Fall 2018 Statistics 201A (Introduction to Probability at an advanced level) - All Lecture Notes

Aditya Guntuboyina

August 15, 2020

Contents

We will start the course by a review of undergraduate probability material. The review will be fairly quick and should be complete in about six lectures.

0.1 Sample spaces, Events, Probability

Probability theory is invoked in situations that are (or can be treated as) chance or random experiments. In such a random experiment, the *sample space* is the set of all possible outcomes and we shall denote it by Ω.

For example, suppose we are tossing a coin 2 times. Then a reasonable sample space is $\{hh, ht, th, tt\}$.

Subsets of the sample space are called *Events*. For example, in the above example, $\{hh, ht\}$ is an Event and it represents the event that the first of the two tosses results in a heads. Similary, the event that at least one of the two tosses results in a heads is represented by $\{hh, ht, th\}$.

Given a collection of events $A_1, A_2, \dots,$

1. A_1^c denotes the event that A_1 does not happen. We say that A_1^c is the complement of the event A_1.
2. $\cup_{i \geq 1} A_i$ denotes the event that at least one of $A_1, A_2, \dots$ happens.
3. $\cap_{i \geq 1} A_i$ denotes the event that all of $A_1, A_2, \dots$ happen.

Probability is defined as a function that maps (or associates) events to real numbers between 0 and 1 and which satisfies certain natural consistency properties. Specifically $\mathbb{P}$ is a probability provided:

1. $0 \leq \mathbb{P}(A) \leq 1$ for every event A.
2. For the empty subset Φ (= the "impossible event"), $\mathbb{P}(\Phi) = 0$

3. For the whole sample space (= the “certain event”), $\mathbb{P}(\Omega) = 1$.
4. If an event A is a **disjoint union** of a sequence of events $A_1, A_2, \dots$ (this means that every point in A belongs to exactly one of the sets $A_1, A_2, \dots$), then $\mathbb{P}(A) = \sum_{i \geq 1} \mathbb{P}(A_i)$.

Example 0.1 (Nontransitive Dice)**.** *Consider the following set of dice:*

1. *Die A has sides 2, 2, 4, 4, 9, 9.*
2. *Die B has sides 1, 1, 6, 6, 8, 8.*
3. *Die C has sides 3, 3, 5, 5, 7, 7.*

What is the probability that A rolls a higher number than B? What is the probability that B rolls higher than C? What is the probability that C rolls higher than A? Assume that, in any roll of dice, all outcomes are equally likely.

0.2 Conditional Probability and Independence

Consider a probability $\mathbb{P}$ and an event B for which $\mathbb{P}(B) > 0$. We can then define $\mathbb{P}(A|B)$ for every event A as

$$\mathbb{P}(A|B) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{1}$$

$\mathbb{P}(A|B)$ is called the conditional probability of A given B.

Here is two very interesting problems from Mosteller’s delightful book (titled *Fifty Challenging Problems in Probability*) illustrating the use of conditional probabilities.

Example 0.2 (From Mosteller’s book (Problem 13; The Prisoner’s Dilemma))**.** *Three prisoners, A, B, and C, with apparently equally good records have applied for parole. The parole board has decided to release two of the three, and the prisoners know this but not which two. A warder friend of prisoner A knows who are to be released. Prisoner A realizes that it would be unethical to ask the warder if he, A, is to be released, but thinks of asking for the name of one prisoner other than himself who is to be released. He thinks that before he asks, his chances of release are* $2/3$. *He thinks that if the warder says ”B will be released,” his own chances have now gone down to* $1/2$, *because either A and B or B and C are to be released. And so A decides not to reduce his chances by asking. However, A is mistaken in his calculations. Explain.*

Example 0.3 (From Mosteller’s book (Problem 20: The Three-Cornered Duel))**.** *A, B, and C are to fight a three-cornered pistol duel. All know that A’s chance of hitting his target is 0.3, C’s is 0.5, and B never misses. They are to fire at their choice of target in succession in the order A, B, C, cyclically (but a hit man loses further turns and is no longer shot at) until only one man is left unhit. What should A’s strategy be?*

We say that two events A and B are independent (under the probability $\mathbb{P}$) if

$$\mathbb{P}(A|B) = \mathbb{P}(A).$$

Equivalently, independence is given by $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$ or $\mathbb{P}(B|A) = \mathbb{P}(B)$.

A very important formula involving conditional probabilities is the Bayes’ rule. This says that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}. \tag{2}$$

Can you derive this from the definition (1) of conditional probability?

Here is a very interesting application of Bayes’ rule.

Example 0.4. *Consider a clinical test for cancer that can yield either a positive (+) or negative (-) result. Suppose that a patient who truly has cancer has a 1% chance of slipping past the test undetected. On the other hand, suppose that a cancer-free patient has a 5% probabiliity of getting a positive test result. Suppose also that 2% of the population has cancer. Assuming that a patient who has been given the test got a positie test result, what is the probability that they have cancer?*

Suppose C and C^c are the events that the patient has cancer and does not have cancer respectively. Also suppose that $+$ and $-$ are the events that the test yields a positive and negative result respectively. By the information given, we have

$$\mathbb{P}(-|C) = 0.01 \qquad \mathbb{P}(+|C^c) = 0.05 \qquad \mathbb{P}(C) = 0.02.$$

We need to compute $\mathbb{P}(C|+)$. By Bayes rule, we have

$$\mathbb{P}(C|+) = \frac{\mathbb{P}(+\cap C)}{\mathbb{P}(+)} = \frac{\mathbb{P}(+|C)\mathbb{P}(C)}{\mathbb{P}(+|C)\mathbb{P}(C) + \mathbb{P}(+|C^c)P(C^c)} = \frac{0.99 * 0.02}{0.99 * 0.02 + 0.05 * 0.98} = 0.2878.$$

Therefore the probability that this patient has cancer (given that the test gave a positive result) is about 29%. This means, in particular, that it is still unlikely that they have cancer even though the test gave a positive result (note though that the probability of cancer increased from 2% to 29%).

Another interesting aspect of the above calculation is that

$$\mathbb{P}(+) = \mathbb{P}(+|C)\mathbb{P}(C) + \mathbb{P}(+|C^c)P(C^c) = 0.99 * 0.02 + 0.05 * 0.98 = 0.0688.$$

This means that test will yield a positive result about 7% of the time (note that only 2% of the population has cancer).

Suppose now that $\mathbb{P}(C) = 0.001$ (as opposed to $\mathbb{P}(C) = 0.02$) and assume that $\mathbb{P}(-|C)$ and $\mathbb{P}(+|C^c)$ stay at 0.01 and 0.05 as before. Then

$$\mathbb{P}(+) = \mathbb{P}(+|C)\mathbb{P}(C) + \mathbb{P}(+|C^c)P(C^c) = 0.99 * 0.001 + 0.05 * 0.999 = 0.0509.$$

Here the true cancer rate of 0.001 has yielded in an apparent rate of 0.05 (which is an increase by a factor of 50). Think about this in the setting where the National Rifle Association is taking a survey by asking a sample of citizens whether they used a gun in self-defense during the past year. Take C to be true usage and $+$ to be reported usage. If only one person in a thousand had truly used a gun in self-defense, it will appear that one in twenty did. These examples are taken from the amazing book titled "Understanding Uncertainty" by Dennis Lindley (I feel that every student of probability and statistics should read this book).

0.3 Random Variables

A random variable is a function that attaches a number to each element of the sample space. In other words, it is a function mapping the sample space to real numbers.

For example, in the chance experiment of tossing a coin 50 times, the number of heads is a random variable. Another random variable is the number of heads before the first tail. Another random variable is the number of times the pattern *hththt* is seen.

Many real-life quantities such as (a) The average temperature in Berkeley tomorrow, (b) The height of a randomly chosen student in this room, (c) the number of phone calls that I will receive tomorrow, (d) the number of accidents that will occur on Hearst avenue in September, etc. can be treated as random variables.

For every event A (recall that events are subsets of the sample space), one can associate a random variable which take the value 1 if A occurs and 0 if A does not occur. This is called the *indicator* random variable corresponding to the event A and is denoted by $I(A)$.

The *distribution* of a random variable is, informally, a description of the set of values that the random variable takes and the probabilities with which it takes those values.

If a random variable X takes a finite or countably infinte set of possible values (in this case, we say that X is a *discrete* random variable), its distribution is described by a listing of the values $a_1, a_2, \dots$ that it takes together with a specification of the probabilities:

$$\mathbb{P}\{X = a_i\} \qquad \text{for } i = 1, 2, \dots.$$

The function which maps a_i to $\mathbb{P}\{X = a_i\}$ is called the *probability mass function* (pmf) of the discrete random variable X.

If a random variable X takes a continuous set of values, its *distribution* is often described by a function called the *probability density function* (pdf). The pdf is a function f on $\mathbb{R}$ that satisfies $f(x) \geq 0$ for every $x \in \mathbb{R}$ and

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

The pdf f of a random variable can be used to calculate $\mathbb{P}\{X \in A\}$ for every set A via

$$\mathbb{P}\{X \in A\} = \int_A f(x)dx.$$

Note that if X has density f, then for every $y \in \mathbb{R}$,

$$\mathbb{P}\{X = y\} = \int_y^y f(x)dx = 0.$$

It is important to remember that a density function $f(x)$ of a random variable does not represent probability (in particular, it is quite common for $f(x)$ to take values much larger than one). Instead, the value $f(x)$ can be thought of as a constant of proportionality. This is because usually (as long as f is continuous at x):

$$\lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{P}\{x \leq X \leq x + \delta\} = f(x).$$

1 Random Variables, Expectation and Variance

A random variable is a function that attaches a number to each element of the sample space. In other words, it is a function mapping the sample space to real numbers.

For example, in the chance experiment of tossing a coin 50 times, the number of heads is a random variable. Another random variable is the number of heads before the first tail. Another random variable is the number of times the pattern *hththt* is seen.

Many real-life quantities such as (a) The average temperature in Berkeley tomorrow, (b) The height of a randomly chosen student in this room, (c) the number of phone calls that I will receive tomorrow, (d) the number of accidents that will occur on Hearst avenue in September, etc. can be treated as random variables.

For every event A (recall that events are subsets of the sample space), one can associate a random variable which take the value 1 if A occurs and 0 if A does not occur. This is called the *indicator* random variable corresponding to the event A and is denoted by $I(A)$.

The *distribution* of a random variable is, informally, a description of the set of values that the random variable takes and the probabilities with which it takes those values.

If a random variable X takes a finite or countably infinte set of possible values (in this case, we say that X is a *discrete* random variable), its distribution is described by a listing of the values $a_1, a_2, \dots$ that it takes together with a specification of the probabilities:

$$\mathbb{P}\{X = a_i\} \qquad \text{for } i = 1, 2, \dots.$$

The function which maps a_i to $\mathbb{P}\{X = a_i\}$ is called the *probability mass function* (pmf) of the discrete random variable X.

