STAT 830: Statistical Theory I

Richard Lockhart

September 5, 2013

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 $A^c = \{\omega \in \Omega : \omega \notin A\} \in \mathcal{F}$.
  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

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

for all $(x_1, \ldots, x_p) \in R^p$.
Jargon and notation: we write $P(X \in A)$ for $P(\{\omega \in \Omega : X(\omega) \in A\})$ and define the **distribution** of X to be the map

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

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

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

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. $0 \leq F(x) \leq 1$.
2. $x > y \Rightarrow F(x) \geq F(y)$ (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

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

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

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

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 (**??**) 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 $X_1, \ldots, X_p$ are **independent** if, for any $A_1, \ldots, A_p$, 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*

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

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

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*

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

and

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

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

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

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

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

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

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

The inside integral is just

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

so

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

Assembling the results we get

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

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

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

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 $0 \leq X_1 \leq X_2 \leq \cdots$ 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 $P(X_{n+1} \geq X_n) = 1$ and $P(X_1 \geq 0) = 1$.

Theorem 5 (Dominated Convergence) *If $|X_n| \leq Y_n$ 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

$$g(x) = \int yf(y|x)dy$$

satisfies these definitions.

Proof:

$$\begin{aligned}
E(R(X)g(X)) &= \int R(x)g(x)f_X(x)dx \\
&= \int R(x) \int yf(y|x)dyf_X(x)dx \\
&= \int \int R(x)yf_X(x)f(y|x)dydx \\
&= \int \int R(x)yf_{X,Y}(x, y)dydx \\
&= E(R(X)Y)
\end{aligned}$$

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:

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

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:

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

- 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

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

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 Saddle-point 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 $\text{Beta}(2, 3)$:

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

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

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

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 $-\pi < \theta \leq \theta$. 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 **??**

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

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

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

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

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

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

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

where $y = A^tx$. 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
$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix},$$
$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$
and
$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$
then $X \sim MVN(\mu, \Sigma) \Rightarrow X_1 \sim MVN(\mu_1, \Sigma_{11})$.

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:

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

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

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

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

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

$$\begin{bmatrix} \frac{\cos\theta_1}{2R} & -R\sin\theta_1 & 0 \\ \frac{\sin\theta_1\cos\theta_2}{2R} & R\cos\theta_1\cos\theta_2 & -R\sin\theta_1\sin\theta_2 \\ \frac{\sin\theta_1\sin\theta_2}{2R} & R\cos\theta_1\sin\theta_2 & R\sin\theta_1\cos\theta_2 \end{bmatrix} .$$

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.

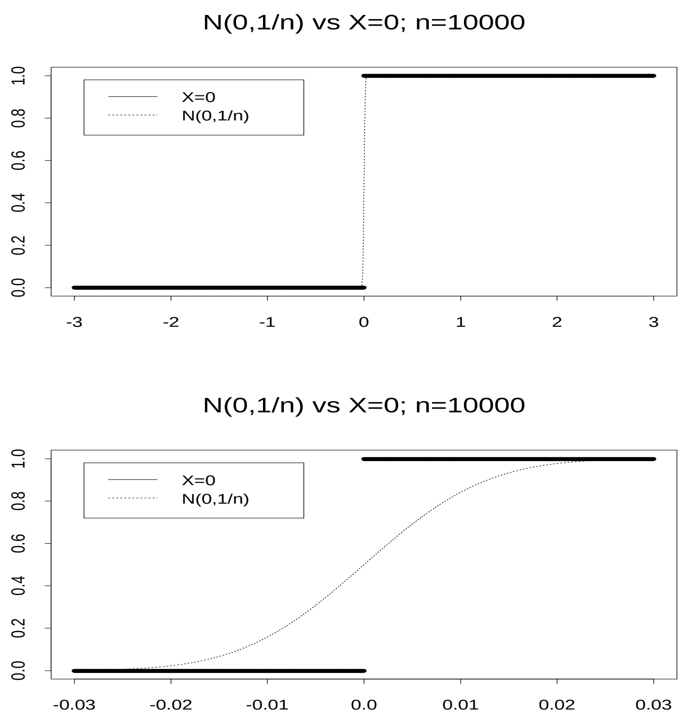


Figure 5.2: Comparison of the $N(0, 1/n)$ distribution and the $N(1/n, 1/n)$ distribution.

