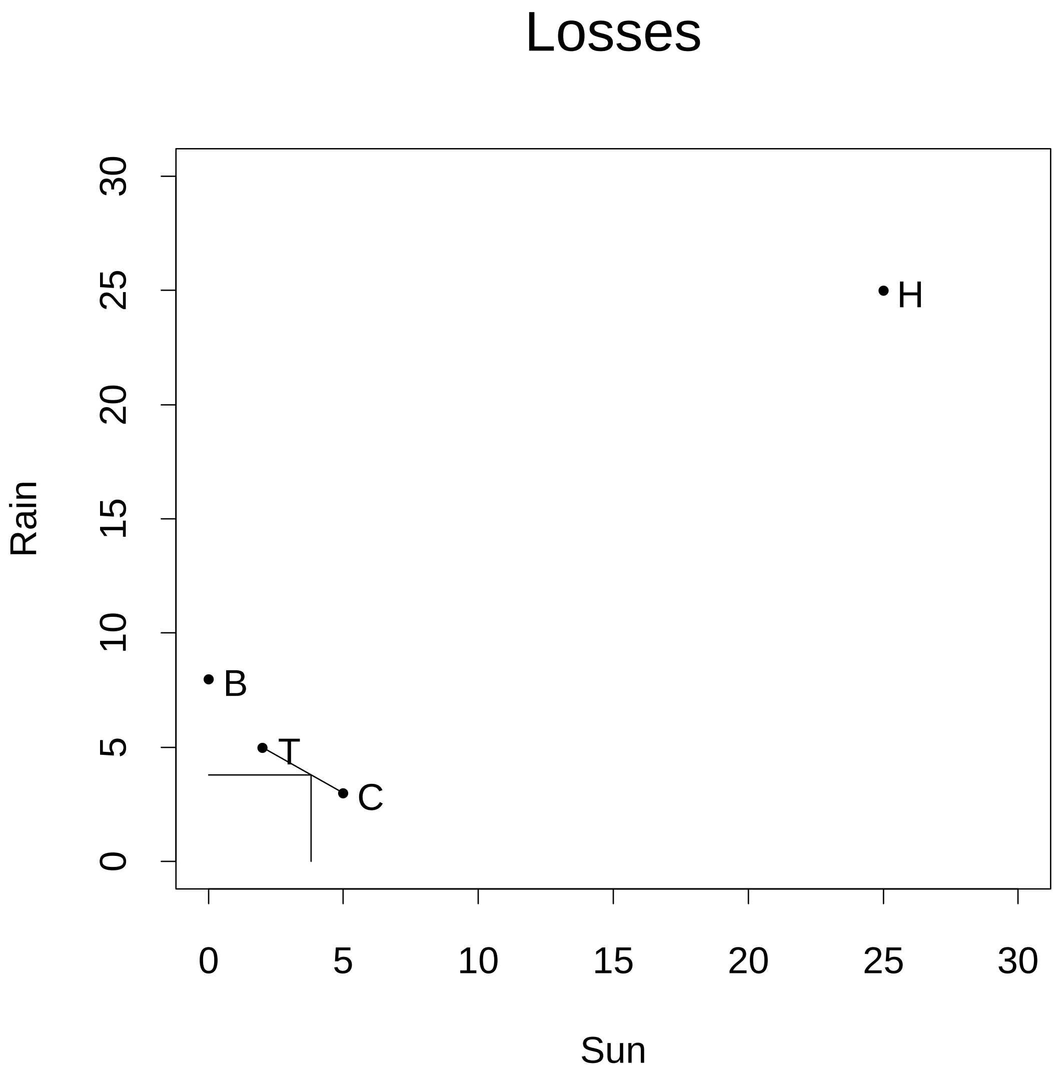
Example: My transportation problem has no data so the only possible (non-randomized) decisions are the four possible actions B, C, T, H. For B and T the worst case is rain. For the other two actions Rain and Sun are equivalent. We have the following table:

	C	B	T	H
R	3	8	5	25
S	5	0	2	25
Maximum	5	8	5	25

To get the smallest maximum: take car, or transit. Thus the minimax action is either to take the car or to take public transit.

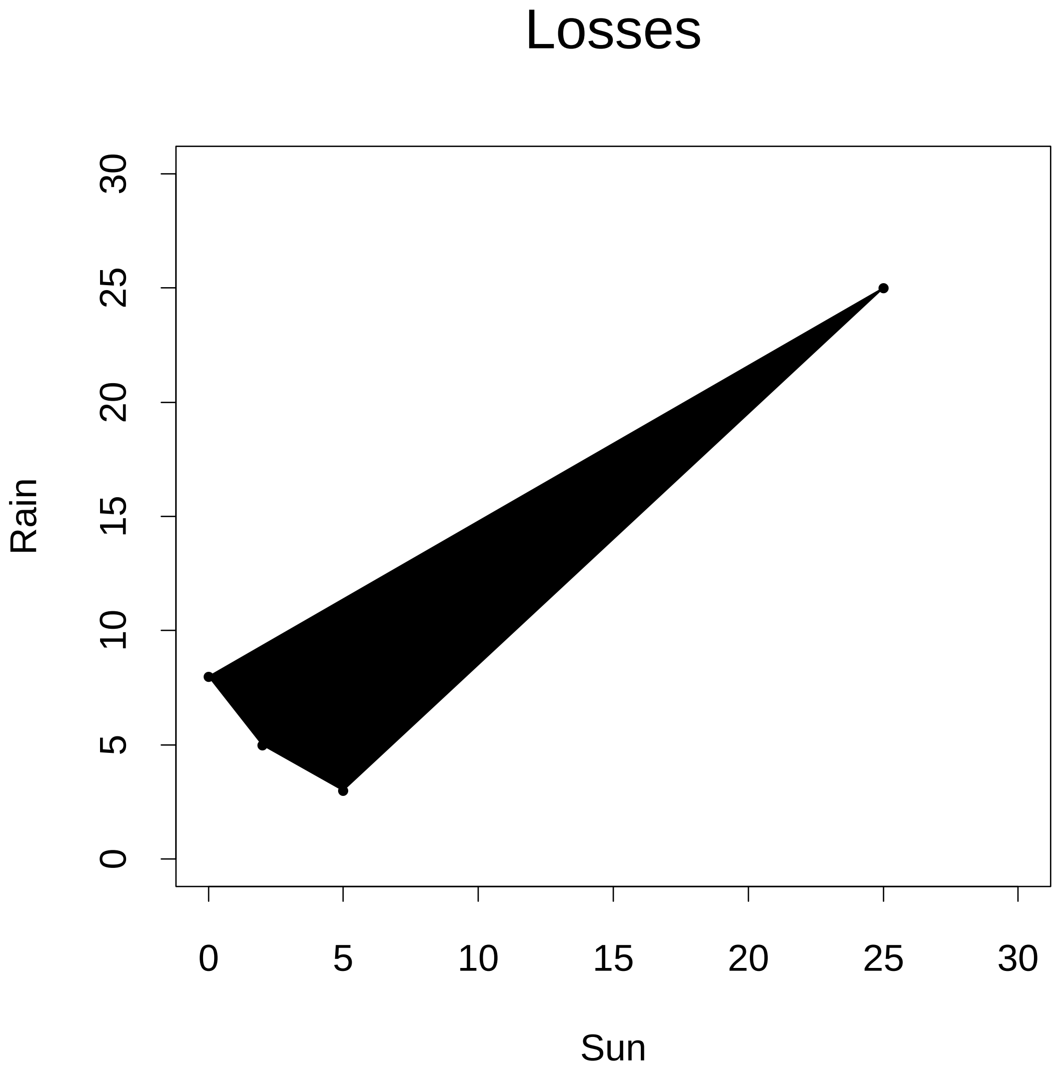
Now imagine I toss a coin with probability λ of getting Heads and take my car if I get Heads, otherwise take transit. The long run average daily loss would be $3\lambda + 5(1 - \lambda)$ when it rains and $5\lambda + 2(1 - \lambda)$ when it is Sunny. Call this procedure d_λ; add it to graph for each value of λ. 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 worst case risk is for sun; for larger λ worst case risk is for rain.