If a random variable X takes a continuous set of values, its *distribution* is often described by a function called the *probability density function* (pdf). The pdf is a function f on $\mathbb{R}$ that satisfies $f(x) \geq 0$ for every $x \in \mathbb{R}$ and

$$\int_{-\infty}^{\infty} f(x)dx = 1.$$

The pdf f of a random variable can be used to calculate $\mathbb{P}\{X \in A\}$ for every set A via

$$\mathbb{P}\{X \in A\} = \int_A f(x)dx.$$

Note that if X has density f, then for every $y \in \mathbb{R}$,

$$\mathbb{P}\{X = y\} = \int_y^y f(x)dx = 0.$$

It is important to remember that a density function $f(x)$ of a random variable does not represent probability (in particular, it is quite common for $f(x)$ to take values much larger than one). Instead, the value $f(x)$ can be thought of as a constant of proportionality. This is because usually (as long as f is continuous at x):

$$\lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{P}\{x \leq X \leq x + \delta\} = f(x).$$

The *cumulative distribution function* (cdf) of a random variable X is the function defined as

$$F(x) := \mathbb{P}\{X \leq x\} \qquad \text{for } -\infty < x < \infty.$$

This is defined for all random variables discrete or continuous. If the random variable X has a density, then its cdf is given by

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

The cdf has the following properties: (a) It is non-decreasing, (b) right-continuous, (c) $\lim_{x \downarrow -\infty} F(x) = 0$ and $\lim_{x \uparrow +\infty} F(x) = 1$.

1.1 Expectations of Random Variables

Let X be a discrete random variable and let g be a real-valued function on the range of X. We then say that $g(X)$ has finite expectation provided

$$\sum_x |g(x)| \mathbb{P}\{X = x\} < \infty$$

where the summation is over all possible values x of X. Note the presence of the absolute value on $g(x)$ above.

When $g(X)$ has finite expectation, we define $\mathbb{E}g(X)$ as

$$\mathbb{E}g(X) := \sum_x g(x) \mathbb{P}\{X = x\} < \infty \tag{3}$$

where again the summation is over all possible values x of X.

Analogously, if X is a continuous random variable with density (pdf) f, then we say that $g(X)$ has finite expectation provided

$$\int_{-\infty}^{\infty} |g(x)| f(x)dx < \infty.$$

When $g(X)$ has finite expectation, we define $\mathbb{E}g(X)$ as

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx. \tag{4}$$

Why do we need to ensure finiteness of the absolute sums (or integrals) before defining expectation? Because otherwise the sum in (3) (or the integral in (4)) might be ill-defined. For example, when X is the discrete random variable which takes the values $\dots, -3, -2, -1, 1, 2, 3, \dots$ with probabilities

$$\mathbb{P}\{X = i\} = \frac{3}{\pi^2}\frac{1}{i^2} \qquad \text{for } i \in \mathbb{Z}, i \neq 0.$$

Then the sum in (3) for $g(x) = x$ becomes

$$\mathbb{E}X = \frac{3}{\pi^2} \sum_{i \in \mathbb{Z}: i \neq 0} \frac{1}{i}$$

which can not be made any sense of. it is easy to see here that $\sum_x |x|\mathbb{P}\{X = x\} = \infty$.

If A is an event, then recall that $I(A)$ denotes the corresponding indicator random variable that equals 1 when A holds and 0 when A does not hold. It is convenient to note that the expectation of $I(A)$ precisely equals $\mathbb{P}(A)$.

An important thing to remember is that Expectation is a *linear operator* i.e.,

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

for any two random variables X and Y with finite expectations and real numbers a and b.

1.2 Variance

A random variable X is said to have finite variance if X^2 has finite expectation (do you know that when X^2 has finite expectation, X also will have finite expectation? how will you prove this?). In that case, the variance of X^2 is defined as

$$Var(X) := \mathbb{E}(X - \mu_X)^2 = \mathbb{E}(X^2) - \mu_X^2 \qquad \text{where } \mu_X := \mathbb{E}(X).$$

It is clear from the definition that Variance of a random variable X measures the average variability in the values taken by X around its mean $\mathbb{E}(X)$.

Suppose X is a discrete random variables taking finitely many values $x_1, \dots, x_n$ with equal probabilities. Then what is the variance of X?

The square root of the variance of X is called the standard deviation of X and is often denoted by $SD(X)$.

The Expectation of a random variable X has the following variational property: it is the value of a that minimizes the quantity $\mathbb{E}(X - a)^2$ over all real numbers a. Do you know how to prove this?

If the variance of a random variable X is small, then X cannot deviate much from its mean ($= \mathbb{E}(X) = \mu$). This can be made precise by Chebyshev's inequality which states the following.

Chebyshev's Inequality: Let X be a random variable with finite variance and mean μ. Then for every $\epsilon > 0$, the following inequality holds:

$$\mathbb{P}\left\{|X - \mu| \geq \epsilon\right\} \leq \frac{Var(X)}{\epsilon^2}.$$

In other words, the probability that X deviates by more than ϵ from its mean is bounded from above by $Var(X)/\epsilon^2$.

Proof of Chebyshev's inequality: Just argue that

$$I\{|X-\mu| \geq \epsilon\} \leq \frac{(X-\mu)^2}{\epsilon^2}$$

and take expectations on both sides (on the left hand side, we have the Indicator random variable that takes the value 1 when $|X-\mu| \geq \epsilon$ and 0 otherwise).

2 Independence of Random Variables

We say that two random variables X and Y are independent if conditioning on any event involving Y does not change the probability of any event involving X i.e.,

$$\mathbb{P}\{X \in A | Y \in B\} = \mathbb{P}\{X \in A\}$$

for every A and B.

Equivalently, independence of X and Y is same as

$$\mathbb{P}\{X \in A, Y \in B\} = \mathbb{P}\{X \in A\}\mathbb{P}\{Y \in B\}$$

for every A and B.

The following are consequences of independence. If X and Y are independent, then

1. $g(X)$ and $h(Y)$ are independent for every pair of functions g and h.
2. $\mathbb{E}\left(g(X)h(Y)\right) = \mathbb{E}g(X)\mathbb{E}h(Y)$ for every pair of functions g and h.

More generally, we say that random variables $X_1, \dots, X_k$ are (mutually) independent if, for every $1 \leq i \leq k$, conditioning on any event involving $X_j, j \neq i$ does not change the probability of any event involving X_i. From here one can easily derive properties of independence such as

$$\mathbb{P}\{X_1 \in A_1, \dots, X_k \in A_k\} = \mathbb{P}\{X_1 \in A_1\}\mathbb{P}\{X_2 \in A_2\} \dots \mathbb{P}\{X_k \in A_k\}$$

for all possible choices of events $A_1, \dots, A_k$.

3 Common Distributions

3.1 $Ber(p)$ Distribution

A random variable X is said to have the $Ber(p)$ (Bernoulli with parameter p) distribution if it takes the two values 0 and 1 with $\mathbb{P}\{X = 1\} = p$.

Note then that $\mathbb{E}X = p$ and $Var(X) = p(1-p)$. For what value of p is X most variable? least variable?

3.2 $Bin(n, p)$ Distribution

A random variable X is said to have the Binomial distribution with parameters n and p (n is a positive integer and $p \in [0, 1]$) if it takes the values $0, 1, \dots, n$ with pmf given by

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} \qquad \text{for every } k = 0, 1, \dots, n.$$

Here $\binom{n}{k}$ is the binomial coefficient:

$$\binom{n}{k} := \frac{n!}{k!(n-k)!}.$$

The main example of a $Bin(n,p)$ random variable is the number of heads obtained in n *independent* tosses of a coin with probability of heads equalling p.

Here is an interesting problem about the Binomial distribution from Mosteller's book (you can easily calculate these probabilities in R).

Example 3.1 (From Mosteller's book (Problem 19: Issac Newton helps Samuel Pepys))**.** *Pepys wrote Newton to ask which of three events is more likely: that a person get (a) at least I six when 6 dice are rolled, (b) at least 2 sixes when 12 dice are rolled, or (c) at least 3 sixes when 18 dice are rolled What is the answer?*

Let X denote the number of heads in n independent tosses of a coin with probability of heads being p. Then we know that $X \sim Bin(n,p)$. If, now, X_i denotes the binary random variable that takes 1 if the i^{th} toss is a heads and 0 if the i^{th} toss is a tail, then it should be clear that

$$X = X_1 + \cdots + X_n.$$

Note that each X_i is a $Ber(p)$ random variable and that $X_1, \ldots, X_n$ are independent. Therefore $Bin(n,p)$ random variables can be viewed as sums of n independent $Ber(p)$ random variables. The Central Limit Theorem (which we will study in detail later in the class) implies that the sum of a large number of i.i.d (what is i.i.d?) random variables is approximately normal. This means that when n is large and p is held fixed, the $Bin(n,p)$ distribution looks like a normal distribution. We shall make this precise later. In particular, this means that Binomial probabilities can be approximately calculated via normal probabilities for n large and p fixed. From this point of view, what is the probability of getting k or more sixes from $6k$ rolls of a die when k is large?

What is the mean of the $Bin(n,p)$ distribution? What is an *unbiased* estimate of the probability of heads based on n independent tosses of a coin? What is the variance of $Bin(n,p)$?

3.3 Poisson Distribution

A random variable X is said to have the Poisson distribution with parameter $\lambda > 0$ (denoted by $Poi(\lambda)$) if X takes the values $0, 1, 2, \ldots$ with pmf given by

$$\mathbb{P}\{X = k\} = e^{-\lambda}\frac{\lambda^k}{k!} \qquad \text{for } k = 0, 1, 2, \ldots.$$

The main utility of the Poisson distribution comes from the following fact:

Fact: The binomial distribution $Bin(n,p)$ is well-approximated by the Poisson distribution $Poi(np)$ provided that the quantity np^2 small.

To intuitively see why this is true, just see that

$$\mathbb{P}\left\{Bin(n,p) = 0\right\} = (1-p)^n = \exp\left(n\log(1-p)\right).$$

Note now that np^2 being small implies that p is small (note that p can be written as $\sqrt{np^2/n} \leq \sqrt{np^2}$ so small np^2 will necessarily mean that p is small). When p is small, we can approximate $\log(1-p)$ as $-p-p^2/2$ so we get

$$\mathbb{P}\left\{Bin(n,p) = 0\right\} = \exp\left(n\log(1-p)\right) \approx \exp\left(-np\right)\exp\left(-np^2/2\right).$$

Now because np^2 is small, we can ignore the second term above to obtain that $\mathbb{P}\{Bin(n,p) = 0\}$ is approximated by $\exp(-np)$ which is precisely equal to $\mathbb{P}\{Poi(np) = 0\}$. One can similarly approximate $\mathbb{P}\{Bin(n,p) = k\}$ by $\mathbb{P}\{Poi(np) = k\}$ for every fixed $k = 0, 1, 2, \ldots$.