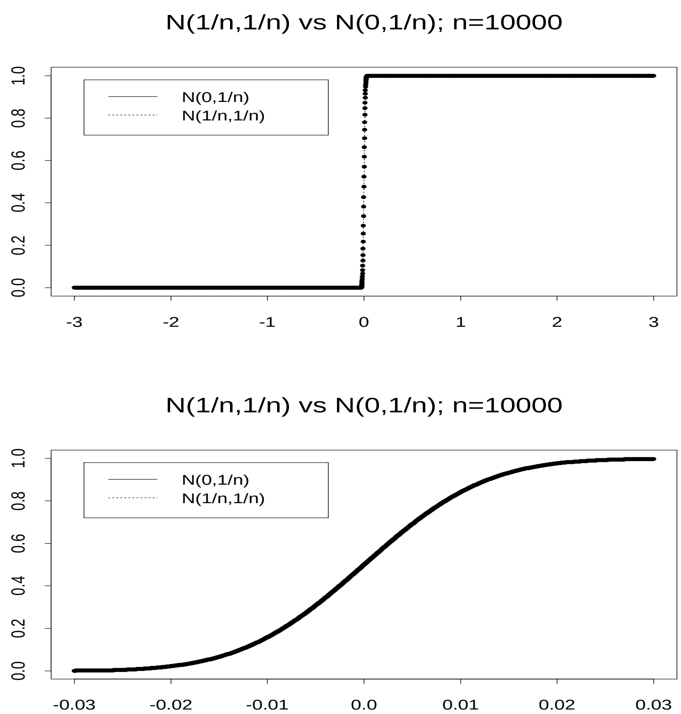
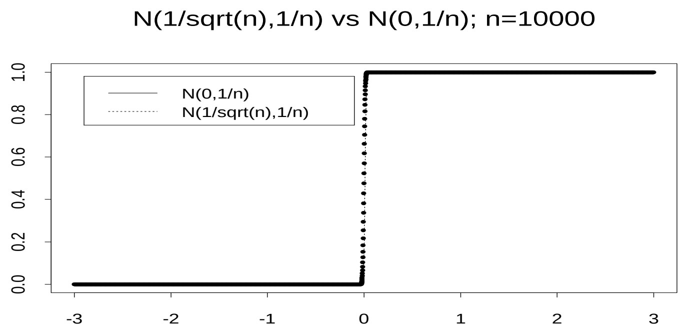
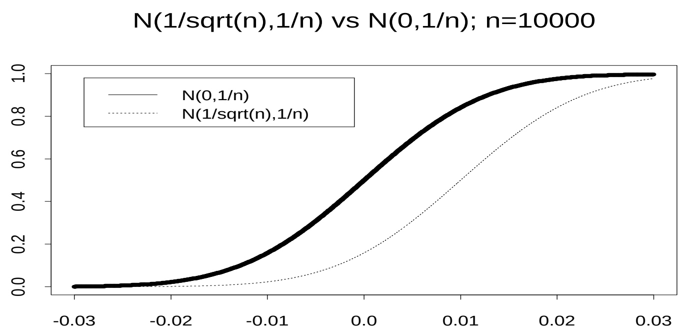


Figure 5.3: Comparison of the $N(n^{-1/2}, 1/n)$ distribution and the $N(0, 1/n)$ distribution.





- 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

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

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

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

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

which converges in distribution to $(1,-2\mu)Y$. 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

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

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

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

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

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

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.