

Probability and Statistics: An Introduction in One-Dimension

Matthew Hoelle
Econ 897.1
July 2007

Note: The reference for this material will be
*Casella and Berger. Statistical Inference, Second Edition. Duxbury
(2002). Chapters 1-2 (mostly)*

1 Basics of Probability

We will use the basic tenets of set theory as discussed last week as the starting point for probability. We are interested in assigning probability to 'events' within the sample space S . Thus, we define axiomatically a collection of events that will allow us to assign probability to events, complements of events, and unions and intersections of events.

Definition 1 A collection of subsets of S is called a sigma algebra (or Borel field), denoted by \mathcal{B} , if it satisfies the three properties:

- a. $\emptyset \in \mathcal{B}$
- b. If $A \in \mathcal{B}$, then $A^C \in \mathcal{B}$ (closed under complementation)
- c. If $A_1, A_2, \dots \in \mathcal{B}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$ (closed under countable unions)

Note: Some of you may be familiar with measure theory, but we will keep the exposition simpler by not introducing it at this time. I believe Professor Felix Kubler will touch on measure theory in 898 and any measure theory that you may need for the spring Macro courses will be covered at that time.

Example 2 If S is finite or countable, then the sigma algebra \mathcal{B} is the power set. If S has n elements, then \mathcal{B} has 2^n sets.

Example 3 If $S = \mathbb{R}$. Let \mathcal{I} be the set of all open intervals in \mathbb{R} . Then the important sigma algebra \mathcal{B} is defined as

$$\mathcal{B} = \cap \{ \mathcal{B}_0 : \mathcal{I} \subset \mathcal{B}_0 \text{ and } \mathcal{B}_0 \text{ is a sigma algebra} \}$$
$$\mathcal{B} \text{ contains sets of the form } [a, b], (a, b], (a, b), \text{ and } [a, b) \text{ for all real numbers } a, b$$

With the sample space S and the sigma algebra \mathcal{B} , we can axiomatically define the third component of our probability triplet, the probability function $P(\cdot)$.

Definition 4 Given a probability space S and associated sigma algebra \mathcal{B} , a probability function is a function $P(\cdot)$ with domain \mathcal{B} if it satisfies the three properties:

1. $P(A) \geq 0$ for all $A \in \mathcal{B}$
2. $P(S) = 1$
3. If $A_1, A_2, \dots \in \mathcal{B}$ are pairwise disjoint, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

Exercise #1 (taken from Casella and Berger exercise 1.11)

Let S be a sample space.

(a) Show that the collection $\mathcal{B} = \{\emptyset, S\}$ is a sigma algebra.

(b) Let $\mathcal{B} = \{\text{all subsets of } S, \text{ including } S \text{ itself}\}$. Show that \mathcal{B} is a sigma algebra.

(c) Show that the intersection of two sigma algebras is a sigma algebra.

Number of possible arrangements of size r from n objects:

	w/o replacement	with replacement
ordered	$\frac{n!}{(n-r)!}$	n^r
unordered	$\binom{n}{r}$	$\binom{n+r-1}{r}$

Definition 5 For $n \geq r$, $\binom{n}{r} = \frac{n!}{r!(n-r)!}$

Intuition: Ordered without replacement

Think of a lottery situation where 6 numbers must be in the correct order to win and none of the 44 possible numbers can be repeated.

$$\begin{aligned} \# \text{ Possibilities (\# possible tickets)} &= 44 \times 43 \times 42 \times 41 \times 40 \times 39 \\ &= \frac{44!}{38!} \end{aligned}$$

Intuition: Ordered with replacement

Now, the lottery situation is the same as above, but the numbers can be repeated.

$$\begin{aligned} \# \text{ Possibilities} &= 44 \times 44 \times 44 \times 44 \times 44 \times 44 \\ &= 44^6 \end{aligned}$$

Intuition: Unordered without replacement

The lottery is the same as the first case, but now a winning ticket is any combination of the distinct numbers. We already know that there are $\frac{44!}{38!}$ possibilities if order matters, but in this situation $\{1, 2, 3, 4, 5, 6\}$ is equivalent to $\{2, 1, 3, 4, 5, 6\}$, etc. How many combinations of six numbers are there?

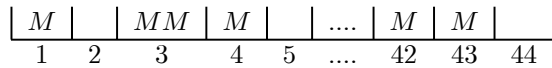
combinations of six numbers = 6!

Thus, the number of possibilities is smaller by this factor

$$\# \text{ Possibilities} = \frac{44!}{38!6!} = \binom{44}{6}$$

Intuition: Unordered with replacement

The lottery is the same as the third case, except now numbers can be repeated. Think of placing 6 markers across the 44 possible numbers.



We can thus think of the ordering of the 43 inner bin walls (black vertical lines) and the 6 markers. There are 49! ways to place this combination of walls and markers and since each of the 43 walls is indistinguishable and the order of the 6 markers does not matter, then the total number of possibilities is given by

$$\# \text{ Possibilities} = \frac{49!}{43!6!} = \binom{49}{6}$$

Example 6 Consider the word 'economics'. How many possible ways can the 9 letters in the word be organized?

There are 9! ways to order 9 objects, but the letters are not all distinct. The 2 'o's and 2 'c's lead to the solution:

$$\# \text{ Possibilities} = \frac{9!}{2!2!}$$

Note: This is related to the multinomial distribution.

Example 7 Consider a 5-card poker hand from a traditional deck of 52 cards. What is the probability that the best hand to play is one pair?

Thinking

13 possible denominations for the pair (2,3,Ace, etc.)

With only a pair, could have a spade matching a heart, spade matching a club, etc.

So, to choose 2 suits out of 4 allows for $\binom{4}{2}$ possibilities (unordered without replacement)

Now, we do not have a better hand than one pair, so we need to choose the 3 remaining cards in the hand from the 12 remaining denominations without allowing for an additional pair.

$$\# \text{ Possibilities} = \binom{12}{3} \text{ (unordered without replacement)}$$

Finally, we do not care about the suit of these 3 cards since a flush cannot occur along with a pair, so there are 4^3 ways to have 4 suits among 3 cards of different denominations (ordered with replacement)

Obviously, there are $\binom{52}{5}$ possible hands in poker (unordered without replacement), so

$$P(\text{one pair}) = \frac{13 \binom{4}{2} \binom{12}{3} 4^3}{\binom{52}{5}} = 0.423$$

Exercise #2 (Flush vs. Straight)

Why in a 5-card poker game does a flush beat a straight, i.e. show that a flush is probabilistically harder to draw than a straight.

1.1 Conditional Probabilities

Definition 8 If A and B are events in S , and $P(B) > 0$, then the conditional probability of A given B , written $P(A|B)$, is $P(A|B) = \frac{P(A \cap B)}{P(B)}$

Example 9 Consider the following scenario. A game show contestant can pick one of three doors numbered 1-3. Behind one of the doors is a dream vacation, while the other doors hide gag prizes of no value. The contestant picks a door. Then the host reveals the gag prize behind one of the un-picked doors. The contestant is then given the option to switch doors. Is switching a smart move probabilistically?