There is a formal theorem (known as Le Cam's theorem) which rigorously proves that $Bin(n,p) \approx Poi(np)$ when np^2 is small. This is stated without proof below (its proof is beyond the scope of this class).

Theorem 3.2 (Le Cam's Theorem)**.** *Suppose $X_1, \dots, X_n$ are independent random variables such that $X_i \sim Ber(p_i)$ for some $p_i \in [0,1]$ for $i = 1, \dots, n$. Let $X = X_1 + \cdots + X_n$ and $\lambda = p_1 + \dots p_n$. Then*

$$\sum_{k=0}^{\infty} |\mathbb{P}\{X = k\} - \mathbb{P}\{Poi(\lambda) = k\}| < 2\sum_{i=1}^{n} p_i^2.$$

In the special case when $p_1 = \cdots = p_n = p$, the above theorem says that

$$\sum_{k=0}^{\infty} |\mathbb{P}\{Bin(n,p) = k\} - \mathbb{P}\{Poi(np) = k\}| < 2np^2$$

and thus when np^2 is small, the probability $\mathbb{P}\{Bin(n,p) = k\}$ is close to $\mathbb{P}\{Poi(np) = k\}$ for each $k = 0, 1, \dots$.

An implication of this fact is that for every fixed $\lambda > 0$, we have

$$Poi(\lambda) \approx Bin\left(n, \frac{\lambda}{n}\right) \qquad \text{when } n \text{ is large.}$$

This is because when $p = \lambda/n$, we have $np^2 = \lambda^2/n$ which will be small when n is large.

This approximation property of the Poisson distribution is the reason why the Poisson distribution is used to model counts of rare events. For example, the number of phone calls a telephone operator receives in a day, the number of accidents in a particular street in a day, the number of typos found in a book, the number of goals scored in a football game can all be modelled as $Poi(\lambda)$ for some $\lambda > 0$. Can you justify why these real-life random quantities can be modeled by the Poisson distribution?

The following example presents another situation where the Poisson distribution provides a good approximation.

Example 3.3. *Consider n letters numbered $1, \dots, n$ and n envelopes numbered $1, \dots, n$. The right envelope for letter i is the envelope i. Suppose that I take a random permutation $\sigma_1, \dots, \sigma_n$ of $1, \dots, n$ and then place the letter σ_i in the envelope i. Let X denote the number of letters which are in their right envelopes. What is the distribution of X?*

Let X_i be the random variable which takes the value 1 when the i^{th} letter is in the i^{th} envelope and 0 otherwise. Then clearly $X = X_1 + \cdots + X_n$. Note that

$$\mathbb{P}\{X_i = 1\} = \frac{1}{n} \qquad \textit{for each } i = 1, \dots, n.$$

This is because the i^{th} letter is equally likely to be in any of the n envelopes. This means therefore that

$$X_i \sim Ber(1/n) \qquad \textit{for } i = 1, \dots, n.$$

If the X_i's were also independent, then $X = X_1 + \cdots + X_n$ will be $Bin(n, 1/n)$ which is very close to $Poi(1)$ for large n. But the X_i's are not independent here because for $i \neq j$,

$$\mathbb{P}\{X_i = 1 | X_j = 1\} = \frac{1}{n-1} \neq \frac{1}{n} = \mathbb{P}\{X_i = 1\}.$$

However, the dependence is relatively weak and it turns out that the distribution of X is quite close to $Poi(1)$. We shall demonstrate this by showing that $\mathbb{P}\{X = 0\}$ is approximately equal to $\mathbb{P}\{Poi(1) = 0\} = e^{-1}$. I will leave as an exercise to show that $\mathbb{P}\{X = k\} \approx \mathbb{P}\{Poi(1) = k\}$ for every fixed k. To compute $\mathbb{P}\{X = 0\}$, we *can write*

$$\begin{aligned}
\mathbb{P}\{X=0\} &= \mathbb{P}\{\prod_{i=1}^{n}(1-X_i)=1\} \\
&= \mathbb{E}\prod_{i=1}^{n}(1-X_i) \\
&= \mathbb{E}\left\{1-\sum_{i=1}^{n}X_i+\sum_{i<j}X_iX_j-\sum_{i<j<k}X_iX_jX_k+\dots+(-1)^nX_1\dots X_n\right\} \\
&= 1-\sum_{i}\mathbb{E}(X_i)+\sum_{i<j}\mathbb{E}(X_iX_j)-\sum_{i<j<k}\mathbb{E}(X_iX_jX_k)+\dots+(-1)^n\mathbb{E}(X_1,\dots,X_n).
\end{aligned}$$

Note now that for every $i_1 < \dots < i_k$, we have

$$\mathbb{E}X_{i_1}X_{i_2}\dots X_{i_k} = \mathbb{P}\{X_{i_1}=1, X_{i_2}=1,\dots,X_{i_k}=1\} = \frac{(n-k)!}{n!}.$$

This gives

$$\mathbb{P}\{X=0\} = \sum_{k=0}^{n}(-1)^k\binom{n}{k}\frac{(n-k)!}{n!} = \sum_{k=0}^{n}(-1)^k\frac{1}{k!} \approx e^{-1} = \mathbb{P}\{Poi(1)=0\}.$$

It is an easy exercise to show that the expectation and variance of a $Poi(\lambda)$ random variable are both equal to λ. This also makes sense because of the connection:

$$Poi(\lambda) \approx Bin(n, \lambda/n)$$

as

$$\mathbb{E}(Bin(n,\lambda/n)) = \lambda \quad \text{and} \quad var(Bin(n,\lambda/n)) = n\frac{\lambda}{n}\left(1-\frac{\lambda}{n}\right) \to \lambda \text{ as } n\to\infty.$$

When modeling count data via the Poisson distribution, it is possible to empirically check the assumption that the variance is equal to the mean. If the empirical variance seems much higher than the mean, then it is said that there is *overdispersion* in which case Poisson may not be a good model for the data.

4 Covariance, Correlation and Regression

Given two random variables X and Y (such that X^2 and Y^2 have finite expectation), the covariance between X and Y is denoted by $Cov(X,Y)$ and is defined as

$$Cov(X,Y) := \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] \tag{5}$$

where $\mu_X := \mathbb{E}(X)$ and $\mu_Y := \mathbb{E}(Y)$. In other words, $Cov(X,Y)$ is defined as the Expectation of the random variable $(X-\mu_X)(Y-\mu_Y)$ (but does this random variable have finite expectation ? Can you verify that this is a consequence of the assumption that X^2 and Y^2 have finite expectation?).

It is important to note that Covariance is a bilinear operator i.e.,

$$Cov(\sum_i a_iX_i, \sum_j b_jY_j) = \sum_i\sum_j a_ib_jCov(X_i,Y_j). \tag{6}$$

Can you prove this as a consequence of the definition (5) of Covariance and the linearity of the Expectation operator?

When $X=Y$, it is easy to see that $Cov(X,X)$ is simply the Variance of X. Using this connection between Covariance and Variance and (6), can you deduce the following standard properties of Variance:

1. $Var(aX + b) = a^2 Var(X)$.

2. $Var(\sum_i a_i X_i) = \sum_i a_i^2 Var(X_i) + \sum_{i \neq j} a_i a_j Cov(X_i, X_j)$.

The correlation between two random variables X and Y (which are such that X^2 and Y^2 have finite variance) is defined as:

$$\rho_{X,Y} := \frac{Cov(X,Y)}{SD(X)SD(Y)} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

If $\rho_{X,Y} = 0$, we say that X and Y are *uncorrelated.*

Proposition 4.1. *Two facts about correlation:*

1. *The correlation $\rho_{X,Y}$ always lies between -1 and 1.*

2. *$\rho_{aX+b,cX+d} = \frac{a}{|a|}\frac{c}{|c|}\rho_{X,Y}$ for every $a, b, c, d \in (-\infty, \infty)$. In words, correlation is invariant (up to sign flips) under linear transformations.*

Proof. Write

$$\rho_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \mathbb{E}\left(\frac{X - \mu_X}{\sqrt{Var(X)}}\frac{Y - \mu_Y}{\sqrt{Var(Y)}}\right).$$

Use the standard inequality:

$$ab \leq \frac{a^2 + b^2}{2} \tag{7}$$

with $a = (X - \mu_X)/\sqrt{Var(X)}$ and $b = (Y - \mu_Y)/\sqrt{Var(Y)}$ to obtain

$$\rho_{X,Y} \leq \mathbb{E}\left(\frac{(X - \mu_X)^2}{2Var(X)} + \frac{(Y - \mu_Y)^2}{2Var(Y)}\right) = \frac{Var(X)}{2Var(X)} + \frac{Var(Y)}{2Var(Y)} = 1.$$

This proves that $\rho_{X,Y} \leq 1$. To prove that $\rho_{X,Y} \geq -1$, argue similarly by using

$$ab \geq \frac{-a^2 - b^2}{2}. \tag{8}$$

The fact about correlations and linear functions is left as an exercise. □

Cauchy-Schwartz Inequality: The fact that correlation $\rho_{X,Y}$ lies between -1 and 1 is sometimes proved via the Cauchy-Schwartz inequality which states the following: For every pair of random variables Z_1 and Z_2, we have

$$|\mathbb{E}(Z_1 Z_2)| \leq \sqrt{\mathbb{E}(Z_1^2)}\sqrt{\mathbb{E}(Z_2^2)} \tag{9}$$

The fact that $|\rho_{X,Y}| \leq 1$ is deduced from the above inequality by taking $Z_1 = X - \mu_X$ and $Z_2 = Y - \mu_Y$.

Can you prove the Cauchy-Schwarz inequality (9) using (7) and (8)?

Uncorrelatedness and Independence: The following summarizes the relation between uncorrelatedness and independence:

1. Two independent random variables X and Y are uncorrelated.

2. There exist numerous examples of pairs of uncorrelated random variables X and Y that are **NOT** independent. Can you think of a few?

3. Two random variables X and Y are independent **if and only if** $g(X)$ and $h(Y)$ are uncorrelated for **every** pair of functions g and h.

5 Correlation and Regression

An important property of $\rho_{X,Y}$ is that it measures the strength of *linear association* between X and Y. This is explained in this section. Consider the problem of *approximating* the random variable Y by a **linear** function $\beta_0 + \beta_1 X$ of X. For given numbers β_0 and β_1, let us measure the accuracy of approximation of Y by $\beta_0 + \beta_1 X$ by the *mean-squared error*:

$$L(\beta_0, \beta_1) := \mathbb{E}\left(Y - \beta_0 - \beta_1 X\right)^2.$$

If $\beta_0 + \beta_1 X$ is a good approximation of Y, then $L(\beta_0, \beta_1)$ should be low. Conversely, if $\beta_0 + \beta_1 X$ is a poor approximation of Y, then $L(\beta_0, \beta_1)$ should be high. What is the smallest possible value of $L(\beta_0, \beta_1)$ as β_0 and β_1 vary over all real numbers.