Added to graph: loss functions for each d_λ, (straight line) and set of (x, y) pairs for which $\min(x, y) = 3.8$ — 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:



Randomization in decision problems permits the assumption that the set of possible risk functions is convex — an important technical conclusion used to prove many basic decision theory results.

The graph shows that many points in the picture correspond to bad decision procedures. Rain or shine not taking my car to work has a lower loss than staying home; the decision to stay home is *inadmissible*.

Definition: A decision rule δ is **inadmissible** if there is a rule δ^* such that

$$R_{\delta^*}(\theta) \le R_\delta(\theta)$$

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**.

Admissible procedures have risks on lower left of graphs, i.e., lines connecting B to T and T to C are the admissible procedures.

10.0.6 Connection between Bayes procedures and admissible procedures

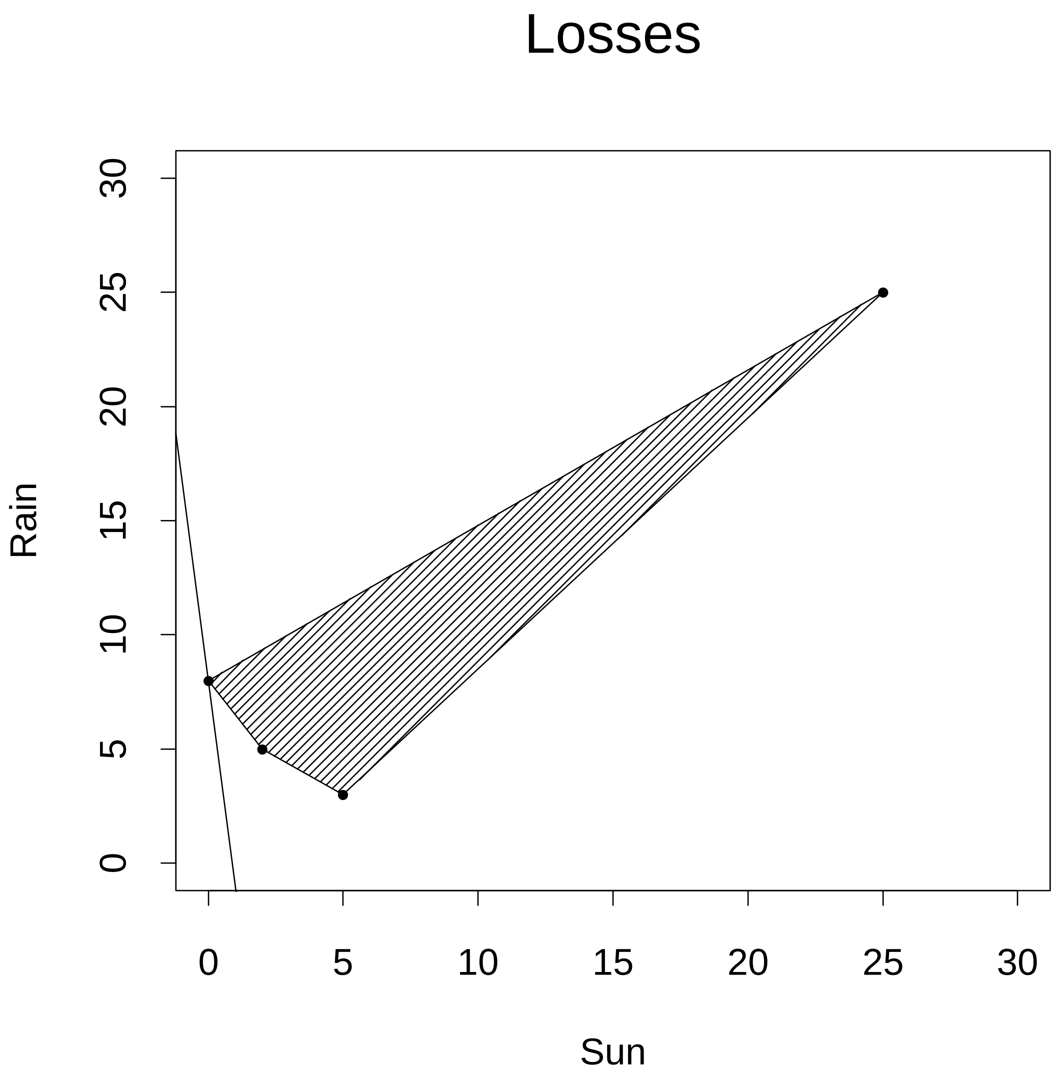
A prior distribution in the example is specified by two probabilities, π_S and π_R which add up to 1. If $L = (L_S, L_R)$ is the risk function for some procedure then the Bayes risk is

$$r_\pi = \pi_R L_R + \pi_S L_S.$$

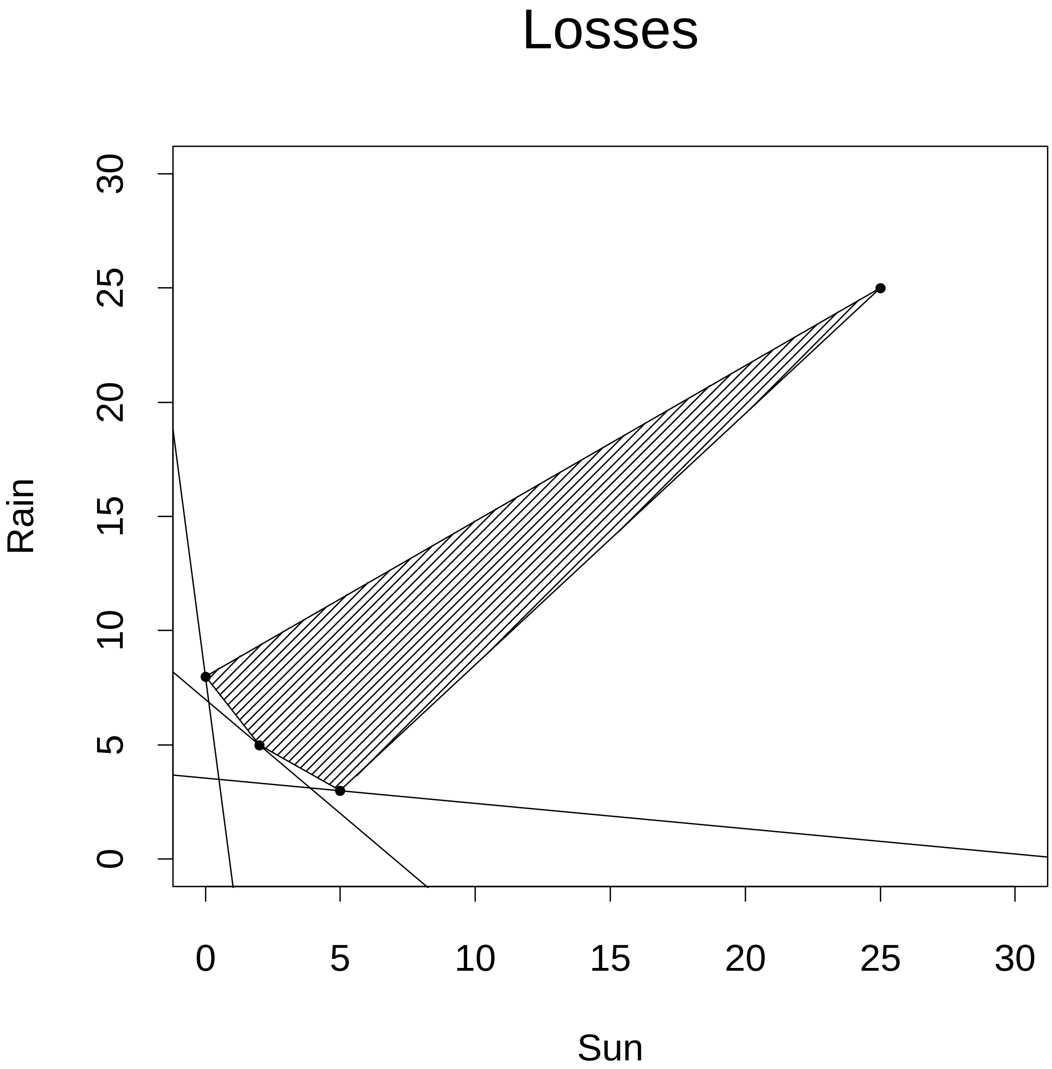
Consider the set of L such that this Bayes risk is equal to some constant. On our picture this is a line with slope $-\pi_S/\pi_R$.