The initial probability of guessing the door correctly is obviously $1/3$.

Conditional on knowing that a certain door does not hide the dream vacation, the conditional probability of guessing the door correctly is

$$\begin{aligned} P(\text{correct} \mid \text{knowledge of a wrong door}) &= \\ P(\text{door } x \text{ is correct} \mid \text{door } y \neq x \text{ is wrong}) &= \\ = \frac{P(\text{door } x \text{ is correct})}{P(\text{door } y \neq x \text{ is wrong})} &= \frac{1/3}{2/3} = \frac{1}{2} \end{aligned}$$

Thus, it is advantageous for the contestant to switch doors.

1.2 Bayes' Rule

Theorem 10 Let A_1, A_2, \dots be a partition of the sample space, and let B be any set. Then for each $i = 1, 2, \dots$ $P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}$

Example 11 In a certain factory, Machines I, II, and III are all producing springs of the same length. Machines I, II, and III produce 1%, 4%, and 2% defective springs, respectively. Of the total production of springs in the factory, Machine I produces 30%, Machine II produces 25%, and Machine III produces 45%.

(i) If one spring is selected at random, determine the probability that it is defective.

(ii) Given that the selected spring is defective, find the conditional probability that it was produced by Machine II.

- (i) $P(\text{defective}) = (.01)(.3) + (.04)(.25) + (.02)(.45) = 0.022$
(ii) $P(\text{Machine II} \mid \text{defective}) = \frac{(.04)(.25)}{P(\text{defective})} = \frac{0.01}{0.022} = 0.4545$

Exercise #3 (Color Blind)

Suppose that 5% of men and .25% of women are color blind. A person is chosen at random and that person is color-blind. What is the probability that the person is male? (Assume males and females to be in equal numbers.)

Definition 12 Two events, A and B , are statistically independent if $P(A \cap B) = P(A)P(B)$.

Theorem 13 If A and B are independent events, then the following pairs are also independent:

- a. A and B^C
b. A^C and B
c. A^C and B^C

Proof (a):
$$\begin{aligned} P(A \cap B^C) &= P(A) - P(A \cap B) \\ &\quad (\text{since } P(A) = P(A \cap B^C) \cup P(A \cap B)) \\ &= P(A) - P(A)P(B) \\ &= P(A)(1 - P(B)) \\ &= P(A)P(B^C) \end{aligned}$$

Exercise #4 (Finish proof)
Prove (c).

2 Random Variables

Definition 14 Consider the sample space S . A function X , which assigns to each element $s \in S$ one and only one real number $X(s) = x$ is called a random variable.

Notation 15 Typically, an upper case letter refers to a random variable and a lower case variable refers to the realization of the random variable.

Example 16 $S = \{\text{results in terms of heads, tails when a coin is tossed 5 times}\}$

Define $X = \text{number of heads}$
let $s_1 = \{H, H, T, H, H\}$ $s_2 = \{H, T, H, H, H\}$
 $X(s_1) = x$ $X(s_2) = x$ here $x = 4$

2.1 Distribution functions

Definition 17 The cumulative distribution function or cdf of a random variable X , denoted by $F_X(x)$, is defined by $F_X(x) = P_X(X \leq x)$, for all x

Theorem 18 The function $F_X(x)$ is a cdf if and only if the following three conditions hold:

- $\lim_{x \rightarrow -\infty} F_X(x) = 0$ and $\lim_{x \rightarrow \infty} F_X(x) = 1$
- $F_X(x)$ is a non-decreasing function of x
- $F_X(x)$ is right-continuous; that is, for every number x_0 ,
 $\lim_{x \downarrow x_0} F_X(x) = F_X(x_0)$

Exercise #5 (Finish proof)

Prove the necessity of these three properties.

Example 19 Let X be the lifetime of a mechanical part in years. Assume that X has a cdf $F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - e^{-x} & x \geq 0 \end{cases}$

The probability that the lifetime of the part is between 1 and 3 years is given by $P(1 < X \leq 3) = F_X(3) - F_X(1) = e^{-1} - e^{-3}$

Definition 20 A random variable X is continuous if $F_X(x)$ is a continuous function of x . A random variable X is discrete if $F_X(x)$ is a step function of x .

Identically distributed - 2 equivalent definitions

$$X \stackrel{i.d.}{\sim} Y \quad \text{iff} \quad \forall A \in \mathcal{B}, P(X \in A) = P(Y \in A)$$

$$\Updownarrow$$

$$F_X(x) = F_Y(x) \quad \forall x$$

2.2 Discrete random variables

Definition 21 The probability mass function (pmf) of a discrete random variable X is given by $f_X(x) = P(X = x)$ for all x

Transformation

If X is a random variable and $g(X)$ is any function of X , then $g(X)$ is also a random variable. We will denote the sample space of X as \mathcal{X} and the sample space of $g(X)$ as \mathcal{Y} . Then $g(x) : \mathcal{X} \rightarrow \mathcal{Y}$, where \mathcal{Y} is the sample space of $g(X)$. Notice the implications: X discrete $\Rightarrow \mathcal{X}$ countable $\Rightarrow \mathcal{Y}$ countable $\Rightarrow g(X)$ discrete random variable.

If we define the inverse mapping $g^{-1}(A) = \{x \in \mathcal{X} : g(x) \in A\}$, then $P(g(X) \in A) = P(X \in g^{-1}(A))$. Thus if we set $Y = g(X)$ for simplicity, then $f_Y(y) = P(Y = y) = \sum_{x \in g^{-1}(y)} P(X = x) = \sum_{x \in g^{-1}(y)} f_X(x)$ for $y \in \mathcal{Y}$

Example 22 We will illustrate how to use the pmf of the random variable X to get the pmf of the random variable $Y = g(X)$. Let the discrete random variable X have the binomial distribution (denoted $X \sim \text{binomial}(n, p)$), i.e. $f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \dots, n$. (Think of p as the probability of a coin landing on heads and X as the # of heads in n tosses.) Define $Y = g(X)$ where $g(x) = n - x$. Then $f_Y(y) = \sum_{x \in g^{-1}(y)} f_X(x)$ notice that g is one-to-one, so the inverse mapping g^{-1} is a function

$$\begin{aligned} f_Y(y) &= f_X(n - y) = \binom{n}{n - y} p^{n-y} (1-p)^{n-(n-y)} \\ &= \binom{n}{y} (1-p)^y p^{n-y} \\ \text{so } Y &\sim \text{binomial}(n, 1-p) \end{aligned}$$

2.3 Continuous random variables

Definition 23 The probability density function or pdf, $f_X(x)$, of a continuous random variable X is a function that satisfies $F_X(x) = \int_{-\infty}^x f_X(t) dt$ for all x

Using the Fundamental Theorem of Calculus, if $f_X(x)$ is continuous, then $f_X(x) = \frac{d}{dx} F_X(x)$

Theorem 24 A function $f_X(x)$ is a pdf (or pmf) of a random variable X iff

- a. $f_X(x) \geq 0$ for all x
- b. for a pmf $\sum_{-\infty}^{\infty} f_X(x) = 1$
for a pdf $\int_{-\infty}^{\infty} f_X(x) = 1$

Transformation

Theorem 25 Let X have cdf $F_X(x)$, let $Y = g(X)$ and let \mathcal{X} and \mathcal{Y} be defined such that $\mathcal{X} = \{x : f_X(x) > 0\}$ and $x \mapsto g(x)$ is onto, i.e. $\mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}$.