It can be shown that

$$\min_{\beta_0, \beta_1} L(\beta_0, \beta_1) = Var(Y)\left(1 - \rho_{X,Y}^2\right). \tag{10}$$

Do you know how to prove the above?

The fact (10) precisely captures the interpretation that correlation measures the strength of linear association between Y and X. This is because $\min_{\beta_0,\beta_1} L(\beta_0, \beta_1)$ represents the smallest possible mean squared error in approximating Y by a linear combination of X and (10) says that it is directly related to the correlation between Y and X.

Can you explicitly write down the values of β_0 and β_1 which minimize $L(\beta_0, \beta_1)$?

Does any of the above remind you of linear regression? In what way?

6 Back to Common Distributions

In the last class, we looked at the Binomial and Poisson distributions. Both of these are discrete distributions that can be motivated by coin tossing ($Bin(n, p)$ is the distribution of the number of heads in n independent tosses of a coin with probability of heads p and $Poi(\lambda) \approx Bin(n, \lambda/n)$). We shall now revise two more discrete distributions which arise from coin tossing: the geometric distribution and the negative binomial distribution.

6.1 Geometric Distribution

We say that X has the Geometric distribution with parameter $p \in [0, 1]$ (written as $X \sim Geo(p)$) if X takes the values $1, 2, \dots$ with the probabilities:

$$\mathbb{P}\{X = k\} = (1 - p)^{k-1} p \qquad \text{for } k = 1, 2, \dots.$$

It is easy to see that the number of independent tosses (of a coin with probability of heads p) required to get the first head has the $Geo(p)$ distribution.

The $Geo(p)$ distribution has the interesting property of memorylessness i.e., if $X \sim Geo(p)$, then

$$\mathbb{P}\left\{X > m + n | X > n\right\} = \mathbb{P}\left\{X > m\right\}. \tag{11}$$

This is easy to check as $\mathbb{P}\left\{X > m\right\} = (1 - p)^m$. It is also interesting that the Geometric distribution is the only distribution on $\{1, 2, \dots\}$ which satisfies the memorylessness property (11). To see this, suppose that X is a random variable satisfying (11) which takes values in $\{1, 2, \dots\}$. Let $G(m) := \mathbb{P}\{X > m\}$ for $m = 1, 2, \dots$. Then (11) is the same as

$$G(m + n) = G(m)G(n).$$

This clearly gives $G(m) = (G(1))^m$ for each $m = 1, 2, \ldots$. Now $G(1) = \mathbb{P}\{X > 1\} = 1 - \mathbb{P}\{X = 1\}$. If $p = 1 - \mathbb{P}\{X = 1\}$, then

$$G(m) = (1-p)^m$$

which means that $\mathbb{P}\{X = k\} = \mathbb{P}\{X > k-1\} - \mathbb{P}\{X > k\} = p(1-p)^{k-1}$ for every $k \geq 1$ meaning that X is $Geo(p)$.

6.2 Negative Binomial Distribution

Let X denote the number of tosses (of a coin with probability of heads p) required to get the k^{th} head. The distribution of X is then given by the following. X takes the values $k, k+1, \ldots$ and

$$\begin{aligned}
\mathbb{P}\{X = k+i\} &= \binom{k+i-1}{i} p^k (1-p)^i \\
&= \frac{(k+i-1)(k+i-2)\ldots(k+1)k}{i!} p^k (1-p)^i \\
&= (-1)^i \frac{(-k)(-k-1)(-k-2)\ldots(-k-i+1)}{i!} p^k (1-p)^i \\
&= (-1)^i \binom{-k}{i} p^k (1-p)^i \qquad \text{for } i = 0, 1, 2, \ldots.
\end{aligned}$$

This is called the Negative Binomial distribution with parameters k and p (denoted by $NB(k,p)$). If $G_1, \ldots, G_k$ are independent $Geo(p)$ random variables, then $G_1 + \cdots + G_k \sim NB(k,p)$ (can you prove this?).

7 Continuous Distributions

7.1 Normal or Gaussian Distribution

A random variable X has the normal distribution with mean μ and variance $\sigma^2 > 0$ if it has the following pdf:

$$\phi(x; \mu, \sigma^2) := \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

We write $X \sim N(\mu, \sigma^2)$. When $\mu = 0$ and $\sigma^2 = 1$, we say that X has the *standard* normal distribution and the standard normal pdf is simply denote by $\phi(\cdot)$:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right).$$

Do you know why $\phi(\cdot)$ is a valid density i.e., why $\int e^{-x^2/2} dx = \sqrt{2\pi}$?

The standard normal cdf is denoted by $\Phi(x)$:

$$\Phi(x) := \int_{-\infty}^{x} \phi(t) dt.$$

If $X \sim N(\mu, \sigma^2)$, then $\mathbb{E}(X) = \mu$ and $Var(X) = \sigma^2$. See the corresponding wikipedia page for a list of numerous properties of the normal distribution. The Central Limit Theorem is the main reason why the normal distribution is the most prominent distribution in statistics.

7.2 Uniform Distribution

A random variable U is said to have the uniform distribution on $(0, 1)$ if it has the following pdf:

$$f(x) = \begin{cases} 1 & : 0 < x < 1 \\ 0 & : \text{for all other } x \end{cases}$$

We write $U \sim U[0, 1]$. What is the mean of U? What is the variance of U? Where do uniform distributions arise in statistics? The p-values under the null distribution are usually distributed according to the $U[0, 1]$ distribution (more on this later).

More generally, given an interval (a, b), we say that a random variable U has the uniform distribution on (a, b) if it has the following pdf:

$$f(x) = \begin{cases} \frac{x-a}{b-a} & : a < x < b \\ 0 & : \text{for all other } x \end{cases}$$

We write this as $U \sim U(a, b)$.

7.3 The Exponential Density

The exponential density with rate parameter $\lambda > 0$ (denoted by $Exp(\lambda)$) is given by

$$f(x) = \lambda e^{-\lambda x} I\{x > 0\}.$$

It is arguably the simplest density for modeling random quantities that are constrained to be nonnegative. It is used to model things such as the time of the first phone call that a telephone operator receives starting from now. More generally, it arises as the distribution of inter-arrival times in a Poisson process (more on this later when we study the Poisson Process).

The cdf of $Exp(\lambda)$ is easily seen to be

$$F(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x} \qquad \text{for } x > 0.$$

The exponential density has the memorylessness property (just like the Geometric distribution). Indeed,

$$\mathbb{P}\{X > a + b | X > b\} = \frac{\mathbb{P}\{X > a + b\}}{\mathbb{P}\{X > b\}} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda b}} = e^{-\lambda a} = \mathbb{P}\{X > a\}.$$

In fact, the exponential density is the only density on $(0, \infty)$ that has the memorylessness property (proof left as exercise). In this sense, the Exponential distributionx can be treated as the continuous analogue of the Geometric distribution.

7.4 The Gamma Density

It is customary to talk about the Gamma density after the exponential density. The Gamma density with shape parameter $\alpha > 0$ and rate parameter $\lambda > 0$ is given by

$$f(x) \propto x^{\alpha-1} e^{-\lambda x} I\{x > 0\}. \tag{12}$$

To find the proportionality constant above, we need to evaluate

$$\int_0^\infty x^{\alpha-1} e^{-\lambda x} dx = \frac{1}{\lambda^\alpha} \int_0^\infty u^{\alpha-1} e^{-u} du.$$

Now the function

$$\Gamma(\alpha) := \int_0^\infty u^{\alpha-1} e^{-u} du \qquad \text{for } \alpha > 0$$

is called the Gamma function in mathematics. So the constant of proportionality in (12) is given by

$$\frac{\lambda^\alpha}{\Gamma(\alpha)}$$

so that the Gamma density has the formula:

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I\{x > 0\}.$$

We shall refer to this as the $Gamma(\alpha, \lambda)$ density.

Note that the $Gamma(\alpha, \lambda)$ density reduces to the $Exp(\lambda)$ density when $\alpha = 1$. Therefore, Gamma densities can be treated as a generalization of the Exponential density. In fact, the Gamma density can be seen as the continuous analogue of the negative binomial distribution because if $X_1, \dots, X_k$ are independent $Exp(\lambda)$ random variables, then $X_1 + \cdots + X_n \sim Gamma(k, \lambda)$ (thus the Gamma distribution arises as the sum of i.i.d exponentials just as the Negative Binomial distribution arises as the sum of i.i.d Geometric random variables).

Here are some elementary properties of the Gamma function that will be useful to us later. The Gamma function does not have a closed form expression for arbitrary $\alpha > 0$. However when α is a positive integer, it can be shown that

$$\Gamma(n) = (n-1)! \qquad \text{for } n \geq 1. \tag{13}$$

The above inequality is a consequence of the property

$$\Gamma(\alpha + 1) = \alpha \Gamma(\alpha) \qquad \text{for } \alpha > 0 \tag{14}$$

and the trivial fact that $\Gamma(1) = 1$. You can easily verify (14) by integration by parts.

Another easy fact about the Gamma function is that $\Gamma(1/2) = \sqrt{\pi}$ (this is a consequence of the fact that $\int e^{-x^2/2} dx = \sqrt{2\pi}$).

8 Variable Transformations

It is often common to take functions or transformations of random variables. Consider a random variable X and apply a function $u(\cdot)$ to X to transform X into another random variable $Y = u(X)$. How does one find the distribution of $Y = u(X)$ from the distribution of X?

If X is a discrete random variable, then $Y = u(X)$ will also be discrete and then the pmf of Y can be written directly in terms of the pmf of X:

$$\mathbb{P}\{Y = y\} = \mathbb{P}\{u(X) = y\} = \sum_{x : u(x) = y} \mathbb{P}\{X = x\}.$$

If X is a continuous random variable with density f and $u(\cdot)$ is a smooth function, then it is fairly straightforward to write down the density of $Y = u(X)$ in terms of f. There are some general formulae for doing this but it is better to learn how to do it from first principles. I will illustrate the general idea using the following two examples.