Now consider three priors: $\pi_1 = (0.9, 0.1)$, $\pi_2 = (0.5, 0.5)$ and $\pi_3 = (0.1, 0.9)$. For 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.

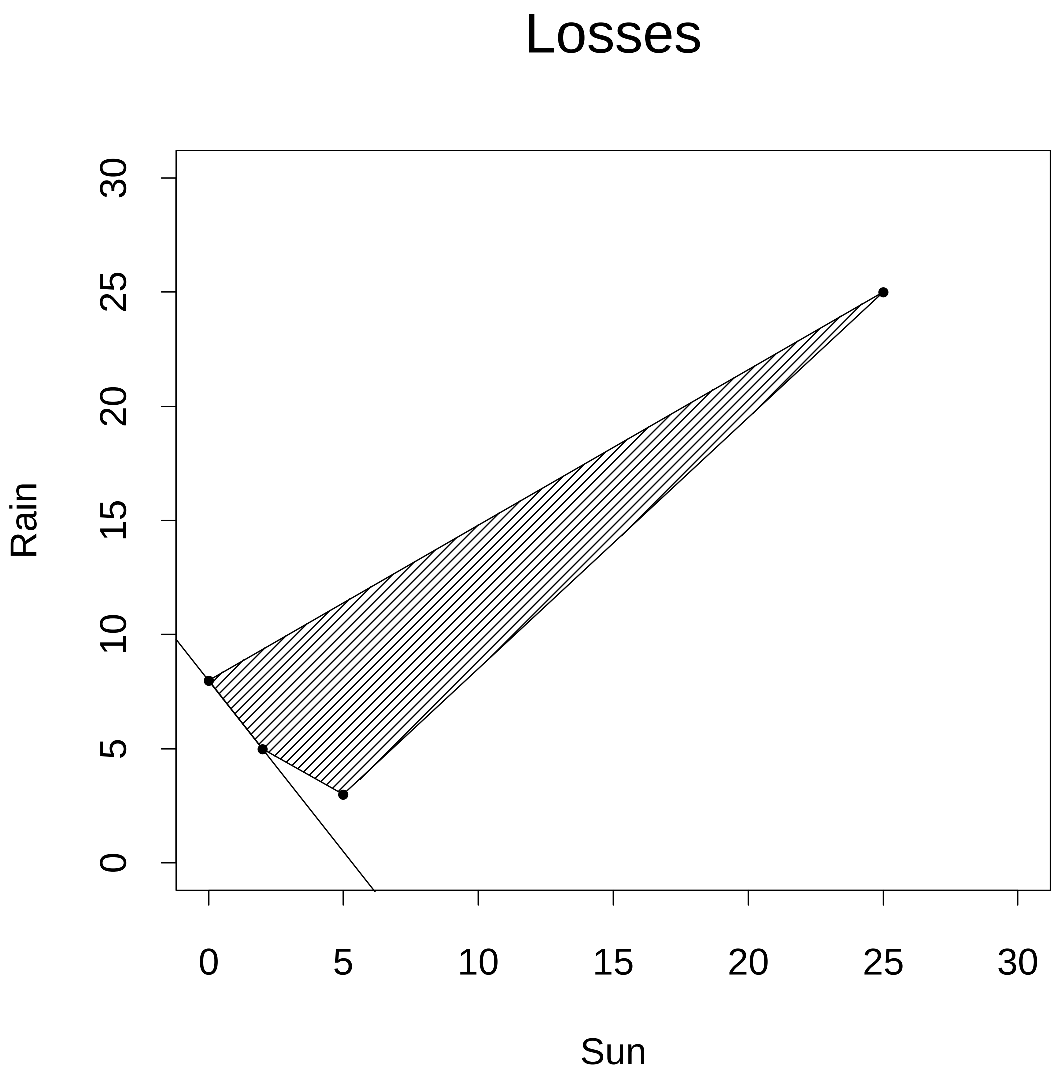
Sliding this line to the right corresponds to making r_π larger and larger so that when it just touches the convex set we have found the Bayes procedure.



Here is a picture showing the same lines for the three priors above.



The Bayes procedure for π_1 (a prior which says you're 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. Prior $(0.6, 0.4)$ produces the line shown here:



Any point on line BT is Bayes for this prior.

Decision Theory and Bayesian Methods
Summary for no data case

- 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

  $$r_\pi(\delta) = E_\pi(L(\delta,\theta)) = \int L(\delta,\theta)\pi(\theta)d\theta$$

  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

  $$r_\pi(\delta^*) \le r_\pi(\delta)$$

  for any decision δ.
- For infinite parameter spaces: $\pi(\theta) > 0$ on Θ is a proper prior if $\int \pi(\theta)d\theta < \infty$; divide π by integral to get a density. If $\int \pi(\theta)d\theta = \infty$ π is an **improper** prior density.
- Decision δ is **inadmissible** if there is δ^* such that

  $$L(\delta^*,\theta) \le L(\delta,\theta)$$

  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 prior with density > 0 for all θ) is admissible.

  Proof: If δ is Bayes for π but not admissible there is a δ^* such that

  $$L(\delta^*,\theta) \le L(\delta,\theta)$$

  Multiply by the prior density; integrate:

  $$r_\pi(\delta^*) \le r_\pi(\delta)$$

  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: the theorem actually requires the extra hypotheses: positive density, and risk functions of δ and δ^* 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.

Decision Theory and Bayesian Methods
Summary when there is data

- 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 $L = L(d, \theta)$: loss I incur if I do d and θ is true state of nature.

- Add data $X \in \mathcal{X}$ with model $\{P_\theta; \theta \in \Theta\}$: model density is $f(x|\theta)$.

- A *procedure* is a map $\delta : \mathcal{X} \mapsto \mathcal{D}$.

- The risk function for δ is the expected loss:

$$R_\delta(\theta) = R(\delta, \theta) = \mathrm{E}\left[L\{\delta(X), \theta\}\right].$$

- We call δ^* minimax if

$$\max_\theta R(\delta^*, \theta) = \min_\delta \max_\theta R(\delta, \theta).$$

- A **prior** is a probability distribution π on Θ,.

- **Bayes risk** of decision δ for prior π is

$$\begin{aligned} r_\pi(\delta) &= E_\pi(R(\delta, \theta)) \\ &= \int L(\delta(x), \theta) f(x|\theta)\pi(\theta) dx d\theta \end{aligned}$$

  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

$$r_\pi(\delta^*) \le r_\pi(\delta)$$

  for any decision δ.

- For infinite parameter spaces: $\pi(\theta) > 0$ on Θ is a **proper** prior if $\int \pi(\theta)d\theta < \infty$; divide π by integral to get a density. If $\int \pi(\theta)d\theta = \infty$ π is an **improper** prior density.

- Decision δ is **inadmissible** if there is δ^* such that

$$R(\delta^*, \theta) \le R(\delta, \theta)$$

  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.
- If every risk function is continuous then every Bayes procedure with finite Bayes risk (for prior with density > 0 for all θ) is admissible.
- A minimax procedure is admissible.
- The minimax procedure has constant risk. The admissible minimax procedure is Bayes for some π; its risk is constant on the set of θ for which the prior density is positive.