- a. If g is an increasing function on \mathcal{X} , $F_Y(y) = F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$
- b. If g is a decreasing function on \mathcal{X} and X is a continuous random variable, $F_Y(y) = 1 - F_X(g^{-1}(y))$ for $y \in \mathcal{Y}$

Intuition for a: $x \mapsto g(x)$ is a bijection, thus $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y))$

Intuition for b: $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y))$
 (since g is decreasing)
 $= 1 - P(X \leq g^{-1}(y))$
 (for X being a continuous random variable)

Theorem 26 Let X have pdf $f_X(x)$ and let $Y = g(X)$, where g is a monotone function. Let \mathcal{X} and \mathcal{Y} be defined by $\mathcal{X} = \{x : f_X(x) > 0\}$ and $\mathcal{Y} = \{y : y = g(x) \text{ for some } x \in \mathcal{X}\}$. Suppose that $f_X(x)$ is continuous on \mathcal{X} and that $g^{-1}(y)$ has a continuous derivative on \mathcal{Y} . Then the pdf of Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

Proof: if g is increasing, from Theorem above, $F_Y(y) = F_X(g^{-1}(y))$
 by the Chain Rule, $f_Y(y) = \frac{d}{dy} F_Y(y) = f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y)$
 if g is decreasing, from Theorem above, $F_Y(y) = 1 - F_X(g^{-1}(y))$
 by the Chain Rule, $f_Y(y) = \frac{d}{dy} F_Y(y) = -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) =$
 $f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$

Example 27 (Normal-chi squared relationship)

Let X have the standard normal distribution (as we will see later, this is given by:

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad -\infty < x < \infty$$

Consider the transformation $Y = X^2$. $g(x) = x^2$ is monotone on $(-\infty, 0)$ and $(0, \infty)$

$$\begin{aligned} \mathcal{Y} &= (0, \infty) \\ \text{for } \mathcal{X}_1 &= (-\infty, 0) & g^{-1}(y) &= -\sqrt{y} \\ \text{for } \mathcal{X}_2 &= (0, \infty) & g^{-1}(y) &= \sqrt{y} \end{aligned}$$

From the Theorem above,

$$\begin{aligned} f_Y(y) &= \frac{1}{\sqrt{2\pi}} e^{-(-\sqrt{y})^2/2} \left| -\frac{1}{2\sqrt{y}} \right| + \frac{1}{\sqrt{2\pi}} e^{-(\sqrt{y})^2/2} \left| \frac{1}{2\sqrt{y}} \right| \\ f_Y(y) &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2} \quad 0 < y < \infty \end{aligned}$$

This distribution is the chi-squared (with 1 degree of freedom). This is denoted $Y \sim \chi_1^2$

Exercise #6 (Cauchy distribution)

Let X have the uniform pdf $f_X(x) = \frac{1}{\pi}$ for $-\frac{\pi}{2} < x < \frac{\pi}{2}$. Find the pdf of $Y = \tan(X)$. This is the pdf of a Cauchy distribution.

Theorem 28 Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(x)$. Then Y is uniformly distributed on $(0, 1)$, that is $P(Y \leq y) = y$, $0 < y < 1$.

Proof: First, define the inverse as $F_X^{-1}(y) = \inf\{x : F_X(x) = y\}$. For $y = F_X(x)$, we have $0 < y < 1$

$$\begin{aligned} P(Y \leq y) &= P(F_X(X) \leq y) = P(F_X^{-1}[F_X(x)] \leq F_X^{-1}(y)) \\ &= P(X \leq F_X^{-1}(y)) \quad (\text{is this true?}) \end{aligned}$$

for flat regions (i.e. $\exists x \in (x_1, x_2)$ s.t. $F_X(x) = F_X(x_1) = F_X(x_2)$)

then by definition, $F_X^{-1}[F_X(x)] = x_1$

but $P(x_1 \leq F_X^{-1}(y)) = P(x \leq F_X^{-1}(y))$ since $f_X(x') = 0$ for $x' \in [x_1, x_2]$

$$\begin{aligned} P(X \leq F_X^{-1}(y)) &= F_X(F_X^{-1}(y)) \quad \text{by definition of } F_X \\ &= y \quad \text{by continuity of } F_X \end{aligned}$$

2.4 Expectation

Definition 29 The expected value or mean of a random variable $g(X)$, denoted

$$\text{by } Eg(X), \text{ is } Eg(X) = \begin{cases} \int_{-\infty}^{\infty} g(x)f_X(x)dx & \text{if } X \text{ is cont.} \\ \sum_{x \in \mathcal{X}} g(x)f_X(x) & \text{if } X \text{ is discrete} \end{cases}$$

provided that $E|g(X)| < \infty$

Example 30 Recall the pdf of a Cauchy random variable X

$$\begin{aligned} f_X(x) &= \frac{1}{\pi} \frac{1}{1+x^2} \quad -\infty < x < \infty \\ E|X| &= \int_{-\infty}^{\infty} \frac{|x|}{\pi} \frac{1}{1+x^2} dx = \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2} dx \end{aligned}$$

For any $M > 0$,

$$\int_0^M \frac{x}{1+x^2} dx = \left[\frac{\log(1+x^2)}{2} \right]_0^M = \frac{\log(1+M^2)}{2}$$

$$E|X| = \lim_{M \rightarrow \infty} \frac{1}{\pi} \log(1+M^2) = \infty$$

so EX does not exist.