Example 8.1. *Suppose $X \sim U(\pi/2, \pi/2)$. What is the density of $Y = \tan(X)$? Here is the method for doing this from first principles. Note that the range of $\tan(x)$ as x ranges over $(-\pi/2, \pi/2)$ is $\mathbb{R}$ so fix $y \in \mathbb{R}$ and we shall find below the density g of Y at y.*

The formula for $g(y)$ is

$$g(y) = \lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{P}\{y < Y < y + \delta\}$$

so that

$$\mathbb{P}\{y < Y < y + \delta\} \approx g(y)\delta \qquad \textit{when } \delta \textit{ is small.} \tag{15}$$

Now for small δ,

$$\begin{aligned}
\mathbb{P}\{y < Y < y + \delta\} &= \mathbb{P}\{y < \tan(X) < y + \delta\} \\
&= \mathbb{P}\{\arctan(y) < X < \arctan(y + \delta)\} \\
&\approx \mathbb{P}\{\arctan(y) < X < \arctan(y) + \delta \arctan'(y)\} \\
&= \mathbb{P}\{\arctan(y) < X < \arctan(y) + \frac{\delta}{1 + y^2}\} \\
&\approx f(\arctan(y)) \frac{\delta}{1 + y^2}.
\end{aligned}$$

where f is the density of X. Comparing the above with (15), we can conclude that

$$g(y) = f(\arctan(y)) \frac{1}{1 + y^2}$$

Using now the density of $X \sim U(-\pi/2, \pi/2)$, we deduce that

$$g(y) = \frac{1}{\pi(1 + y^2)} \qquad \textit{for } y \in \mathbb{R}.$$

This is the ***Cauchy*** *density. Can you rigorize the above argument? If you were taught formulae for calculating the densities of transformations, does the formula match the answer here?*

Example 8.2. *Suppose $X \sim N(0, 1)$ so that X has the standard normal density $\phi(\cdot)$. What is the density of $Y = X^2$? The following method does this from first principles. The range of X^2 as X ranges over $(-\infty, \infty)$ is $[0, \infty)$ so let us fix $y > 0$. We shall find the density g of Y at y. Write*

$$\begin{aligned}
\mathbb{P}\{y < Y < y + \delta\} &= \mathbb{P}\{\sqrt{y} < X < \sqrt{y + \delta}\} + \mathbb{P}\{-\sqrt{y + \delta} < X < -\sqrt{y}\} \\
&\approx \mathbb{P}\{\sqrt{y} < X < \sqrt{y} + \delta \frac{d\sqrt{y}}{dy}\} + \mathbb{P}\{-\sqrt{y} - \delta \frac{d\sqrt{y}}{dy} < X < -\sqrt{y}\} \\
&= \mathbb{P}\{\sqrt{y} < X < \sqrt{y} + \frac{\delta}{2\sqrt{y}}\} + \mathbb{P}\{-\sqrt{y} - \frac{\delta}{2\sqrt{y}} < X < -\sqrt{y}\} \\
&\approx \phi(\sqrt{y}) \frac{\delta}{2\sqrt{y}} + \phi(-\sqrt{y}) \frac{\delta}{2\sqrt{y}} = \frac{\phi(\sqrt{y})}{\sqrt{y}} \delta.
\end{aligned}$$

This gives

$$g(y) = \frac{\phi(\sqrt{y})}{\sqrt{y}} = (2\pi)^{-1/2} y^{-1/2} e^{-y/2} \qquad \textit{for } y > 0.$$

This is the density of the chi-squared random variable with 1 degree of freedom (or the Gamma random variable with shape parameter $\alpha = 1/2$ and scale parameter $\beta = 1/2$).

9 Distribution Functions and the Quantile Transform

The distribution function (cdf) of a random variable X is defined as

$$F(x) := \mathbb{P}\{X \le x\} \qquad \text{for } -\infty < x < \infty.$$

The cdf has the following properties:

1. It is non-decreasing i.e., $F(x) \leq F(y)$ whenever $x \leq y$.

2. It takes values between 0 and 1.

3. For every x, we have

$$\lim_{y \downarrow x} F(y) = F(x) = \mathbb{P}\{X \leq x\} \quad \text{and} \quad \lim_{y \uparrow x} F(y) = \mathbb{P}\{X < x\} = F(x) - \mathbb{P}\{X = x\}.$$

This implies, in particular, that F is right continuous and that continuity of F is equivalent to $\mathbb{P}\{X = x\} = 0$ for every x.

4. The function $F(x)$ approaches 0 as $x \to -\infty$ and approaches 1 as $x \to +\infty$.

The above properties characterize cdfs in the sense that every function F on $(-\infty, \infty)$ that satisfies these four properties equals the cdf of some random variable. One way to prove this is via the *Quantile Transform.*

Given a function F satisfying the four properties listed above, the corresponding Quantile Transform (or Quantile Function) q_F is a real-valued function on $(0, 1)$ defined as

$$q_F(u) := \inf\{x \in \mathbb{R} : F(x) \geq u\}. \tag{16}$$

The quantile transform can be seen as some kind of an inverse of the cdf F. Indeed, when the cdf F is continuous and strictly increasing, the quantile function q_F is exactly equal to F^{-1}.

The fundamental property of the quantile transform is the following. For $x \in \mathbb{R}$ and $0 < u < 1$:

$$F(x) \geq u \quad \text{if and only if} \quad x \geq q_F(u) \tag{17}$$

Here is the proof of (17). When $F(x) \geq u$, then it is obvious from the definition of q_F that $x \geq q_F(u)$.

On the other hand, again by the definition of $q_F(u)$ and the fact that F is non-decreasing, we have that $F(q_F(u) + \epsilon) \geq u$ for every $\epsilon > 0$. Letting $\epsilon \downarrow 0$ and using the right-continuity of F, we deduce that $F(q_F(u)) \geq u$. This implies that when $x \geq q_F(u)$, we have $F(x) \geq F(q_F(u)) \geq u$. This proves (17).

An important fact about the quantile transform is:

$$\mathbb{P}\{X \leq q_F(u)\} \geq u \quad \text{and} \quad \mathbb{P}\{X < q_F(u)\} \leq u. \tag{18}$$

Therefore, $q_F(u)$ is a quantile of X (when $u = 0.5$, it follows that $q_F(0.5)$ is a median of X). Hence the name quantile transform. The first inequality above is a consequence of

$$\mathbb{P}\{X \leq q_F(u)\} = F(q_F(u)) \geq u.$$

To prove the second inequality, note that by definition of $q_F(u)$,

$$F(q_F(u) - \epsilon) < u \qquad \text{for every } \epsilon > 0.$$

Letting $\epsilon \downarrow 0$, we obtain that $\mathbb{P}\{X < q_F(u)\} \leq u$ which proves (18).

The following result is an important application of the use of the quantile transform.

Proposition 9.1. *The following two statements are true.*

1. *Suppose U is a random variable distributed according to the uniform distribution on $(0, 1)$. Then $q_F(U)$ has cdf F.*

2. *Suppose X is a random variable with a* ***continuous*** *cdf F. Then $F(X)$ has the uniform distribution on $(0, 1)$.*

Proof. For the first part, note first that the cdf F_U of the uniform random variable U satisfies $F_U(u) = u$ for $0 \leq u \leq 1$. Thus the cdf F_X of $X = q_F(U)$ is given by

$$F_X(x) = \mathbb{P}\{X \leq x\} = \mathbb{P}\{q_F(U) \leq x\} = \mathbb{P}\{U \leq F(x)\} = F(x)$$

where we crucially used (17). This proves that X has cdf F.

For the second part, assume that F is continuous. Note that for every $\epsilon > 0$, the definition of q_F implies that $F(q_F(u) - \epsilon) < u$. Letting $\epsilon \to 0$ and using the continuity of F, we deduce that $F(q_F(u)) \leq u$. Combining with (17), this gives $F(q_F(u)) = u$. Therefore for every $0 < u < 1$, we have

$$\mathbb{P}\{F(X) \geq u\} = \mathbb{P}\{X \geq q_F(u)\} = 1 - \mathbb{P}\{X < q_F(u)\} = 1 - \mathbb{P}\{X \leq q_F(u)\} = 1 - F(q_F(u)) = 1 - u$$

where we have used the continuity of F (which implies that $\mathbb{P}\{X = x\}$ for every x). The fact that $\mathbb{P}\{F(X) \geq u\} = 1 - u$ for every $0 < u < 1$ implies that $F(X) \sim U(0, 1)$. □

Example 9.2 (p-values corresponding to test statistics having continuous distributions have uniform distributions under the null hypothesis)**.** *Statistical hypothesis testing problems are usually formed by calculating a relevant test statistic based on data. Suppose T_{obs} is the observed value of the statistic calculated from the data. The p-value corresponding to the test is defined as the probability, under the null hypothesis, of observing a value for the statistic that is more extreme compared to T_{obs}. Usually this is calculated as*

$$p = 1 - F_0(T_{obs})$$

where F_0 is the cdf of the test statistic under the null hypothesis. If F_0 is a continuous cdf, then it should be clear that p is distributed according to $U(0, 1)$ when $T_{obs} \sim F_0$. In other words, under the null distribution (i.e., $T_{obs} \sim F_0$), the p-value has the standard uniform distribution.

10 Joint Densities

Joint densities are used to describe the distribution of a finite set of continuous random variables. We focus on bivariate joint densities (i.e., when there are two continuous variables X and Y). The ideas are the same for the case of more than two variables.

The following are the main points to remember about joint densities:

1. $f(\cdot, \cdot)$ is called a joint density if

$$f(x, y) \geq 0 \quad \text{for all } x, y \qquad \text{and} \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1.$$

2. We say that two random variables X and Y have joint density $f(\cdot, \cdot)$ if

$$\mathbb{P}\left\{(X, Y) \in B\right\} = \int \int_B f(x, y) dx dy = \int \int I\{(x, y) \in B\} f(x, y) dx dy.$$

   for every subset B of $\mathbb{R}^2$. We shall often denote the joint density of (X, Y) by $f_{X,Y}$.

3. If Δ is a small region in $\mathbb{R}^2$ around a point (x_0, y_o), we have (under some regularity condition on the behavior of $f_{X,Y}$ at (x_0, y_0))

$$\mathbb{P}\{(X, Y) \in \Delta\} \approx (\text{area of } \Delta)\, f_{X,Y}(x_0, y_0).$$

   More formally,

$$\lim_{\Delta \downarrow (x_0, y_0)} \frac{\mathbb{P}\{(X, Y) \in \Delta\}}{\text{area of } \Delta} = f_{X,Y}(x_0, y_0)$$

   where the limit is taken as Δ shrinks to (x_0, y_0).

4. If (X, Y) have joint density $f_{X,Y}$, then the density of X is given by f_X and the density of Y is f_Y where

$$f_X(x) = \int f_{X,Y}(x,y)dy \quad \text{and} \quad f_Y(y) = \int f_{X,Y}(x,y)dx.$$

   The densities f_X and f_Y are referred to as the *marginal* densities of X and Y respectively.

5. **Independence and Joint Densities**: The following statements are equivalent:
   (a) The random variables X and Y are independent.
   (b) The joint density $f_{X,Y}(x)$ factorizes into the product of a function depending on x alone and a function depending on y alone.
   (c) $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x, y.