10.1 Bayesian Estimation

In this section I will focus on the problem of estimation of a 1 dimensional parameter, θ. Earlier we discussed comparing estimators in terms of Mean Squared Error. In the language of decision theory Mean Squared Error corresponds to using

$$L(d,\theta) = (d-\theta)^2$$

which is called squared error loss. The multivariate version would be

$$L(d,\theta) = ||d-\theta||^2$$

or possibly the more general formula

$$L(d,\theta) = (d-\theta)^T\mathbf{Q}(d-\theta)$$

for some positive definite symmetric matrix $\mathbf{Q}$. The risk function of a procedure (estimator) $\hat\theta$ is

$$R_{\hat\theta}(\theta) = E_\theta[(\hat\theta-\theta)^2].$$

Now consider prior with density $\pi(\theta)$. The Bayes risk of $\hat\theta$ is

$$\begin{aligned} r_\pi &= \int R_{\hat\theta}(\theta)\pi(\theta)d\theta \\ &= \int\int (\hat\theta(x)-\theta)^2 f(x;\theta)\pi(\theta)dxd\theta \end{aligned}$$

For a Bayesian the problem is then to choose $\hat\theta$ to minimize r_π? This problem will turn out to be analogous to the calculations I made when I minimized $\beta+\lambda\alpha$ in hypothesis testing. First recognize that $f(x;\theta)\pi(\theta)$ is really a joint density

$$\int\int f(x;\theta)\pi(\theta)dxd\theta = 1$$

For this joint density: conditional density of X given θ is just the model $f(x;\theta)$. This justifies the standard notation $f(x|\theta)$ for $f(;\theta)$¿ Now I will 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** of θ given the data $X = x$. This posterior density may be 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

$$r_\pi(\hat{\theta}) = \int \left[\int (\hat{\theta}(x) - \theta)^2 \pi(\theta|x) d\theta \right] f(x) dx$$

[REMEMBER the meta-theorem: when you see a double integral it is always written in the wrong order. Change the order of integration to learn something useful.] Notice that by writing the integral in this order you see that you 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 $\hat{\theta}(x)$; it is 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: estimating 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 collect 25 measurements of the speed of sound. Assume: 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. So model is: given μ, $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ variables.

The joint density of the data and μ is then

$$(2\pi)^{-n/1}\sigma^{-n} \exp\{-\sum (X_i - \mu)^2/(2\sigma^2)\} \times (2\pi)^{-1/2}\tau^{-1} \exp\{-(\mu - \nu)^2/\tau^2\}.$$

Thus $(X_1, \ldots, X_n, \mu) \sim MVN$. Conditional distribution of θ given $X_1, \ldots, X_n$ is normal. We can now use standard MVN formulas to calculate conditional means and variances.

Alternatively: the exponent in joint density has 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

$$\frac{\psi}{\gamma^2} = \frac{\sum X_i}{\sigma^2} + \frac{\nu}{\tau^2}$$

So: the conditional distribution 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: 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.

Example: $N(\mu, \sigma^2)$; consider prior density

$$\pi(\mu) \equiv 1.$$

This "density" integrates to ∞; using Bayes' theorem to compute the posterior would give

$$\pi(\mu|X) = \frac{(2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i - \mu)^2/(2\sigma^2)\}}{\int (2\pi)^{-n/2}\sigma^{-n}\exp\{-\sum(X_i - \xi)^2/(2\sigma^2)\}d\xi}$$

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 corresponding to 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: Given p, X has a Binomial(n,p) distribution.

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

$$\binom{n}{X} p^X (1-p)^{n-X} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} ;$$

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.

This is Beta$(X+\alpha, n-X+\beta)$ density. Mean of Beta(α,β) distribution is $\alpha/(\alpha+\beta)$.

So 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: again weighted average of prior mean and mle.

Notice: prior is proper for $\alpha > 0$ and $\beta > 0$.

To get $w = 1$ take $\alpha = \beta = 0$; use improper prior

$$\frac{1}{p(1-p)}$$

Again: 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 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

$$w^2 \text{Var}(\hat{p}) + (wp + (1-w)p - p)^2 = w^2 p(1-p)/n + (1-w)^2(p-p_0)^2 .$$

Risk function constant if coefficients of p^2 and p in risk are 0.

Coefficient of p^2 is

$$-w^2/n + (1-w)^2$$

so $w = n^{1/2}/(1+n^{1/2})$.

Coefficient of p is then

$$w^2/n - 2p_0(1-w)^2$$

which vanishes if $2p_0 = 1$ or $p_0 = 1/2$.

Working backwards: to get these values for w and p_0 require $\alpha = \beta$. Moreover

$$w^2/(1-w)^2 = n$$

gives

$$n/(\alpha+\beta) = \sqrt{n}$$

or $\alpha = \beta = \sqrt{n}/2$. Minimax estimate of p is

$$\frac{\sqrt{n}}{1+\sqrt{n}}\hat{p} + \frac{1}{1+\sqrt{n}}\frac{1}{2}$$

Example: $X_1, \ldots, X_n$ iid $MVN(\mu, \Sigma)$ with Σ known.

Take improper prior for μ which is constant.

Posterior of μ given X is then $MVN(\bar{X}, \Sigma/n)$.

Multivariate estimation: 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 risk is sum of MSEs of individual components.

Bayes estimate is again posterior mean. 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.

Fact first demonstrated by James and Stein who produced an estimate which is better, in terms of this risk function, for every μ.

So-called **James Stein** estimator is essentially never used.

10.2 Bayesian Hypothesis Testing

Hypothesis Testing and Decision Theory

Decision analysis of hypothesis testing takes $D = \{0, 1\}$ and

$$L(d, \theta) = 1(\text{make an error})$$

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 $\delta = P(\text{reject})$ so $\mathcal{D} = [0, 1]$. The loss function of $\delta \in [0, 1]$ is

$$L(\delta, \theta) = (1-\delta)\ell_1 1(\theta \in \Theta_1) + \delta \ell_0 1(\theta \in \Theta_0)\,.$$

Definition: **Simple hypotheses**: Prior is $\pi_0 > 0$ and $\pi_1 > 0$ with $\pi_0 + \pi_1 = 1$.

Definition: Procedure: map from sample space to $\mathcal{D}$ – a test function.

Risk function of 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

$$\phi(X) = 1(f_1(X)/f_0(X) > \pi_0 \ell_0/(\pi_1 \ell_1))$$

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.

10.3 Optimal Estimation Theory

10.4 Unbiased Estimation Theory

The Binomial problem we considered shows a general phenomenon, namely, an estimator can be good for some values of θ and bad for others. To compare $\hat\theta$ and $\tilde\theta$, two estimators of θ we way $\hat\theta$ is better than $\tilde\theta$ if it has *uniformly* smaller MSE:

$$MSE_{\hat\theta}(\theta) \le MSE_{\tilde\theta}(\theta)$$

for **all** θ. Normally we also require that the inequality be strict for at least one θ.

Warning: Of course we could measure the quality of an estimator in some way other than MSE!

Question is there a *best* estimate – one which is better than every other estimator?

Answer NO. Suppose $\hat\theta$ were such a best estimate. Fix a θ^* in Θ and let $\tilde\theta \equiv \theta^*$. [This is a crazy estimator – it just ignores the data and guesses $\hat\theta^*$. But consider for a minute tossing an ordinary coin to estimate the probability, p, that it lands Heads up. How many tosses would it take to convince you that you should use the traditional estimator of p instead of just assuming the coin is fair and estimating that p is 1/2?] Then the MSE of $\tilde\theta$ is 0 when $\theta = \theta^*$. Since $\hat\theta$ is better than $\tilde\theta$ we must have

$$MSE_{\hat\theta}(\theta^*) = 0$$

so that $\hat\theta = \theta^*$ with probability equal to 1 when the true value of θ is θ^*. So $\hat\theta = \tilde\theta$.