Theorem 31 Let X be a random variable and let a, b , and c be constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist,

- $E(ag_1(X) + bg_2(X) + c) = aEg_1(X) + bEg_2(X) + c$
- If $g_1(x) \geq 0 \quad \forall x$, then $Eg_1(X) \geq 0$
- If $g_1(x) \geq g_2(x) \quad \forall x$, then $Eg_1(X) \geq Eg_2(X)$
- If $a \leq g_1(x) \leq b \quad \forall x$, then $a \leq Eg_1(X) \leq b$

Proof (a): Consider the continuous case (discrete case is similar)

$$\begin{aligned} E(ag_1(X) + bg_2(X) + c) &= \int_{-\infty}^{\infty} (ag_1(x) + bg_2(x) + c) f_X(x) dx \\ &= \int_{-\infty}^{\infty} ag_1(x) f_X(x) dx + \int_{-\infty}^{\infty} bg_2(x) f_X(x) dx + \int_{-\infty}^{\infty} c f_X(x) dx \\ &= a \int_{-\infty}^{\infty} g_1(x) f_X(x) dx + b \int_{-\infty}^{\infty} g_2(x) f_X(x) dx + c \int_{-\infty}^{\infty} f_X(x) dx \\ &= aEg_1(X) + bEg_2(X) + c \end{aligned}$$

Exercise #7 (Finish proof- properties of expectation)

Last name A-H, prove (b)

Last name I-Q, prove (c)

Last name R-Z, prove (d)

2.5 Moment generating functions

Definition 32 The n^{th} moment of X is EX^n

The n^{th} central moment of X is $E(X - \mu)^n$ where $\mu = EX$

Definition 33 The variance of a random variable X is its second central moment, $\text{Var}X = E(X - EX)^2$. The standard deviation of X is the positive square root of $\text{Var}X$.

Fact 1: $\text{Var}(a + bX) = a^2\text{Var}X$ for constants a, b

$$\begin{aligned} \text{Proof: } \text{Var}(a + bX) &= E(aX + b - aEX - b)^2 \\ &= a^2E(X - EX)^2 = a^2\text{Var}X \end{aligned}$$

Fact 2: $\text{Var}X = EX^2 - (EX)^2$

$$\begin{aligned} \text{Proof: } \text{Var}X &= E(X - EX)^2 = E[X^2 - 2XEX + (EX)^2] \\ &= EX^2 - 2(EX)^2 + (EX)^2 = EX^2 - (EX)^2 \end{aligned}$$

where we use the fact $E[aX] = aE[X]$ since $a = 2EX$ is a constant

Definition 34 Let X be a random variable with cdf F_X . The moment generating function (mgf) of X , denoted by $M_X(t)$ is $M_X(t) = Ee^{tX}$ provided that the expectation exists for t in some neighborhood of 0 (that is, $\exists h > 0$ s.t. $\forall t \in (-h, h)$, Ee^{tX} exists).

Example 35 Recall that the series $\frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \dots$ converges to $\frac{\pi^2}{6}$. Consider

$$\text{the distribution } f_X(x) = \begin{cases} \frac{6}{\pi^2 x^2} & x = 1, 2, 3 \\ 0 & \text{elsewhere} \end{cases}$$

$$M_X(t) = E(e^{tX}) = \sum_x e^{tx} f_X(x) = \sum_{x=1}^{\infty} \frac{6e^{tx}}{\pi^2 x^2}$$

$E|e^{tX}| \rightarrow \infty$ for any $t > 0$, thus Ee^{tX} does not exist, so this pmf does not have a mgf.

Theorem 36 Let X and Y be random variables with moment generating functions M_X and M_Y , respectively, existing in open intervals around 0. Then $F_X(z) = F_Y(z)$ for all z iff $M_X(t) = M_Y(t)$ for all $t \in (-h, h)$ for some $h > 0$.

Fact: The mgf of the sum of independent random variables is the product of their mgfs.

Theorem 37 If X has mgf $M_X(t)$, then $EX^n = M_X^{(n)}(0)$ where we define $M_X^{(n)}(0) = \frac{d^n}{dt^n} M_X(t)|_{t=0}$

Before we get to this proof of the above theorem, we need to add 3 important results to our toolkit that are fundamental in probability theory.

2.6 Addition to the toolkit

Dominated Convergence, Monotone Convergence, and Fatou's Lemma

These results will be stated without proof and all involve exchanging limits and expectations (recall that a derivative is a limit and an expectation is an integral).

Theorem 38 *Dominated Convergence Theorem (DCT)*

If $P(X_n \rightarrow X) = 1$ and $|X_n| \leq Y$ for $1 \leq n < \infty$ where Y satisfies $E(Y) < \infty$, then $E(|X|) < \infty$ and $\lim_{n \rightarrow \infty} E(X_n) = E\left(\lim_{n \rightarrow \infty} X_n\right)$

Theorem 39 *Monotone Convergence Theorem (MCT)*

If $0 \leq X_n \leq X_{n+1}$ for all $n \geq 1$, then $\lim_{n \rightarrow \infty} E(X_n) = E\left(\lim_{n \rightarrow \infty} X_n\right)$ where the limits may take infinity as a possible value.

Theorem 40 *Fatou's Lemma*

If $0 \leq X_n$ for all $n \geq 1$, then $E\left(\liminf_{n \rightarrow \infty} X_n\right) \leq \liminf_{n \rightarrow \infty} E(X_n)$ where the limits may take infinity as a possible value.

Now, for the proof:

$$\begin{aligned} &\text{Using DCT (since } \exists Y \text{ large s.t. } |e^{tX_n}| \leq Y \quad \forall n \quad \text{and } E(Y) < \infty) \\ &\text{then } \frac{d^n}{dt^n} M_X(t) = \int_{-\infty}^{\infty} \frac{d^n}{dt^n} e^{tx} f_X(x) dx \\ &\quad = \int_{-\infty}^{\infty} x^n e^{tx} f_X(x) dx = E[X^n e^{tX}] \\ &\text{so } \frac{d^n}{dt^n} M_X(t)|_{t=0} = E[X^n] \end{aligned}$$

More generally,

Definition 41 *The characteristic function of the random variable X , denoted $\varphi_X(t)$, is defined by, $\varphi_X(t) = E[e^{itX}]$*

This expectation exists for every distribution

$$\begin{aligned} |E(e^{itX})| &= \left| \int_{-\infty}^{\infty} e^{itx} f_X(x) dx \right| \leq \int_{-\infty}^{\infty} |e^{itx} f_X(x)| dx \\ &\text{Notice } |e^{itx} f_X(x)| = |e^{itx}| f_X(x) \\ |e^{itx}| &= |\cos tx + i \sin tx| \quad \text{this is Euler's formula} \\ &\text{By definition of absolute value of complex number,} \\ |\cos tx + i \sin tx| &= \sqrt{\cos^2 tx + \sin^2 tx} = 1 \\ \text{So, } |\varphi_X(t)| &\leq \int_{-\infty}^{\infty} f_X(x) dx = 1 \end{aligned}$$

2.7 Another addition to the toolkit: Famous Inequalities

Markov, Chebychev, Jensen's, and Cauchy-Schwarz

Theorem 42 *Let X be a random variable and let m be a positive integer. Suppose $E[X^m]$ exists. If k is an integer and $k \leq m$, then $E[X^k]$ exists.*

Exercise #8 (Finish proof- existence of expectations)

Prove the above theorem.

Theorem 43 (*Markov's Inequality*)

Let $u(X)$ be a nonnegative function of the random variable X . If $E[u(X)]$ exists, then for every positive constant c , $P[u(X) \geq c] \leq \frac{E[u(X)]}{c}$

Proof: We consider continuous random variable X only.