Example 10.1. *Consider the function*

$$f(x,y) = \begin{cases} 1 & : 0 \le x \le 1 \textit{ and } 0 \le y \le 1 \\ 0 & : \textit{otherwise} \end{cases}$$

Check that this is indeed a density function. This density takes the value 1 on the unit square. If the random variables X and Y have this density f, then we say that they are uniformly distributed on the unit square. Using indicator functions, we can write this density also as:

$$f(x,y) = I\{0 \le x \le 1, 0 \le y \le 1\} = I\{0 \le x \le 1\}I\{0 \le y \le 1\}$$

The factorization above immediately says that if $f = f_{X,Y}$, then X and Y are independent. The marginal densities of X and Y are uniform densities on $[0,1]$.

Question: If X, Y have this density f, calculate $\mathbb{P}\{X^2 + Y^2 \le 1\}$ (Ans: $\pi/4$).

Example 10.2. *Suppose X, Y have the joint density*

$$f_{XY}(x,y) = \frac{1}{\pi} I\{x^2 + y^2 \le 1\}.$$

Show that the marginal density of X is given by

$$f_X(x) = \frac{2}{\pi}\sqrt{1 - x^2} I\{-1 \le x \le 1\}.$$

Are X and Y independent? (Ans: No. Why?)

11 Joint Densities under Transformations

We address the following general question. Suppose X and Y have the joint density $f_{X,Y}$. Suppose now that we consider two new random variables defined by

$$(U, V) := T(X, Y)$$

where $T : \mathbb{R}^2 \to \mathbb{R}^2$ is a differentiable and invertible function. What is the joint density $f_{U,V}$ of U, V in terms of $f_{X,Y}$?

The following simple example will nicely motivate the general ideas.

Example 11.1. *Suppose X, Y have joint density $f_{X,Y}$. What is the joint density of U and V where $U = X$ and $V = X + Y$?*

We see that $(U, V) = T(X, Y)$ where $T(x, y) = (x, x + y)$. This transformation T is clearly invertible and its inverse is given by $S(u, v) = T^{-1}(u, v) = (u, v - u)$. In order to determine the joint density of (U, V) at a point (u, v), let us consider

$$\mathbb{P}\{u \le U \le u + \delta, v \le V \le v + \epsilon\} \approx \delta\epsilon f_{U,V}(u, v). \tag{19}$$

Let R denote the rectangle joining the points $(u, v), (u + \delta, v), (u, v + \epsilon)$ and $(u + \delta, v + \epsilon)$. Then the above probability is the same as

$$\mathbb{P}\{(U, V) \in R\} = \mathbb{P}\{(X, Y) \in S(R)\}.$$

What is the region $S(R)$? It is easy to see that this is the parallelogram joining the points $(u, v - u), (u + \delta, v - u - \delta)$, $(u, v - u + \epsilon)$ and $(u + \delta, v - u + \epsilon - \delta)$. When δ and ϵ are small, $S(R)$ is clearly a small region around $(u, v - u)$ which allows us to write

$$\mathbb{P}\{(U, V) \in R\} = \mathbb{P}\{(X, Y) \in S(R)\} \approx f_{X,Y}(u, v - u)\,(\text{area of } S(R))\,.$$

The area of the parallelogram $S(R)$ can be computed to be $\delta\epsilon$ (using the formula that the area of a parallelogram equals base times height) so that

$$\mathbb{P}\{(U, V) \in R\} \approx f_{X,Y}(u, v - u)\delta\epsilon.$$

Comparing with (19), we obtain

$$f_{U,V}(u, v) = f_{X,Y}(u, v - u).$$

This gives the formula for the joint density of (U, V) in terms of the joint density of (X, Y).

11.1 Detour to Convolutions

We shall come back to the general problem of finding densities of transformations after taking a short detour to convolutions.

We proved in the above example the joint density of $U = X$ and $V = X + Y$ is given by

$$f_{U,V}(u, v) = f_{X,Y}(u, v - u)$$

where $f_{X,Y}$ is the joint density of (X, Y). As a consequence, we see that the density of $V = X + Y$ is given by

$$f_{X+Y}(v) = \int_{-\infty}^{\infty} f_{U,V}(u, v)du = \int_{-\infty}^{\infty} f_{X,Y}(u, v - u)du.$$

Suppose now that X and Y are independent. Then $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ and consequently

$$f_{X+Y}(v) = \int_{-\infty}^{\infty} f_X(u)f_Y(v - u)du = \int_{-\infty}^{\infty} f_X(v - w)f_Y(w)dw \tag{20}$$

where the last equality is a consequence of a simple change of variable $v - u = w$.

Definition 11.2 (Convolution)**.** *Given two densities f_1 and f_2, we define their convolution, $f_1 \star f_2$ to be the density:*

$$(f_1 \star f_2)(v) := \int f_1(u)f_2(v - u)du = \int f_1(v - u)f_2(u)du.$$

The equation (20) therefore says, in words, that the density of $X + Y$, where $X \sim f_X$ and $Y \sim f_Y$ are independent, equals the convolution of f_X and f_Y.

Example 11.3. *Suppose X and Y are independent random variables which are exponentially distributed with rate parameter λ. What is the distribution of $X + Y$?*

By the convolution formula,

$$\begin{aligned} f_{X+Y}(v) &= \int_{-\infty}^{\infty} f_X(u) f_Y(v-u) du \\ &= \int_{-\infty}^{\infty} \lambda e^{-\lambda u} I\{u > 0\} \lambda e^{-\lambda(v-u)} I\{v - u > 0\} du \\ &= \int_{-\infty}^{\infty} \lambda^2 e^{-\lambda v} I\{0 < u < v\} du = \lambda^2 v e^{-\lambda v} I\{v > 0\}. \end{aligned}$$

This shows that $X + Y$ has the Gamma distribution with shape parameter 2 and rate parameter λ.

Example 11.4. *Suppose X and Y are independent random variables that are uniformly distributed on $[0, 1]$. What is the density of $X + Y$?*

By the convolution formula,

$$\begin{aligned} f_{X+Y}(v) &= \int_{-\infty}^{\infty} f_X(u) f_Y(v-u) du \\ &= \int I\{0 \le u \le 1\} I\{0 \le v - u \le 1\} du \\ &= \int_{-\infty}^{\infty} I\{0 \le u \le 1, 0 \le v - u \le 1\} du \\ &= \int_{-\infty}^{\infty} I\{\max(v-1, 0) \le u \le \min(v, 1)\} du. \end{aligned}$$

This integral is non-zero only when $\max(v-1, 0) \le \min(v, 1)$ which is easily seen to be equivalent to $0 \le v \le 2$. When $0 \le v \le 2$, we have

$$f_{X+Y}(v) = \min(v, 1) - \max(v - 1, 0)$$

which can be simplified as

$$f_{X+Y}(v) = \begin{cases} v & : 0 \le v \le 1 \\ 2 - v & : 1 \le v \le 2 \\ 0 & : \textit{otherwise} \end{cases}$$

This is called the triangular density.

12 Joint Densities under transformations

In the last class, we calculated the joint density of $(X, X + Y)$ in terms of the joint density of (X, Y). In this lecture, we generalize the idea behind that calculation by first calculating the joint density of a linear and invertible transformation of a pair of random variables. We also deal with the case of a non-linear and invertible transformation.

In the next subsection, we shall recall some standard properties of linear transformations.

12.1 Linear Transformations

By a linear transformation $L : \mathbb{R}^2 \to \mathbb{R}^2$, we mean a function that is given by

$$L(x, y) := M \begin{pmatrix} x \\ y \end{pmatrix} + c \tag{21}$$

where M is a 2×2 matrix and c is a 2×1 vector. The first term on the right hand side above involves multiplication of the 2×2 matrix M with the 2×1 vector with components x and y.

We shall refer to the 2×2 matrix M as the matrix corresponding to the linear transformation L and often write M_L for the matrix M.

The linear transformation L in (21) is invertible if and only if the matrix M is invertible. We shall only deal with invertible linear transformations in the sequel. The following are two standard properties of linear transformations that you need to familiar with for the sequel.

1. If P is a parallelogram in $\mathbb{R}^2$, then $L(P)$ is also a parallelogram in $\mathbb{R}^2$. In other words, linear transformations map parallelograms to parallelograms.
2. For every parallelogram P, the following identity holds:
$$\frac{\text{area of } L(P)}{\text{area of } P} = |\det(M_L)|.$$
In other words, the ratio of the areas of $L(P)$ to that of P is given by the absolute value of the determinant of the matrix M_L.

12.2 Invertible Linear Transformations

Suppose X, Y have joint density $f_{X,Y}$ and let $(U, V) = T(X, Y)$ for a linear and invertible transformation $T : \mathbb{R}^2 \to \mathbb{R}^2$. Let the inverse transformation of T be denoted by S. In the example of the previous lecture, we hacd $T(x, y) = (x, x + y)$ and $S(u, v) = (u, v - u)$. The fact that T is assumed to be linear and invertible means that S is also linear and invertible.

To compute $f_{U,V}$ at a point (u, v), we consider
$$\mathbb{P}\{u \le U \le u + \delta, v \le V \le v + \epsilon\} \approx f_{U,V}(u, v)\delta\epsilon$$
for small δ and ϵ. Let R denote the rectangle joining the points $(u, v), (u + \delta, v), (u, v + \epsilon)$ and $(u + \delta, v + \epsilon)$. Then the above probability is the same as
$$\mathbb{P}\{(U, V) \in R\} = \mathbb{P}\{(X, Y) \in S(R)\}.$$
What is the region $S(R)$? Clearly now $S(R)$ is a small region (as δ and ϵ are small) around the point $S(u, v)$ so that
$$\mathbb{P}\{(U, V) \in R\} = \mathbb{P}\{(X, Y) \in S(R)\} \approx f_{X,Y}(S(u, v)) \, (\text{area of } S(R)) \, .$$
By the facts mentioned in the previous subsection, we now note that $S(R)$ is a parallelogram whose area equals $|det(M_S)|$ multiplied by the area of R (note that the area of R equals $\delta\epsilon$). We thus have
$$f_{U,V}(u, v)\delta\epsilon \approx \mathbb{P}\left\{(U, V) \in R\right\} = \mathbb{P}\left\{(X, Y) \in S(R)\right\} = f_{X,Y}(S(u, v))|\det(M_S)|\delta\epsilon$$
which allows us to deduce that
$$f_{U,V}(u, v) = f_{X,Y}(S(u, v)) \, |\det M_S| \, . \tag{22}$$
It is helpful to remember here that M_S is the 2×2 matrix corresponding to the linear transformation S.