[Pedants might need more convincing; I have proved that when the true value of θ is θ^* the estimator has variance 0. How do I know this is true for some other value of θ? Let's imagine that the data X has density $f(x;\theta)$. Then for any set A in the range of X we have

$$P_\theta(X \in A) = \int_A f(x,\theta)dx$$

I claim that if this integral is 0 when $\theta = \theta^*$ then it is 0 for all θ. To see this write

$$\int_A f(x,\theta)dx = \int_A \frac{f(x,\theta)}{f(x,\theta^*)} f(x,\theta^*)dx$$

If $\int_A f(x,\theta^*)dx = 0$ then the integrand is 0 (technically 0 almost everywhere) on A. That means that the product

$$\frac{f(x,\theta)}{f(x,\theta^*)} f(x,\theta^*) = 0$$

almost everywhere on A – unless the first term is actually $+\infty$. So I modify my assertion slightly: if all the different densities $f(x,\theta)$ are positive on the same set (jargon – they all have the same *support*) then

$$P_{\theta^*}(X \in A) = 0$$

implies

$$P_\theta(X \in A) = 0$$

for all θ. Letting A be $\{x : \hat\theta(x) \neq \theta^*\}$ finishes our pedantry.]

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,

$$E_\theta(\hat\theta) \equiv \theta\,.$$

Warning: In my view the Principle of Unbiasedness is a load of hog wash.

For an unbiased estimate the MSE is just the variance. This means that if we only allow unbiased estimates we can compare them by comparing their variances. Sometimes this is enough of a restriction on the set of possible estimators to mean that there is a single overall best estimator.

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

$$\text{Var}_\theta(\hat\phi) \le \text{Var}_\theta(\tilde\phi)$$

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.

10.4.1 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 the score function has mean 0 (at the true parameter value)

$$\mathrm{Cov}_\theta(T(X),U(\theta)) = 1$$

Remember that correlations are between -1 and 1 so that

$$\begin{aligned} 1 &= |\mathrm{Cov}_\theta(T(X),U(\theta))| \\ &\le \sqrt{\mathrm{Var}_\theta(T)\mathrm{Var}_\theta(U(\theta))}\,. \end{aligned}$$

Squaring gives the so-called Cramér Rao Lower Bound (often given the acronym CRLB):

$$\mathrm{Var}_\theta(T) \ge \frac{1}{I(\theta)}\,.$$

The inequality is strict unless the correlation is ±1 so that

$$U(\theta) = A(\theta)T(X) + B(\theta)$$

for 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 other 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 $\text{Var}_\theta(T(X)) \equiv 1/I(\theta)$ then $T(X)$ is the UMVUE. In the $N(\mu, 1)$ example the Fisher information is n and $\text{Var}(\overline{X}) = 1/n$ 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.
- Densities of exponential form (called *exponential family*) given above are somehow special.
- Usually inequality is strict — strict unless score is affine function of a statistic T and T (or T/c for constant c) 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

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

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.

Now here is a different tactic: 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 this equivalence sign is a polynomial function of p as is the right hand side. 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 $k = 0$; its coefficient is

$$h(0)\binom{n}{0} = h(0)\,.$$

Thus $h(0) = 0$. Now $p^1 = p$ occurs only in term $k = 1$ with coefficient $nh(1)$ so $h(1) = 0$. Since terms with $k = 0$ or 1 are 0 the quantity p^2 occurs only in $k = 2$ term; coefficient is

$$n(n-1)h(2)/2$$

so $h(2) = 0$. Continue in this way to see that $h(k) = 0$ for each k.

Conclusion: So the *only* unbiased estimate which is a function of X is X/n.

Note: If there is only one unbiased estimate then of course that one estimate is the best possible unbiased estimate!

A Binomial random variable is a sum of n iid Bernoulli(p) rvs. If $Y_1,\ldots,Y_n$ iid Bernoulli(p) then $X=\sum Y_i$ is Binomial(n,p). That leads to the question: could we do better than $\hat{p}=X/n$ by trying $T(Y_1,\ldots,Y_n)$ for some other function T?

I am going to try to build some insight by studying small values of n before presenting the general theory. Try $n=2$. 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

$$E_p(h(Y_1,Y_2))\equiv 0$$

and we have

$$\begin{aligned} E_p(h(Y_1,Y_2)) &= h(0,0)(1-p)^2 \\ & \quad +[h(1,0)+h(0,1)]p(1-p) \\ & \quad +h(1,1)p^2 \,. \end{aligned}$$

This can be rewritten in the form

$$\sum_{k=0}^n w(k)\binom{n}{k}p^k(1-p)^{n-k}$$

where

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

So, as before $w(0)=w(1)=w(2)=0$. This argument can be used to prove:

For any unbiased estimate $T(Y_1,\ldots,Y_n)$: the average value of $T(y_1,\ldots,y_n)$ over those values $y_1,\ldots,y_n$ which have exactly k 1s and $n-k$ 0s is k/n.

Now let's look at the variance of T:

$$\begin{aligned} &\mathrm{Var}(T) \\ &\quad = E_p([T(Y_1,\ldots,Y_n)-p]^2) \\ &\quad = E_p([T(Y_1,\ldots,Y_n)-X/n+X/n-p]^2) \\ &\quad = E_p([T(Y_1,\ldots,Y_n)-X/n]^2)+ \\ &\qquad 2E_p([T(Y_1,\ldots,Y_n)-X/n][X/n-p]) \\ &\quad + E_p([X/n-p]^2) \end{aligned}$$

I claim 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 $T(Y_1,\ldots,Y_n)\equiv X/n$). Compute the cross product term by writing

$$\begin{aligned} &E_p([T(Y_1,\ldots,Y_n)-X/n][X/n-p]) \\ &\qquad = \sum_{y_1,\ldots,y_n}[T(y_1,\ldots,y_n)-\sum y_i/n][\sum y_i/n-p] \\ &\qquad\qquad \times p^{\sum y_i}(1-p)^{n-\sum y_i} \end{aligned}$$

Sum over those $y_1, \ldots, y_n$ whose sum is an integer x; then sum over x:

$$\begin{aligned}
&E_p([T(Y_1,\ldots,Y_n) - X/n][X/n-p]) \\
&\qquad = \sum_{x=0}^{n} \sum_{\sum y_i = x} [T(y_1,\ldots,y_n) - \sum y_i/n] \\
&\qquad\quad \times [\sum y_i/n - p] p^{\sum y_i}(1-p)^{n-\sum y_i} \\
&\qquad = \sum_{x=0}^{n} \left[\sum_{\sum y_i = x} [T(y_1,\ldots,y_n) - x/n] \right] [x/n - p] \\
&\qquad\qquad \times p^x(1-p)^{n-x}
\end{aligned}$$

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 rewrite

$$\begin{aligned}
&E_p\{T(Y_1,\ldots,Y_n)\} \\
&\qquad = \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) \\
&\qquad\quad \times P(Y_1 = y_1, \ldots, Y_n = y_n) \\
&\qquad = \sum_{x=0}^{n} \sum_{\sum y_i = x} T(y_1,\ldots,y_n) \\
&\qquad\quad \times P(Y_1 = y_1, \ldots, Y_n = y_n | X = x) P(X = x) \\
&\qquad = \sum_{x=0}^{n} \frac{\sum_{\sum y_i = x} T(y_1,\ldots,y_n)}{\binom{n}{x}} \binom{n}{x} p^x (1-p)^{n-x}
\end{aligned}$$

Note: : the large fraction is the average value of T over all those y such that $\sum y_i = x$. Notice also that the weights in average do not depend on p. Finally notice that this average is actually

$$\begin{aligned}
&E\{T(Y_1,\ldots,Y_n)|X = x\} \\
&\qquad = \sum_{y_1,\ldots,y_n} T(y_1,\ldots,y_n) \\
&\qquad\qquad \times P(Y_1 = y_1, \ldots, Y_n = y_n | X = x)
\end{aligned}$$

And then see 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: 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 the CRLB is strict except for multiples of X. Since the only unbiased multiple of X is $\hat{p} = X/n$ the UMVUE of p is $\hat{p}$.