Let $A = \{x : u(x) \geq c\}$.

$$E[u(X)] = \int_{-\infty}^{\infty} u(x)f_X(x)dx = \int_A u(x)f_X(x)dx + \int_{A^c} u(x)f_X(x)dx$$

Note that each integral is non-negative, thus

$$E[u(X)] \geq \int_A u(x)f_X(x)dx$$

Since $x \in A$, then $u(x) \geq c$ so

$$\int_A u(x)f_X(x)dx \geq c \int_A f_X(x)dx = cP(X \in A) = cP[u(X) \geq c]$$

$$\Rightarrow \frac{E[u(X)]}{c} \geq P[u(X) \geq c]$$

Theorem 44 (*Chebychev's Inequality*)

Let the random variable X have a probability distribution about which we will assume only that there is a finite variance σ^2 (this implies $\mu = E(X)$ exists). Then for every $k > 0$, $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$

Proof: Use Markov's inequality with $u(X) = (X - \mu)^2$ $c = k^2\sigma^2$

then, we have

$$P[(X - \mu)^2 \geq k^2\sigma^2] \leq \frac{E[(X - \mu)^2]}{k^2\sigma^2}$$

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

$$P[|X - \mu| \geq k\sigma] \leq \frac{1}{k^2} \quad (\text{an equivalent representation})$$

Theorem 45 (*Jensen's Inequality*)

If ϕ is convex on an open interval \mathcal{I} and X is a random variable whose support is contained in \mathcal{I} and has finite expectation, then $\phi[E(X)] \leq E[\phi(X)]$

Proof: For simplification, assume that ϕ has a second derivative, but in general only convexity is required. Expand $\phi(x)$ using Taylor series about $\mu = E[X]$:

$$\phi(x) = \phi(\mu) + \phi'(\mu)(x - \mu) + \frac{\phi''(\hat{x})(x - \mu)^2}{2} \quad \hat{x} \in [x, \mu]$$

by convexity, $\phi''(\hat{x}) \geq 0$ for any \hat{x}

$$\text{thus, } \frac{\phi''(\hat{x})(x - \mu)^2}{2} \geq 0$$

$\phi(x) \geq \phi(\mu) + \phi'(\mu)(x - \mu)$
 taking expectations yields
 $E[\phi(X)] \geq E[\phi(\mu)] + \phi'(\mu)E[X - \mu] = \phi(\mu) + 0$
 Thus, $E[\phi(X)] \geq \phi[E(X)]$

Exercise #9 (Application of inequality)

A random variable X is defined by $Z = \log X$ where $EZ = 0$. Is EX greater than, less than, or equal to 1?

Theorem 46 (Hölder's Inequality- generalization of Cauchy-Schwarz)

Let X and Y be any two random variables,
 and let p and q be s.t. $\frac{1}{p} + \frac{1}{q} = 1$ (*)

Then $E|XY| \leq (E|X|^p)^{1/p} (E|Y|^q)^{1/q}$

Proof: We will use the following lemma in this proof

$$\frac{1}{p}a^p + \frac{1}{q}b^q \geq ab$$

Subproof: Define $g(a) = \frac{1}{p}a^p + \frac{1}{q}b^q - ab$
 $g'(a) = 0 \Rightarrow a^{p-1} - b = 0 \Rightarrow b = a^{p-1}$

Is this a^* a minimum?

$$g''(a) = (p-1)a^{p-2} > 0$$

$$g^*(a) = \frac{1}{p}a^p + \frac{1}{q}(a^{p-1})^q - a^p$$

$$\text{from (*)} \quad \frac{1}{q} = 1 - \frac{1}{p} = \frac{p-1}{p}$$

$$q = \frac{p}{p-1}$$

$$g^*(a) = \frac{1}{p}a^p + \frac{p-1}{p}(a^{p-1})^{\frac{p}{p-1}} - a^p$$

$$= a^p \left(\frac{1}{p} + \frac{p-1}{p} - \frac{p}{p} \right) = 0$$

Thus, $g(a) \geq 0 \quad \forall a$

Define $a = \frac{|X|}{(E|X|^p)^{1/p}}$

$$b = \frac{|Y|}{(E|Y|^q)^{1/q}}$$

from the lemma:

$$\frac{1}{p} \frac{|X|^p}{(E|X|^p)} + \frac{1}{q} \frac{|Y|^q}{(E|Y|^q)} \geq \frac{|XY|}{(E|X|^p)^{1/p} (E|Y|^q)^{1/q}}$$

take expectations:

$$E(LHS) = 1$$

$$\text{thus, } E|XY| \leq (E|X|^p)^{1/p} (E|Y|^q)^{1/q}$$

Theorem 47 (Cauchy-Schwarz Inequality)

For any two random variables X and Y , $E|XY| \leq (E|X|^2)^{1/2} (E|Y|^2)^{1/2}$

Proof: From Hölder's Inequality, set $p = q = 1/2$

2.8 Convergence in probability and distribution

This section is a little digression from typical probability and statistics, but everyone should be exposed to this material before being thrown into a quick-paced first-year econometrics sequence.

Definition 48 A sequence of random variables $\{Z_n\}$ converges in probability to a constant α if for any $\epsilon > 0$, $\lim_{n \rightarrow \infty} P(|Z_n - \alpha| > \epsilon) = 0$. This is denoted $Z_n \xrightarrow{p} \alpha$.

Definition 49 A sequence of random variables $\{Z_n\}$ converges almost surely to a constant α if $P\left(\lim_{n \rightarrow \infty} Z_n = \alpha\right) = 1$. This is denoted $Z_n \xrightarrow{a.s.} \alpha$.

Note: If we can show that a sequence converges almost surely, then the sequence converges in probability (the convergence in probability being the typical result stated in theory).

Definition 50 A sequence of random variables $\{Z_n\}$ converges in mean square to α if $\lim_{n \rightarrow \infty} E[(Z_n - \alpha)^2] = 0$. This is denoted $Z_n \xrightarrow{m.s.} \alpha$.

Note: If we can show that a sequence converges in mean square, then the sequence converges in probability.

Definition 51 Let $\{Z_n\}$ be a sequence of random variables with cumulative distribution F_n . We say that $\{Z_n\}$ converges in distribution to a random vector Z with cumulative distribution F , if F_n converges to F at every continuity point of F . This is denoted $Z_n \xrightarrow{d} Z$.