Example 12.1. *Suppose X and Y are independent standard normal random variables. Find the joint density of $U = X + Y$ and $V = X - Y$.*

We can use the formula (22) with $T(x, y) = (x + y, x - y)$ whose inverse transformation is $S(u, v) = (\frac{u+v}{2}, \frac{u-v}{2})$ and clearly the matrix corresponding to S is given by $M_S = \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & -1/2 \end{pmatrix}$. The formula (22) *then gives*

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right)|\det M_S| \\ &= f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right)\left|\det\begin{pmatrix}1/2 & 1/2 \\ 1/2 & -1/2\end{pmatrix}\right| = \frac{1}{2} f_{X,Y}\left(\frac{u+v}{2}, \frac{u-v}{2}\right). \end{aligned}$$

Because X and Y are independent standard normals, we have

$$f_{X,Y}(x,y) = \frac{1}{2\pi}\exp\left(\frac{-(x^2+y^2)}{2}\right)$$

so that

$$f_{U,V}(u,v) = \frac{1}{4\pi}e^{-u^2/4}e^{-v^2/4}.$$

This implies that U and V are independent $N(0,2)$ random variables.

Example 12.2. *Suppose X and Y are independent standard normal random variables. Then what is the distribution of $(U,V) = T(X,Y)$ where*

$$T(x,y) = (x\cos\theta - y\sin\theta, x\sin\theta + y\cos\theta).$$

Geometrically the transformation T corresponds to rotating the point (x,y) by an angle θ in the counter clockwise direction. The inverse transformation $S := T^{-1}$ of T is given by

$$S(u,v) = (u\cos\theta + v\sin\theta, -u\sin\theta + v\cos\theta)$$

and this corresponds to rotating the point (u,v) clockwise by an angle θ. The matrix corresponding to S is

$$M_S = \begin{pmatrix}\cos\theta & \sin\theta \\ -\sin\theta & \cos\theta\end{pmatrix}$$

The formula (22) then gives

$$\begin{aligned} f_{U,V}(u,v) &= f_{X,Y}(u\cos\theta + v\sin\theta, -u\sin\theta + v\cos\theta)\left|\det\begin{pmatrix}\cos\theta & \sin\theta \\ -\sin\theta & \cos\theta\end{pmatrix}\right| \\ &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}(u\cos\theta + v\sin\theta)^2 - \frac{1}{2}(-u\sin\theta + v\cos\theta)^2\right) = \frac{1}{2\pi}\exp\left(-\frac{u^2}{2} - \frac{v^2}{2}\right). \end{aligned}$$

This means that U and V are independent random variables each having the standard normal distribution.

We shall next study the problem of obtaining the joint densities under differentiable and invertible transformations that are not necessarily linear.

12.3 General Invertible Transformations

Let (X,Y) have joint density $f_{X,Y}$. We transform (X,Y) to two new random variables (U,V) via $(U,V) = T(X,Y)$. What is the joint density $f_{U,V}$? Suppose that T is invertible (having an inverse $S = T^{-1}$) and differentiable. Note that S and T are not necessarily linear transformations.

In order to compute $f_{U,V}$ at a point (u,v), we consider

$$\mathbb{P}\{u \le U \le u+\delta, v \le V \le v+\epsilon\} \approx f_{U,V}(u,v)\delta\epsilon$$

for small δ and ϵ. Let R denote the rectangle joining the points $(u,v), (u+\delta,v), (u,v+\epsilon)$ and $(u+\delta,v+\epsilon)$. Then the above probability is the same as

$$\mathbb{P}\{(U,V) \in R\} = \mathbb{P}\{(X,Y) \in S(R)\}.$$

What is the region $S(R)$? If S is linear then $S(R)$ (as we have seen previously) will be a parallelogram. For general S, the main idea is that, as long as δ and ϵ are small, the region $S(R)$ can be approximated by a parallelogram. This is because S itself can be approximated by a linear transformation on the region R. To see this, let us write the function $S(a,b)$ as $(S_1(a,b), S_2(a,b))$ where S_1 and S_2 map points in $\mathbb{R}^2$ to $\mathbb{R}$. Assuming that S_1 and S_2 are differentiable, we can approximate $S_1(a,b)$ for (a,b) near (u,v) by

$$S_1(a,b) \approx S_1(u,v) + \left(\frac{\partial}{\partial u} S_1(u,v), \frac{\partial}{\partial v} S_1(u,v)\right) \begin{pmatrix} a-u \\ b-v \end{pmatrix} = S_1(u,v) + (a-u)\frac{\partial}{\partial u} S_1(u,v) + (b-v)\frac{\partial}{\partial v} S_1(u,v).$$

Similarly, we can approximate $S_2(a,b)$ for (a,b) near (u,v) by

$$S_2(a,b) \approx S_2(u,v) + \left(\frac{\partial}{\partial u} S_2(u,v), \frac{\partial}{\partial v} S_2(u,v)\right) \begin{pmatrix} a-u \\ b-v \end{pmatrix}.$$

Putting the above two equations together, we obtain that, for (a,b) close to (u,v),

$$S(a,b) \approx S(u,v) + \begin{pmatrix} \frac{\partial}{\partial u} S_1(u,v) & \frac{\partial}{\partial v} S_1(u,v) \\ \frac{\partial}{\partial u} S_2(u,v) & \frac{\partial}{\partial v} S_2(u,v) \end{pmatrix} \begin{pmatrix} a-u \\ b-v \end{pmatrix}.$$

Therefore S can be appromixated by a linear transformation with matrix given by

$$J_S(u,v) := \begin{pmatrix} \frac{\partial}{\partial u} S_1(u,v) & \frac{\partial}{\partial v} S_1(u,v) \\ \frac{\partial}{\partial u} S_2(u,v) & \frac{\partial}{\partial v} S_2(u,v) \end{pmatrix}$$

for (a,b) near (u,v). Note that, in particular, when δ and ϵ are small, that this linear appximation for S is valid over the region R. The matrix $J_S(u,v)$ is called the Jacobian matrix of $S(u,v) = (S_1(u,v), S_2(u,v))$ at the point (u,v).

Because of the above linear approximation, we can write

$$\mathbb{P}\{(X,Y) \in S(R)\} \approx f_{X,Y}(S(u,v)) \left|\det(J_S(u,v))\right| (\text{area of } R)$$

This gives us the important formula

$$f_{U,V}(u,v) = f_{X,Y}(S(u,v)) \left|\det J_S(u,v)\right|. \tag{23}$$

Example 12.3. *Suppose X and Y have joint density $f_{X,Y}$. What is the joint density of $U = X/Y$ and $V = Y$?*

We need to compute the joint density of $(U,V) = T(X,Y)$ where $T(x,y) = (x/y, y)$. The inverse of this transformation is $S(u,v) = (uv, v)$. Then formula (23) gives

$$f_{U,V}(u,v) = f_{X,Y}(uv,v) \left|\det \begin{pmatrix} v & u \\ 0 & 1 \end{pmatrix}\right| = f_{X,Y}(uv,v)|v|.$$

As a consequence, the marginal density of $U = X/Y$ is given by

$$f_U(u) = \int f_{U,V}(u,v)dv = \int f_{X,Y}(uv,v)|v|dv.$$

In the special case when X and Y are independent standard normal random variables, the density of $U = X/Y$ is given by

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{(1+u^2)v^2}{2}\right) |v|dv \\ &= 2\int_0^{\infty} \frac{1}{2\pi} \exp\left(-\frac{(1+u^2)v^2}{2}\right) vdv = \frac{1}{\pi(1+u^2)}. \end{aligned}$$

This is the standard Cauchy density.

Example 12.4. *Suppose X and Y are independent standard normal random variables. Let $R := \sqrt{X^2+Y^2}$ and let Θ denote the angle made by the vector (X,Y) with the positive X-axis in the plane. What is the joint density of (R,Θ)?*

Clearly $(R,\Theta) = T(X,Y)$ where the inverse of T is given by $S(r,\theta) = (r\cos\theta, r\sin\theta)$. The density of f (R,Θ) at (r,θ) is zero unless $r > 0$ and $0 < \theta < 2\pi$. The formula (23) then gives

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta)\left|\det\begin{pmatrix}\cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta\end{pmatrix}\right| = \frac{1}{2\pi}e^{-r^2/2}rI\{r>0\}I\{0<\theta<2\pi\}.$$

It is easy to see from here that Θ is uniformly distributed on $(0,2\pi)$ and R has the density

$$f_R(r) = re^{-r^2/2}I\{r>0\}.$$

Moreover R and Θ are independent. The density of R is called the Rayleigh density.

Example 12.5. *Here is an important fact about Gamma distributions: Suppose $X \sim Gamma(\alpha_1,\lambda)$ and $Y \sim Gamma(\alpha_2,\lambda)$ are independent, then $X+Y \sim Gamma(\alpha_1+\alpha_2,\lambda)$. This can be proved using the convolution formula for densities of sums of independent random variables. A different formula uses the Jacobian formula to derive the joint density of U and V where $V = X/(X+Y)$. The relevant inverse transformation here $S(u,v) = (uv, u-uv)$ so that the Jacobian formula gives:*

$$f_{U,V}(u,v) = f_{X,Y}(uv, u(1-v))u = f_X(uv)f_Y(u(1-v))u.$$

Plugging in the relevant Gamma densities for f_X and f_Y, we can deduce that

$$f_{U,V}(u,v) = \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1+\alpha_2)}u^{\alpha_1+\alpha_2-1}e^{-\lambda u}I\{u>0\}\frac{\Gamma(\alpha_1+\alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)}v^{\alpha_1-1}(1-v)^{\alpha_2-1}I\{0<v<1\}.$$

This implies that $U \sim Gamma(\alpha_1+\alpha_2,\lambda)$. It also implies that $V \sim Beta(\alpha_1,\alpha_2)$, that U and V are independent as well as

$$B(\alpha_1,\alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1+\alpha_2)}$$

where, on the right hand side above, we have the Beta function. Note that because $\Gamma(n) = (n-1)!$ for when n is an integer, the above formiula gives us a way to calculate the Beta function $B(\alpha_1,\alpha_2)$ when α_1 and α_2 are positive integers.

13 Joint Densities under Non-Invertible Transformations

In the last class, we looked at the Jacobian formula for calculating the joint density of a transformed set of continuous random variables in terms of the joint density of the original random variables. This formula assumed that the transformation is invertible. In other words, the formula does not work if the transformation is non-invertible. However, the general method based on first principles (that we used to derive the Jacobian formula) works fine. This is illustrated in the following example.