10.4.2 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: 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 (that is, the conditional distribution is the same for all values of θ.
Intuition: Data tell us about θ **if** different values of θ give different distributions to X. If two different values of θ correspond to same density or cdf for X we cannot distinguish these two values of θ by examining X. As an extension of this notion: if two values of θ give the same conditional distribution of X given T then observing T in addition to X doesn't improve our ability to distinguish the two values.
Mathematically Precise version of this intuition: Suppose $T(X)$ is a sufficient statistic and $S(X)$ is any estimate or confidence interval or .... If you 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: $Y_1, \ldots, Y_n$ iid Bernoulli(p). 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 $\{1, \ldots, n\}$. Chance does not depend on p so $T(Y_1, \ldots, Y_n) = \sum Y_i$ is sufficient statistic.
Example 2: $X_1, \ldots, X_n$ iid $N(\mu, 1)$. Joint distribution of $X_1, \ldots, X_n, \overline{X}$ is MVN. All entries of mean vector are μ. Variance covariance matrix partitioned as

$$\left[\begin{array}{cc} I_{n\times n} & \mathbf{1}_n/n \\ \mathbf{1}_n^t/n & 1/n \end{array}\right]$$

where $\mathbf{1}_n$ is column vector of n 1s and $I_{n\times n}$ is $n \times n$ identity matrix.

Compute conditional means and variances of X_i given $\overline{X}$; use fact that conditional law is MVN. Conclude conditional law of data given $\overline{X} = x$ is MVN. Mean vector has all entries x. Variance-covariance matrix is $I_{n\times n} - \mathbf{1}_n\mathbf{1}_n^t/n$. No dependence on μ so $\overline{X}$ is sufficient.
WARNING: Whether or not statistic is sufficient depends on density function and on Θ.
Theorem: [Rao-Blackwell] Suppose $S(X)$ is a sufficient statistic for model $\{P_\theta, \theta \in \Theta\}$. If T is an estimate of $\phi(\theta)$ then:

1. $E(T|S)$ is a statistic.

2. $E(T|S)$ has the same bias as T; if T is unbiased so is $E(T|S)$.

3. $\text{Var}_\theta(E(T|S)) \le \text{Var}_\theta(T)$ and the inequality is strict unless T is a function of S.

4. MSE of $E(T|S)$ is no more than MSE of T.

Proof: Review conditional distributions: abstract definition of conditional expectation is:
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 $R(X)$. $E(Y|X = x)$ 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

$$g(x) = \int yf(y|x)dy$$

satisfies these definitions.
Proof:

$$\begin{aligned} E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\ &= \int\int R(x)yf_X(x)f(y|x)dydx \\ &= \int\int R(x)yf_{X,Y}(x,y)dydx \\ &= E(R(X)Y) \end{aligned}$$

Think of $E(Y|X)$ as average Y holding X fixed. Behaves like ordinary expected value but functions of X only are like constants:

$$E(\sum A_i(X)Y_i|X) = \sum A_i(X)E(Y_i|X)$$

Example: $Y_1, \ldots, Y_n$ iid Bernoulli(p). Then $X = \sum Y_i$ is Binomial(n, p). Summary of conclusions:

- Log likelihood function of X only not of $Y_1, \ldots, Y_n$.
- Only function of X which is unbiased estimate of p is $\hat{p} = X/n$.
- If $T(Y_1, \ldots, Y_n)$ is unbiased for p then average value of $T(y_1, \ldots, y_n)$ over $y_1, \ldots, y_n$ for which $\sum y_i = x$ is x/n.
- Distribution of T given $\sum Y_i = x$ does not depend on p.
- If $T(Y_1, \ldots, Y_n)$ is unbiased for p then
$$\text{Var}(T) = \text{Var}(\hat{p}) + E[(T - \hat{p})^2]$$
- $\hat{p}$ is the UMVUE of p.

This proof that $\hat{p} = X/n$ is UMVUE of p is special case of general tactic.

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: relies on very useful decomposition. (Total sum of squares = regression sum of squares + residual sum of squares.)

$$\text{Var(Y)} = \text{Var}\{E(Y|X)\} + E[\text{Var}(Y|X)]$$

The conditional variance means

$$\text{Var}(Y|X) = E[\{Y - E(Y|X)\}^2|X]$$

Square out right hand side:

$$\begin{aligned} \text{Var}(E(Y|X)) &= E[\{E(Y|X) - E[E(Y|X)]\}^2] \\ &= E[\{E(Y|X) - E(Y)\}^2] \end{aligned}$$

and

$$E[\text{Var}(Y|X)] = E[\{Y - E(Y|X)\}^2]$$

Adding these together gives

$$E\left[Y^2 - 2YE[Y|X] + 2(E[Y|X])^2 \right.$$
$$\left. -2E(Y)E[Y|X] + E^2(Y)\right]$$

Simplify remembering $E(Y|X)$ is function of X — constant when holding X fixed. So

$$E[Y|X]E[Y|X] = E[YE(Y|X)|X]$$

taking expectations gives

$$\begin{aligned} E[(E[Y|X])^2] &= E[E[YE(Y|X)|X]] \\ &= E[YE(Y|X)] \end{aligned}$$

So 3rd term above cancels with 2nd term.

Fourth term simplifies

$$E[E(Y)E[Y|X]] = E(Y)E[E[Y|X]] = E^2(Y)$$

so that

$$\text{Var}(E(Y|X)) + E[\text{Var}(Y|X)] = E[Y^2] - E^2(Y)$$

Apply to Rao Blackwell theorem to get

$$\text{Var}_\theta(T) = \text{Var}_\theta(E(T|S)) + E[(T - E(T|S))^2]$$

Second term ≥ 0 so variance of $E(T|S)$ is no more than that of T; will be strictly less unless $T = E(T|S)$. This would mean that T is already a function of S. Adding the squares of the biases of T (or of $E(T|S)$) gives the inequality for MSE.

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:

$$
\begin{aligned}
&E(Y_1(1-Y_2)|X=x)\\
&\quad= P(Y_1(1-Y_2)=1|X=x)\\
&\quad= P(Y_1=1, Y_2=0|Y_1+Y_2+\cdots+Y_n=x)\\
&\quad= \frac{P(Y_1=1, Y_2=0, Y_1+\cdots+Y_n=x)}{P(Y_1+Y_2+\cdots+Y_n=x)}\\
&\quad= \frac{P(Y_1=1, Y_2=0, Y_3+\cdots+Y_n=x-1)}{\binom{n}{x}p^x(1-p)^{n-x}}\\
&\quad= \frac{p(1-p)\binom{n-2}{x-1}p^{x-1}(1-p)^{(n-2)-(x-1)}}{\binom{n}{x}p^x(1-p)^{n-x}}\\
&\quad= \frac{\binom{n-2}{x-1}}{\binom{n}{x}}\\
&\quad= \frac{x(n-x)}{n(n-1)}
\end{aligned}
$$

This is simply $n\hat{p}(1-\hat{p})/(n-1)$ (can be bigger than 1/4, 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

Binomial(n,θ): log likelihood $\ell(\theta)$ (part depending on θ) is function of X alone, not of $Y_1, \ldots, Y_n$ as well.

Normal example: $\ell(\mu)$ is, ignoring terms not containing μ,

$$\ell(\mu) = \mu\sum X_i - n\mu^2/2 = n\mu\bar{X} - n\mu^2/2\,.$$

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

$$f(x,\theta) = g(S(x),\theta)h(x)$$

Proof: Find statistic $T(X)$ such that X is a one to one function of the pair S, T. Apply change of variables to the joint density of S and T. If the density factors then

$$f_{S,T}(s,t) = g(s,\theta)h(x(s,t))J(s,t)$$

where J is the Jacobian, so 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 $f_{T|S}$ has no θ in it so joint density of S, T is

$$f_S(s,\theta)f_{T|S}(t|s)$$

Apply change of variables formula to get

$$f_X(x) = f_S(S(x),\theta)f_{T|S}(t(x)|S(x))J(x)$$

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

$$(2\pi)^{-n/2}\sigma^{-n}\times$$
$$\exp\{-\sum X_i^2/(2\sigma^2) + \mu\sum X_i/\sigma^2 - n\mu^2/(2\sigma^2)\}$$

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.
Example: If $Y_1,\ldots,Y_n$ are iid Bernoulli(p) then

$$\begin{aligned} f(y_1,\ldots,y_p;p) &= \prod p^{y_i}(1-p)^{1-y_i} \\ &= p^{\sum y_i}(1-p)^{n-\sum y_i} \end{aligned}$$

Define $g(x,p) = p^x(1-p)^{n-x}$ and $h \equiv 1$ to see that $X = \sum Y_i$ is sufficient by the factorization criterion.