Useful Result #1:

Suppose $a(\cdot)$ is continuous

$$(i) \quad Z_n \xrightarrow{p} \alpha \quad \Rightarrow \quad a(Z_n) \xrightarrow{p} a(\alpha)$$

$$(ii) \quad Z_n \xrightarrow{d} Z \quad \Rightarrow \quad a(Z_n) \xrightarrow{d} a(Z)$$

Useful Result #2 (Slutsky's Theorem):

$$(i) \quad X_n \xrightarrow{d} X, Y_n \xrightarrow{p} \alpha \quad \Rightarrow \quad X_n + Y_n \xrightarrow{d} X + \alpha$$

$$(ii) \quad X_n \xrightarrow{d} X, A_n \xrightarrow{p} A \quad \Rightarrow \quad A_n X_n \xrightarrow{d} AX \quad \text{provided}$$

that A_n and X_n are conformable

Useful Result #3 (Delta method):

Suppose $\{X_n\}$ is a sequence of random variables such that $X_n \xrightarrow{p} \beta$ and $\sqrt{n}(X_n - \beta) \xrightarrow{d} Z$ and suppose $a(\cdot)$ has continuous first derivatives $A(\beta) = \frac{\partial a(\beta)}{\partial \beta}$, then $\sqrt{n}[a(X_n) - a(\beta)] \xrightarrow{d} A(\beta)Z$

Proof: From the Mean-Value Theorem (remember this fella, right?)
 $a(X_n) - a(\beta) = A(Y_n)(X_n - \beta)$ for Y_n between X_n and β
 Multiply both sides by \sqrt{n}
 $\sqrt{n}[a(X_n) - a(\beta)] = A(Y_n)\sqrt{n}(X_n - \beta)$
 since $X_n \xrightarrow{p} \beta$ and Y_n is between X_n and β , then $Y_n \xrightarrow{p} \beta$
 from Useful Result #1, $A(Y_n) \xrightarrow{p} A(\beta)$
 since $\sqrt{n}(X_n - \beta) \xrightarrow{d} Z$, we apply Slutsky to yield
 $\sqrt{n}[a(X_n) - a(\beta)] \xrightarrow{d} A(\beta)Z$

Exercise #10 (Application of Delta Method)

Suppose we have a sequence of random variables $\{X_n\}$ s.t.

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n X_i - \mu \right) \xrightarrow{d} N(0, \sigma^2) \text{ where } \mu = E[X_n]$$

Find the limiting distribution of $\sqrt{n} \left(\frac{1}{\frac{1}{n} \sum_{i=1}^n X_i} - \frac{1}{\mu} \right)$.

3 Common Distributions

3.1 Discrete distributions of the random variable X

- Bernoulli (p)

$$P(X = x|p) = p^x(1-p)^{1-x} \quad x = 0, 1 \quad 0 \leq p \leq 1$$

$$EX = \sum_{x=0,1} xp^x(1-p)^{1-x} = p$$

$$Var X = EX^2 - (EX)^2$$

$$= \sum_{x=0,1} x^2 p^x(1-p)^{1-x} - p^2 = p - p^2 = p(1-p)$$

$$M_X(t) = E[e^{tX}] = \sum_{x=0,1} e^{tx} p^x(1-p)^{1-x} = (1-p) + pe^t$$

- Binomial (n, p)

$$P(X = x|n, p) = \binom{n}{x} p^x(1-p)^{n-x} \quad x = 0, 1, \dots, n \quad 0 \leq p \leq 1$$

$$M_X(t) = E[e^{tX}] = \sum_{x=0}^n e^{tx} \binom{n}{x} p^x(1-p)^{n-x}$$

$$= \sum_{x=0}^n \binom{n}{x} (pe^t)^x (1-p)^{n-x}$$

$$= 1(1-p)^n + n(pe^t)(1-p)^{n-1} + \frac{n(n-1)}{2}(pe^t)^2(1-p)^{n-2} + \dots$$

$$+ (pe^t)^n$$

Note: $[(1-p) + pe^t]^n = (1-p)^n + n(1-p)^{n-1}(pe^t)$

$$+ \frac{n(n-1)}{2}(1-p)^{n-2}(pe^t)^2 + \dots + (pe^t)^n$$

this intuition is the statement of the Binomial Theorem

thus, $M_X(t) = [(1-p) + pe^t]^n$

$$EX = M'_X(0) \quad EX^2 = M''_X(0)$$

$$M'_X(t) = n[(1-p) + pe^t]^{n-1} (pe^t)$$

$$EX = np$$

$$M''_X(t) = n(n-1)[(1-p) + pe^t]^{n-2} (pe^t)^2 + n[(1-p) + pe^t]^{n-1} (pe^t)$$

$$EX^2 = n(n-1)p^2 + np$$

$$VarX = n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p)$$

- Poisson (λ)

$$P(X = x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0, 1, \dots \quad 0 \leq \lambda < \infty$$

$$EX = \sum_{x=0}^{\infty} x \frac{e^{-\lambda}\lambda^x}{x!} = \sum_{x=1}^{\infty} \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} \lambda = 1 \times \lambda = \lambda$$

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{\infty} e^{xt} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{e^{xt}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t\lambda)^x}{x!} \quad \text{by definition, } e^z = \sum_{x=0}^{\infty} \frac{(z)^x}{x!} \end{aligned}$$

$$M_X(t) = e^{-\lambda} e^{\lambda e^t} = e^{\lambda(e^t-1)}$$

$$EX^2 = M''_X(0)$$

$$M'_X(t) = \lambda e^t e^{\lambda(e^t-1)}$$

$$M''_X(t) = (\lambda e^t)^2 e^{\lambda(e^t-1)} + \lambda e^t e^{\lambda(e^t-1)}$$

$$EX^2 = \lambda^2 + \lambda$$

$$VarX = \lambda^2 + \lambda - \lambda^2 = \lambda$$

Exercise #11 (Bernoulli into Binomial)

Show that the sum of n independent and identically distributed random variables $X_i \sim \text{Bernoulli}(p)$ has distribution $\sum_{i=1}^n X_i \sim \text{Binomial}(n, p)$.

Exercise #12 (Poisson into Poisson)

Show that the sum of n independent and identically distributed random variables $X_i \sim \text{Poisson}(\lambda_i)$ has distribution $\sum_{i=1}^n X_i \sim \text{Poisson}(\sum_{i=1}^n \lambda_i)$.

3.2 Continuous distributions of the random variable X

- Uniform (a, b)

$$f_X(x|a, b) = \frac{1}{b-a} \quad a \leq x \leq b$$

$$EX = \int_a^b \frac{1}{b-a} x dx = \frac{1}{b-a} \left[\frac{x^2}{2} \right]_a^b = \frac{(b-a)}{2}$$

$$EX^2 = \int_a^b \frac{1}{b-a} x^2 dx = \frac{1}{b-a} \left[\frac{x^3}{3} \right]_a^b = \frac{(b-a)^2}{3}$$

$$VarX = \frac{(b-a)^2}{3} - \frac{(b-a)^2}{4} = \frac{(b-a)^2}{12}$$

$$M_X(t) = \int_a^b \frac{1}{b-a} e^{xt} dx = \frac{1}{b-a} \left[\frac{1}{t} e^{xt} \right]_a^b = \frac{1}{t(b-a)} e^{t(b-a)}$$

- Exponential (β)

$$f_X(x|\beta) = \frac{1}{\beta} e^{-x/\beta} \quad 0 \leq x < \infty, \beta > 0$$

$$M_X(t) = \int_0^\infty e^{xt} \frac{1}{\beta} e^{-x/\beta} dx = \int_0^\infty \frac{1}{\beta} e^{-x(\frac{1}{\beta}-t)} dx \cdot \frac{\frac{1}{\beta}-t}{\frac{1}{\beta}-t}$$

$$= \frac{1/\beta}{\frac{1}{\beta}-t} \int_0^\infty \left(\frac{1}{\beta} - t \right) e^{-x(\frac{1}{\beta}-t)} dx = \frac{1/\beta}{\frac{1}{\beta}-t} \times 1 = \frac{1}{1-\beta t}$$

$$M'_X(t) = -(1-\beta t)^{-2} (-\beta)$$

$$M''_X(t) = -2(1-\beta t)^{-3} (\beta)(-\beta)$$

$$EX = M'_X(0) = \beta$$

$$EX^2 = M''_X(0) = 2\beta^2$$

$$VarX = 2\beta^2 - \beta^2 = \beta^2$$

- Weibull (γ, β)

$$f_X(x|\gamma, \beta) = \frac{\gamma}{\beta} x^{\gamma-1} e^{-x^\gamma/\beta} \quad 0 \leq x < \infty \quad \gamma > 0, \beta > 0$$

This is a generalization of the Exponential distribution, i.e. Weibull ($1, \beta$) = Exponential (β), and is used surprisingly often in econometrics.

$$EX = \beta^{1/\gamma} \Gamma\left(1 + \frac{1}{\gamma}\right)$$

$$VarX = \beta^{2/\gamma} \left[\Gamma\left(1 + \frac{2}{\gamma}\right) - \Gamma^2\left(1 + \frac{1}{\gamma}\right) \right]$$

where Γ is called the gamma function and is defined by:

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$$

Γ has the following properties:

$$\Gamma(1) = 1; \quad \text{for any } \alpha > 1 \quad \Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1)$$

$$\text{for any positive integer } \alpha > 1 \quad \Gamma(\alpha) = (\alpha-1)!$$

- Normal (μ, σ^2) $N(\mu, \sigma^2)$

$$f_X(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

$$EX = \mu \quad VarX = \sigma^2$$

$$M_X(t) = \int_{-\infty}^{\infty} e^{xt} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} dx$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{(2\sigma^2 xt - x^2 + 2x\mu - \mu^2)/2\sigma^2} dx$$

$$\begin{aligned}
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{(-x^2+2x(\mu+\sigma^2t)-(\mu+\sigma^2t)^2)/2\sigma^2} dx \times e^{(2\mu\sigma^2t+\sigma^4t^2)/2\sigma^2} \\
&= 1 \times e^{\mu t + \sigma^2 t^2 / 2}
\end{aligned}$$

- Chi-squared (p) ($\chi^2(p)$)

$$f_X(x|p) = \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2} \quad 0 \leq x < \infty, \quad p = 1, 2, \dots$$

$$\begin{aligned}
M_X(t) &= \int_0^{\infty} e^{xt} \frac{1}{\Gamma(p/2)2^{p/2}} x^{(p/2)-1} e^{-x/2} dx \\
&= \frac{1}{\Gamma(p/2)2^{p/2}} \int_0^{\infty} x^{(p/2)-1} e^{-x(1-2t)/2} dx \times (1-2t)^{(p/2)-1} (1-2t)^{-(p/2)+1}
\end{aligned}$$

for change of variables, define $y = x(1-2t)$, then $dx = \frac{1}{1-2t} dy$

$$= \frac{1}{\Gamma(p/2)2^{p/2}} \int_0^{\infty} y^{(p/2)-1} e^{-y/2} dy \times \frac{1}{1-2t} \times (1-2t)^{-(p/2)+1}$$

$$= 1 \times (1-2t)^{-(p/2)}$$

$$M'_X(t) = p(1-2t)^{-p/2-1}$$

$$M''_X(t) = p(p/2+1)(2)(1-2t)^{-p/2-2}$$

$$EX = M'_X(0) = p$$

$$EX^2 = M''_X(0) = p^2 + 2p$$

$$VarX = p^2 + 2p - p^2 = 2p$$

Fact: $X_i \sim N(0, 1)$, then $\sum_{i=1}^p X_i^2 \sim \chi^2(p)$

- F_{ν_1, ν_2}

$$f_X(x|\nu_1, \nu_2) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\frac{\nu_1}{2})\Gamma(\frac{\nu_2}{2})} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{x^{(\nu_1-2)/2}}{\left(1+(\frac{\nu_1}{\nu_2})x\right)^{(\nu_1+\nu_2)/2}}$$

$$0 < \nu_1, \nu_2 < \infty$$

$$EX = \frac{\nu_2}{\nu_2-2} \quad \nu_2 > 2$$

$$VarX = 2 \left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{\nu_1+\nu_2-2}{\nu_1(\nu_2-4)} \quad \nu_2 > 4$$

Fact: $F_{\nu_1, \nu_2} = \frac{(\chi^2(\nu_1)/\nu_1)}{(\chi^2(\nu_2)/\nu_2)}$ for independent χ^2 s

- t_ν

$$f_X(x|\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+(\frac{x^2}{\nu})\right)^{(\nu+1)/2}} \quad -\infty < x < \infty, \quad \nu = 1, \dots$$

$$EX = 0, \quad \nu > 1$$

$$VarX = \frac{\nu}{\nu-2}, \quad \nu > 2$$

Fact: $F_{1, \nu} = (t_\nu)^2$

Fact: if $W \sim N(0, 1)$ $V \sim \chi^2(r)$ then $T = \frac{W}{\sqrt{V/r}} \sim t_r$

Exercise #13 (Normal as the limit of t)

Show that if $X \sim t_\nu$, $Y \sim N(0, 1)$, then $\lim_{\nu \rightarrow \infty} X = Y$.

Do this roughly by using the Table of Distributions:

$$\begin{aligned} P(X \leq x) & \quad X \sim t_\nu \\ P(Y \leq y) & \quad Y \sim N(0, 1) \end{aligned}$$

4 Maximum Likelihood and Sufficient Statistics

This final section will take a short journey into estimating the parameters of a distribution given data. Given a vector of random variables $X = (X_1, \dots, X_n)'$, we want to find the parameters θ that maximizes the cross-product of random variable densities. The objective function to be maximized is called the likelihood function and is denoted $L(\theta; X)$.

Definition 52 $L(\theta; X) = \prod_{i=1}^n f_X(x_i; \theta) \quad \theta \in \Omega$

In population, there exists true values for the parameters θ , denoted θ_0 . Unfortunately, we only have a finite sample of the population, so we estimate the parameters θ , denoted θ^* , given this sample.

Definition 53 $\theta^* = \arg \max_{\theta \in \Omega} L(\theta; X)$

The definition above is justified by the Theorem below.