Minimal Sufficiency

In any model $S(X) \equiv X$ is sufficient. (Apply the factorization criterion.) In any iid model the vector $X_{(1)},\ldots,X_{(n)}$ of order statistics is sufficient. (Apply the factorization criterion.) In $N(\mu,1)$ model we have 3 sufficient statistics:

1. $S_1 = (X_1,\ldots,X_n)$.

2. $S_2 = (X_{(1)},\ldots,X_{(n)})$.

3. $S_3 = \bar{X}$.

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.)

Recognize minimal sufficient statistics from ℓ:

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.

Subtraction of $\ell(\theta^*)$ gets rid of irrelevant constants in ℓ. In $N(\mu, 1)$ example:

$$\ell(\mu) = -n\log(2\pi)/2 - \sum X_i^2/2 + \mu \sum X_i - n\mu^2/2$$

depends on $\sum X_i^2$, not needed for 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: from $\sum X_i$ can compute this minimal sufficient statistic and vice versa. Thus $\sum X_i$ is also minimal sufficient.

Completeness

In Binomial(n, p) example only one function of X is unbiased. Rao Blackwell shows UMVUE, if it exists, will be a function of any sufficient statistic.

Q: Can there be more than one such function?

A: Yes in general but no for some models like the binomial.

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

$$E(h(\bar{X})) = \frac{\sqrt{n}e^{-n\mu^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(x)e^{-nx^2/2}e^{n\mu x}dx$$

so that

$$\int_{-\infty}^{\infty} h(x)e^{-nx^2/2}e^{n\mu x}dx \equiv 0\,.$$

Called Laplace transform of $h(x)e^{-nx^2/2}$.

Theorem: 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

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: $N(\mu,\sigma^2)$ model density has form

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

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

$$c(\theta) = -\frac{\mu^2}{2\sigma^2} - \log\sigma\,.$$

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.

FACT: A complete sufficient statistic is also minimal sufficient.

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 Rao-Blackwell, 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

$$E\left[\frac{\sqrt{n-1}s}{\sigma}\right] = \frac{\int_0^\infty x^{1/2}\left(\frac{x}{2}\right)^{(n-1)/2-1} e^{-x/2} dx}{2\Gamma((n-1)/2)}.$$

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

$$E(s) = \sigma \frac{\sqrt{2}\Gamma(n/2)}{\sqrt{n-1}\Gamma((n-1)/2)}.$$

The UMVUE of σ is then

$$s\frac{\sqrt{n-1}\Gamma((n-1)/2)}{\sqrt{2}\Gamma(n/2)}$$

by the Lehmann-Scheffé theorem.

Criticism of Unbiasedness

- UMVUE can be **inadmissible for squared error loss** meaning 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 $\hat\phi$.

- Unbiased estimation may be impossible. Binomial(n, p) log odds 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

- The UMVUE of σ is not the square root of the UMVUE of σ^2. This method of estimation does not have the parametrization equivariance that maximum likelihood does.

- Unbiasedness is irrelevant (unless you average together many estimators).

  Property is an average over possible values of the estimate in which positive errors are allowed to cancel negative errors.

  Exception to criticism: if you average a number of estimators to get a single estimator then it is a problem if all the estimators have the same bias.

  See assignment 5, one way layout example: 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.

Linear Algebra
Linear Algebra Review Notes

Notation: We will need the following notation:

- Vectors $x \in R^n$ are column vectors

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

- An $m \times n$ matrix A has m rows, n columns and entries A_{ij}.

- Matrix and vector addition are defined componentwise:

$$(A+B)_{ij} = A_{ij} + B_{ij}; \qquad (x+y)_i = x_i + y_i$$

- If A is $m \times n$ and B is $n \times r$ then AB is the $m \times r$ matrix

$$(AB)_{ij} = \sum_{k=1}^{n} A_{ik} B_{kj}$$

- The matrix I or sometimes $I_{n \times n}$ which is an $n \times n$ matrix with $I_{ii} = 1$ for all i and $I_{ij} = 0$ for any pair $i \neq j$ is called the $n \times n$ **identity matrix**.

- The **span** of a set of vectors $\{x_1, \ldots, x_p\}$ is the set of all vectors x of the form $x = \sum c_i x_i$. It is a vector space. The **column space** of a matrix, A, is the span of the set of columns of A. The **row space** is the span of the set of rows.

- A set of vectors $\{x_1, \ldots, x_p\}$ is **linearly independent** if $\sum c_i x_i = 0$ implies $c_i = 0$ for all i. The **dimension** of a vector space is the cardinality of the largest possible set of linearly independent vectors.

Definition: The **transpose**, A^T, of an $m \times n$ matrix A is the $n \times m$ matrix whose entries are given by

$$(A^T)_{ij} = A_{ji}$$

so that A^T is $n \times m$. We have

$$(A+B)^T = A^T + B^T$$

and

$$(AB)^T = B^T A^T\,.$$

Definition: The **rank** of a matrix A is the number of linear independent columns of A; we use $\text{rank}(A)$ for notation. We have

$$\begin{aligned}\text{rank}(A) &= \dim(\text{column space of } A)\\ &= \dim(\text{row space of } A)\\ &= \text{rank}(A^T)\end{aligned}$$

If A is $m \times n$ then $\text{rank}(A) \le \min(m,n)$.

Matrix inverses

For this little section all matrices are square $n \times n$ matrices.

If there is a matrix B such that $BA = I_{n\times n}$ then we call B the inverse of A. If B exists it is unique and $AB = I$ and we write $B = A^{-1}$. The matrix A has an inverse if and only if $\text{rank}(A) = n$.

Inverses have the following properties:

$$(AB)^{-1} = B^{-1}A^{-1}$$

(if one side exists then so does the other) and

$$(A^T)^{-1} = (A^{-1})^T$$

Determinants

Again A is $n \times n$. The determinant if a function on the set of $n \times n$ matrices such that:

1. $\det(I) = 1$.
2. If A' is the matrix A with two columns interchanged then

   $$\det(A^{`}prime) = -\det(A)\,.$$

   (Notice that this means that two equal columns guarantees $\det(A) = 0$.)
3. $\det(A)$ is a linear function of each column of A. That is if $A = (a_1, \ldots, a_n)$ with a_i denoting the ith column of the matrix then

   $$\begin{aligned}\det(a_1, \ldots, a_i + b_i, \ldots, a_n) =&\det(a_1, \ldots, a_i, \ldots, a_n)\\ &+ \det(a_1, \ldots, b_i, \ldots, a_n)\end{aligned}$$

Here are some properties of the determinant:

1. $\det(A^T) = \det(A)$.
2. $\det(AB) = \det(A)\det(B)$.
3. $\det(A^{-1})_{=}1/\det(A)$.
4. A is invertible if and only if $\det(A) \neq 0$ if and only if $\text{rank}(A) = n$.
5. Determinants can be computed (slowly) by expansion by minors.

Special Kinds of Matrices

1. A is symmetric if $A^T = A$.
2. A is orthogonal if $A^T = A^{-1}$ (or $AA^T = A^T A = I$).
3. A is idempotent if $AA \equiv A^2 = A$.
4. A is diagonal if $i \neq j$ implies $A_{ij} = 0$.

Inner Products and orthogonal and orthonormal vectors

Definition: Two vectors x and y are **orthogonal** if $x^T y = \sum x_i y_i = 0$.

Definition: The **inner product** or **dot product** of x and y is

$$< x, y >= x^T y = \sum x_i y_i$$

Definition: x and y are **orthogonal** if $x^T y = 0$.

Definition: The **norm** (or length) of x is $||x|| = (x^t x)^{1/2} = (\sum x_i^2)^{1/2}$

A is orthogonal if each column of A has length 1 and is orthogonal to each other column of A.