Theorem 54 Under the following regularity conditions

(i) $\theta \neq \theta' \Rightarrow f_X(x_i; \theta) \neq f_X(x_i; \theta')$ (distinct pdfs)

(ii) The pdfs have common support for all θ

(iii) $\theta_0 \in \text{int}(\Omega)$

Then $\lim_{n \rightarrow \infty} P[L(\theta_0; X) > L(\theta; X)] = 1$ for all $\theta \neq \theta_0$

Proof: By taking the natural log, $L(\theta_0; X) > L(\theta; X)$ is equivalent to

$$\frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_X(x_i; \theta_0)}{f_X(x_i; \theta)} \right] < 0$$

We will need the following Lemma to complete the proof:

Lemma: Law of Large Numbers

Given the true parameters θ_0 , then $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} E_{\theta_0}[X]$

Using this lemma, $\frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_X(x_i; \theta)}{f_X(x_i; \theta_0)} \right] \xrightarrow{p} E_{\theta_0} \left[\log \frac{f_X(x; \theta)}{f_X(x; \theta_0)} \right]$

by Jensen's inequality,

$$\begin{aligned}
E_{\theta_0} \left[\log \frac{f_X(x; \theta)}{f_X(x; \theta_0)} \right] &< \log E_{\theta_0} \left[\frac{f_X(x; \theta)}{f_X(x; \theta_0)} \right] \\
&= \log \int \frac{f_X(x; \theta)}{f_X(x; \theta_0)} f_X(x; \theta_0) dx = \log(1) = 0 \\
\text{by the definition of convergence in probability,} \\
\lim_{n \rightarrow \infty} P \left[\left| \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_X(x_i; \theta)}{f_X(x_i; \theta_0)} \right] - c \right| > \epsilon \right] &= 0 \quad \text{where } c < 0 \\
\lim_{n \rightarrow \infty} P \left[\left| \frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_X(x_i; \theta)}{f_X(x_i; \theta_0)} \right] - c \right| < \epsilon \right] &= 1 \quad \text{any } \epsilon > 0 \\
\text{thus, } \lim_{n \rightarrow \infty} P \left[\frac{1}{n} \sum_{i=1}^n \log \left[\frac{f_X(x_i; \theta)}{f_X(x_i; \theta_0)} \right] < 0 \right] &= 1 \quad \text{and we're done.}
\end{aligned}$$

Example 55 Suppose that the vector of random variables $X = (X_1, \dots, X_n)'$ are i.i.d. with distribution $X_i \sim N(\mu, \sigma^2)$

Here, $\theta = (\mu, \sigma^2)$. The likelihood function is given by (since $f_X(x_i; \theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - \mu)^2 / (2\sigma^2)}$)

$$L(\theta; X) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} e^{-(x_i - \mu)^2 / (2\sigma^2)}$$

it is easier to maximize the log transformation of the likelihood (this is of course allowed since the transformation is monotonic)

$$\begin{aligned}
\log L(\theta; X) &= \sum_{i=1}^n \left[-\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\sigma^2) - (x_i - \mu)^2 / (2\sigma^2) \right] \\
&= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log L(\theta; X)}{\partial \mu} = 0 &\Rightarrow -\frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) = 0 \\
&\Rightarrow \sum_{i=1}^n x_i = n\mu \Rightarrow \mu^* = \frac{1}{n} \sum_{i=1}^n x_i
\end{aligned}$$

$$\begin{aligned}
\frac{\partial \log L(\theta; X)}{\partial \sigma^2} = 0 &\Rightarrow -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 = 0 \\
&\Rightarrow \sigma^{2*} = \frac{1}{n} \sum_{i=1}^n (x_i - \mu^*)^2
\end{aligned}$$

Exercise #14 (Find the MLE)

Suppose that the vector of random variables $X = (X_1, \dots, X_n)'$ are i.i.d. with distribution $X_i \sim \text{Pareto}(\alpha)$ (here $\theta = \alpha$) and the Pareto distribution is defined as $F_X(x_i; \alpha) = 1 - \left(\frac{1}{x_i}\right)^\alpha$ $1 \leq x_i < \infty$, $\alpha > 0$

Derive the MLE α^* .

4.1 Sufficient Statistics

We have a vector of random variables $X = (X_1, \dots, X_n)'$ with distribution parameters θ . From a sample, we can have any large number of statistics to describe the data. However, the statistic defined so that $\theta_i = X_i \quad \forall i = 1, \dots, n$ is not useful for example. Thus, we want the smallest number of statistics that can still allow us to estimate the parameters θ , denoted θ^* .

Definition 56 Let X_1, X_2, \dots, X_n denote random variables from a distribution that has pdf $f_X(x; \theta)$, $\theta \in \Omega$. The statistic $Y_1 = u_1(X_1, \dots, X_n)$ is a sufficient

statistic for θ if and only if we can find two nonnegative functions, k_1 and k_2 , such that

$$f_X(x_1; \theta) \cdots f_X(x_n; \theta) = k_1[u_1(x_1, \dots, x_n); \theta] k_2(x_1, \dots, x_n)$$

where $k_2(x_1, \dots, x_n)$ does not depend upon θ

Example 57 Let X_1, X_2, \dots, X_n denote random variables from a normal distribution, i.e. $X_i \sim N(\mu, \sigma^2)$ $-\infty < \theta < \infty$ where σ^2 is known. We wish to show that $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is a sufficient statistic for θ .

$$\begin{aligned} \text{Preliminary: } \sum_{i=1}^n (x_i - \theta)^2 &= \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \theta)]^2 \\ &= \sum_{i=1}^n [(x_i - \bar{x})^2 + (\bar{x} - \theta)^2 + 2(\bar{x} - \theta)(x_i - \bar{x})] \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \theta)^2 + 2(\bar{x} - \theta) \sum_{i=1}^n (x_i - \bar{x}) \end{aligned}$$

notice that the third term is equal to 0

$$\begin{aligned} \text{thus, } f_X(x_1; \theta) \cdots f_X(x_n; \theta) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\sum_{i=1}^n (x_i - \theta)^2 / 2\sigma^2\right] \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\sum_{i=1}^n (x_i - \bar{x})^2 / 2\sigma^2 - n(\bar{x} - \theta)^2 / 2\sigma^2\right] \\ &= \underbrace{\left\{\exp\left[-n(\bar{x} - \theta)^2 / 2\sigma^2\right]\right\}}_{k_1} \underbrace{\left\{\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left[-\sum_{i=1}^n (x_i - \bar{x})^2 / 2\sigma^2\right]\right\}}_{k_2} \end{aligned}$$

thus, \bar{x} is a sufficient statistic for θ .

Exercise #15 (Find the Sufficient Statistic)

Let X_1, X_2, \dots, X_n denote random variables from a normal distribution, i.e. $X_i \sim N(0, \theta)$ $0 < \theta < \infty$
 Show that $\sum_{i=1}^n X_i^2$ is a sufficient statistic for θ .