Quadratic Forms

Suppose A is an $n \times n$ matrix. The function

$$g(x) = x^T Ax$$

is called a quadratic form. Now

$$\begin{aligned} g(x) &= \sum_{ij} A_{ij} x_i x_j \\ &= \sum_i A_{ii} x_i^2 + \sum_{i<j} (A_{ij} + A_{ji}) x_i x_j \end{aligned}$$

so that $g(x)$ depends only on the total $A_{ij} + A_{ji}$. In fact

$$x^T A x = x^T A^T x = x^T \left(\frac{A + A^T}{2} \right) x$$

Thus we will assume that A is symmetric.

Eigenvalues and eigenvectors

If A is $n \times n$ and $v \neq 0 \in R^n$ and $\lambda \in R$ are such that

$$Av = \lambda v$$

then we say that λ is an **eigenvalue** (or characteristic value or latent value) of A and that v is the corresponding **eigenvector**. Since $Av - \lambda v = (A - \lambda I)v = 0$ we find that $A - \lambda I$ must be singular. Therefore $\det(A - \lambda I) = 0$. Conversely if $A - \lambda I$ is singular then there is a $v \neq 0$ such that $(A - \lambda I)v = 0$. In fact, $\det(A - \lambda I)$ is a polynomial function of λ of degree n. Each root is an eigenvalue. For general A the roots could be multiple roots or complex valued.

Diagonalization

A matrix A is **diagonalized** by a non-singular matrix P is $P^{-1}AP \equiv D$ is a diagonal matrix. If so then $AP = PD$ and each column of P is an eigenvector of A with the ith column having eigenvalue D_{ii}. Thus to be diagonalizable A must have n linearly independent eigenvectors.

Symmetric Matrices

If A is symmetric then

1. Every eigenvalue of A is real (not complex).

2. A is diagonalizable and the columns of P may be taken to be orthogonal to each other and of unit length. In other words, A is diagonalizable by an orthogonal matrix P; in symbols $P^T AP = D$. The diagonal entries in D are the eigenvalues of A.

3. If $\lambda_1 \neq \lambda_2$ are two eigenvalues of A and v_1 and v_2 are corresponding eigenvectors then

$$v_1^T A v_2 = v_1^T \lambda_2 v_2 = \lambda_2 v_1^T v_2$$

and

$$\begin{aligned}(v_1^T A v_2) &= (v_1^T A v_2)^T \\ &= v2^T A^T v_1 \\ &= v_2^T A v_1 \\ &= v_2^T \lambda_1 v_1 \\ &= \lambda_1 v_2^T v_1\end{aligned}$$

Since $(\lambda_1 - \lambda_2) v_1^t v_2 = 0$ and $\lambda_1 \neq \lambda_2$ we see that $v_1^T v_2 = 0$. In other words eigenvectors corresponding to distinct eigenvalues are orthogonal.

Orthogonal Projections

Suppose that S is a vector subspace of R^n and that $a_1, \ldots, a_m$ are a basis for S. Given any $x \in R^n$ there is a unique $y \in S$ which is closest to x. That is, y minimizes

$$(x-y)^T(X-y)$$

over $y \in S$. Any y in S is of the form

$$y = c_1 a_1 + \cdots + c_m a_m = Ac$$

where A is the $n \times m$ matrix with columns $a_1, \ldots, a_m$ and c is the column vector with ith entry c_i. Define

$$Q = A(A^T A)^{-1} A^T$$

(The fact that A has rank m guarantees that $A^T A$ is invertible.) Then

$$\begin{aligned}(x-Ac)^T(x-Ac) &= (x-Qx+Qx-Ac)^T(x-Qx+Qx-Ac)\\ &= (x-Qx)^T(x-Qx)+(Qx-Ac)^T(x-Qx)\\ &\quad + (x-Qx)^T(Qx-Ac)+(Qx-Ac)^T(Qx-Ac)\end{aligned}$$

Note that $x - Qx = (I-Q)x$ and that

$$QAc = A(A^T A)^{-1} A^T Ac = Ac$$

so that

$$Qx - Ac = Q(x-Ac)$$

Then

$$(Qx-Ac)^T(x-Qx) = (x-Ac)^T Q^T (I-Q)x$$

Since $Q^T = Q$ we see that

$$\begin{aligned}Q^T(I-Q) &= Q(I-Q)\\ &= Q - Q^2\\ &= Q - A(A^T A)^{-1} A^T A (A^T A)^{-1} A^T\\ &= Q - Q\\ &= 0\end{aligned}$$

This shows that

$$(x-Ac)^T(x-Ac) = (x-Qx)^T(x-Qx) + (Qx-Ac)^T(Qx-Ac)$$

Now to choose Ac to minimize this quantity we need only minimize the second term. This is achieved by making $Qx = Ac$. Since $Qx = A(A^T A)^{-1} A^T x$ this can be done by taking $c = (A^T A)^{-1} A^T x$. In summary we find that the closest point y in S is

$$y = Qx = A(A^T A)^{-1} A^T x$$

We call y the orthogonal projection of x onto S.

Notice that the matrix Q is idempotent:

$$Q^2 = Q$$

We call Qx the orthogonal projection of x on S because Qx is perpendicular to the residual $x - Qx = (I - Q)x$.

Partitioned Matrices

Suppose that A_{11} is a $p \times r$ matrix, $A_{1,2}$ is $p \times s$, $A_{2,1}$ is $q \times r$ and $A_{2,2}$ is $q \times s$. Then we could make a big $(p+q) \times (r+s)$ matrix by putting together the A_{ij} in a 2 by 2 matrix giving the following picture:

$$A = \left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right]$$

For instance if

$$A_{11} = \left[\begin{array}{cc} 1 & 0 \\ 0 & 1 \end{array}\right]$$

$$A_{12} = \left[\begin{array}{c} 2 \\ 3 \end{array}\right]$$

$$A_{21} = \left[\begin{array}{cc} 4 & 5 \end{array}\right]$$

and

$$A_{22} = [6]$$

then

$$A = \left[\begin{array}{cc|c} 1 & 0 & 2 \\ 0 & 1 & 3 \\ \hline 4 & 5 & 6 \end{array}\right]$$

where I have drawn in lines to indicate the partitioning.

We can work with partitioned matrices just like ordinary matrices always making sure that in products we never change the order of multiplication of things.

$$\left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right] + \left[\begin{array}{cc} B_{11} & B_{12} \\ B_{21} & B_{22} \end{array}\right] = \left[\begin{array}{cc} A_{11}+B_{11} & A_{12}+B_{12} \\ A_{21}+B_{21} & A_{22}+B_{22} \end{array}\right]$$

and

$$\left[\begin{array}{cc} A_{11} & A_{12} \\ A_{21} & A_{22} \end{array}\right] \left[\begin{array}{cc} B_{11} & B_{12} \\ B_{21} & B_{22} \end{array}\right] = \left[\begin{array}{cc} A_{11}B_{11}+A_{12}B_{21} & A_{11}B_{12}+A_{12}B_{22} \\ A_{21}B_{11}+A_{22}B_{21} & A_{21}B_{12}+A_{22}B_{22} \end{array}\right]$$

In these formulas the partitioning of A and B must match. For the addition formula the dimensions of A_{ij} and B_{ij} must be the same. For the multiplication formula A_{12} must have as many columns as B_{21} has rows and so on. In general A_{ij} and B_{jk} must be of the right size for $A_{ij}B_{jk}$ to make sense for each i, j, k.

The technique can be used with more than a 2 by 2 partitioning.

Definition: A **block diagonal** matrix is a partitioned matrix A with pieces A_{ij} for which $A_{ij} = 0$ if $i \neq j$. If

$$A = \left[\begin{array}{cc} A_{11} & 0 \\ 0 & A_{22} \end{array} \right]$$

then A is invertible if and only if each A_{ii} is invertible and then

$$A^{-1} = \left[\begin{array}{cc} A_{11}^{-1} & 0 \\ 0 & A_{22}^{-1} \end{array} \right]$$

Moreover $\det(A) = \det(A_{11})\det(A_{22})$. Similar formulas work for larger matrices.