Example 13.1 (Order Statistics)**.** *Suppose X and Y have joint density $f_{X,Y}$. What is the joint density of $U = \min(X,Y)$ and $V = \max(X,Y)$?*

Let us find the joint density of (U,V) at (u,v). Since $U < V$, the density $f_{U,V}(u,v)$ will be zero when $u \geq v$. So let $u < v$. For δ and ϵ small, let us consider

$$\mathbb{P}\{u \leq U \leq u+\delta, v \leq V \leq v+\epsilon\}.$$

If δ and ϵ are much smaller compared to $v-u$, then the above probability equals

$$\mathbb{P}\{u \leq X \leq u+\delta, v \leq Y \leq v+\epsilon\} + \mathbb{P}\{u \leq Y \leq u+\delta, v \leq X \leq v+\epsilon\}$$

which is further approximately equal to

$$f_{X,Y}(u,v)\delta\epsilon + f_{X,Y}(v,u)\delta\epsilon.$$

We have thus proved that

$$f_{U,V}(u,v) = \begin{cases} f_{X,Y}(u,v) + f_{X,Y}(v,u) & : u < v \\ 0 & : \textit{otherwise} \end{cases}$$

We can generalize this to the case of more than two random variables. Suppose $X_1, \dots, X_n$ are random variables having a joint density $f_{X_1,\dots,X_n}(x_1,\dots,x_n)$. Let $X_{(1)} \le \dots \le X_{(n)}$ denote the ***order statistics*** *of $X_1, \dots, X_n$ i.e., $X_{(1)}$ is the smallest value among $X_1, \dots, X_n$, $X_{(2)}$ is the next smallest value and so on with $X_{(n)}$ denoting the largest value. What then is the joint distribution of $X_{(1)}, \dots, X_{(n)}$. The calculation above for the case of the two variables can be easily generalized to obtain*

$$f_{X_{(1)},\dots,X_{(n)}}(u_1,\dots,u_n) = \begin{cases} \sum_\pi f_{X_1,\dots,X_n}(u_{\pi_1},\dots,u_{\pi_n}) & : u_1 < u_2 < \dots < u_n \\ 0 & : \textit{otherwise} \end{cases}$$

where the sum is over all permutations π (i.e, one-one and onto functions mapping $\{1,\dots,n\}$ to $\{1,\dots,n\}$).

When the variables $X_1, \dots, X_n$ are i.i.d (independent and identically distributed), then it follows from the above that

$$f_{X_{(1)},\dots,X_{(n)}}(u_1,\dots,u_n) = \begin{cases} (n!) f_{X_1}(u_1)\dots f_{X_n}(u_n) & : u_1 < u_2 < \dots < u_n \\ 0 & : \textit{otherwise} \end{cases}$$

14 More on Order Statistics: The density of $X_{(i)}$ for a fixed i

Assume now that $X_1, \dots, X_n$ are i.i.d random variables with a common density f. In the previous section, we derived the joint density of the order statistics $X_{(1)}, \dots, X_{(n)}$. Here we focus on the problem of determining the density of $X_{(i)}$ for a fixed i. The answer is given by

$$f_{X_{(i)}}(u) = \frac{n!}{(n-i)!(i-1)!} (F(u))^{i-1} (1 - F(u))^{n-i} f(u). \tag{24}$$

and there are three standard methods for deriving this.

14.1 Method One

The first method integrates the joint density $f_{X_{(1)},\dots,X_{(n)}}(u_1,\dots,u_{i-1},u,u_{i+1},\dots u_n)$ over $u_1,\dots,u_{i-1},u_{i+1},\dots,u_n$ to obtain $f_{X_{(i)}}(u)$. More precisely,

$$f_{X_{(i)}}(u) = \int \cdots \int n! f(u_1)\dots f(u_{i-1}) f(u) f(u_{i+1}) \dots f(u_n) I\{u_1 < \dots < u_n\} du_1 \dots du_{i-1} du_{i+1} \dots du_n$$

Integrate the above first with respect to u_1 (in the range $(-\infty, u_2)$), then with respect to u_2 (in the range of $(-\infty, u_3)$) and all the way up to the integral with respect to u_{i-1}. Then integrate with respect to u_n, then with respect to u_{n-1} and all the way to u_{i+1}. This will lead to (24).

14.2 Method Two

This method uses multinomial probabilities. Suppose that we repeat an experiment n times and that the outcomes of the n repetitions are independent. Suppose that each individual experiment has k outcomes which we denote by $O_1, \dots, O_k$ and let the probabilities of these outcomes be given by $p_1, \dots, p_k$ (note that these are nonnegative numbers which sum to one).

Now let N_i denote the number of times (over the n repetitions) that the outcome O_i appeared (note that $N_1, \dots, N_k$ are nonnegative integers which sum to n). The joint distribution of $(N_1, \dots, N_k)$ is known as the multinomial distribution with parameters n and $p_1, \dots, p_k$. It is an exercise to show that

$$\mathbb{P}\{N_1 = n_1, N_2 = n_2, \dots, N_k = n_k\} = \frac{n!}{n_1! \dots n_k!} p_1^{n_1} \dots p_k^{n_k} \tag{25}$$

whenever $n_1, \dots, n_k$ are nonnegative integers which sum to n.

Let us now get back to the problem of obtaining the density of $X_{(i)}$. Consider the probability

$$\mathbb{P}\left\{u \le X_{(i)} \le u + \delta\right\}$$

for a fixed u and small δ. If δ is small, then this probability can be approximated by the probability of the event E where E is defined as follows. E is the event where $(i-1)$ observations among $X_1, \dots, X_n$ are strictly smaller than u, one observation among $X_1, \dots, X_n$ lies in $[u, u+\delta]$ and $n-i$ observations among $X_1, \dots, X_n$ are strictly larger than $u + \delta$. This latter probability is a special case of the multinomial probability formula (25) and when δ is small, we get that this probability equals

$$\frac{n!}{(n-i)!(i-1)!} \left(F(u)\right)^{i-1} \left(f(u)\delta\right) \left(1 - F(u)\right)^{n-i}$$

where F is the cdf corresponding to f. The formula (24) then immediately follows.

14.3 Method Three

Here we first compute the cdf of $X_{(i)}$ and then differentiate it to get the pdf. Note that

$$\begin{aligned}
F_{X_{(i)}}(x) &= \mathbb{P}\left\{X_{(i)} \le x\right\} \\
&= \mathbb{P}\{\text{at least } i \text{ of } X_1, \dots, X_n \text{ are } \le x\} \\
&= \sum_{r=i}^{n} \mathbb{P}\{\text{exactly } r \text{ of } X_1, \dots, X_n \text{ are } \le x\} \\
&= \sum_{r=i}^{n} \mathbb{P}\{Bin(n, F(x)) = r\} = \sum_{r=i}^{n} \binom{n}{r} (F(x))^r (1 - F(x))^{n-r}.
\end{aligned}$$

To compute the density, we have to differentiate $F_{X_{(i)}}$ with respect to x. This gives (note that the derivative of F is f)

$$\begin{aligned}
f_{X_{(i)}}(x) &= \sum_{r=i}^{n} \binom{n}{r} \left\{r(F(x))^{r-1} f(x) (1 - F(x))^{n-r} - (F(x))^r (n-r)(1 - F(x))^{n-r-1} f(x)\right\} \\
&= \sum_{r=i}^{n} \frac{n!}{(n-r)!(r-1)!} (F(x))^{r-1} (1 - F(x))^{n-r} f(x) - \sum_{r=i}^{n-1} \frac{n!}{(n-r-1)!r!} (F(x))^r (1 - F(x))^{n-r-1} f(x) \\
&= \sum_{r=i}^{n} \frac{n!}{(n-r)!(r-1)!} (F(x))^{r-1} (1 - F(x))^{n-r} f(x) - \sum_{s=i+1}^{n} \frac{n!}{(n-s)!(s-1)!} (F(x))^{s-1} (1 - F(x))^{n-s} f(x) \\
&= \frac{n!}{(n-i)!(i-1)!} \left(F(u)\right)^{i-1} \left(1 - F(u)\right)^{n-i} f(u)
\end{aligned}$$

and thus we again get the formula (24).

Next we look at some special instances of the formula (24) for the density of individual order statistics.

14.4 Uniform Order Statistics

Suppose $X_1, \dots, X_n$ are i.i.d having the uniform density on $(0, 1)$. Then the formula (24) (by plugging in $f(u) = 1$ and $F(u) = u$ for $0 < u < 1$) gives the following density for $X_{(i)}$:

$$f_{X_{(i)}}(u) = \frac{n!}{(n-i)!(i-1)!} u^{i-1}(1-u)^{n-i} \qquad \text{for } 0 < u < 1. \tag{26}$$

This is a Beta density with parameters i and $n - i + 1$. Generally, a Beta density with parameters $\alpha > 0$ and $\beta > 0$ is given by

$$f(u) = \frac{u^{\alpha-1}(1-u)^{\beta-1}}{\int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx} I\{0 < u < 1\}$$

The integral in the denominator above is called the Beta function:

$$B(\alpha, \beta) := \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx \qquad \text{for } \alpha > 0, \beta > 0.$$

The Beta function does not usually have a closed form expression but we can write it in terms of the Gamma function via the formula

$$B(\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

that we saw in the last class. This formula allows us to write $B(\alpha, \beta)$ in closed form when α and β are integers (note that $\Gamma(n) = (n-1)!$). This gives, for example,

$$B(i, n-i+1) = \frac{n!}{(n-i)!(i-1)!}.$$

14.5 Maximum of Independent Uniforms

Suppose $X_1, \dots, X_n$ are independent random variables having the uniform distribution on the interval $(0, \theta)$ for some $\theta > 0$. It turns out that the maximum order statistic, $X_{(n)}$, is the maximum likelihood estimate of θ. The density of $X_{(n)}$ is easily seen to be (as a consequence of (24)):

$$f_{X_{(n)}}(u) = \frac{nu^{n-1}}{\theta^n} I\{0 < u < \theta\}.$$

What then is $\mathbb{E}(X_{(n)})$? Using the formula for the density above,

$$\mathbb{E}X_{(n)} = \int_0^\theta \frac{unu^{n-1}}{\theta^n} du = \frac{n\theta}{n+1}.$$

This means therefore that $X_{(n)}$ has a slight negative bias of $-\theta/(n+1)$ as an estimator for θ and that $((n+1)/n)X_{(n)}$ is an unbiased estimator of θ.

14.6 Minimum of Independent Exponentials

The exponential density with rate parameter $\lambda > 0$ (denoted by $Exp(\lambda)$) is given by

$$f(x) = \lambda e^{-\lambda x} I\{x > 0\}.$$

It is arguably the simplest density for modeling random quantities that are constrained to be nonnegative. It arises as the distribution of inter-arrival times in a Poisson process (more on this later when we study the Poisson Process).

The cdf of $Exp(\lambda)$ is easily seen to be

$$F(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x} \qquad \text{for } x > 0.$$

Suppose now that $X_1, \dots, X_n$ are i.i.d observations from $Exp(\lambda)$. What is the density of $X_{(1)}$? From the formula (24):

$$f_{X_{(1)}}(u) = n(1 - F(u))^{n-1} f(u) = (n\lambda)e^{-(n\lambda)u} \qquad \text{for } u > 0.$$

Thus $X_{(1)}$ has the Exponential density with rate parameter $n\lambda$.

15 The Central Limit Theorem

I am using the second chapter of the book *Elements of Large Sample Theory* by Erich Lehmann as the reference for our treatment of the CLT.

The Central Limit Theorem (CLT) is not a single theorem but encompasses a variety of results concerned with the sum of a large number of random variables which, suitably normalized, has a normal limit distribution. The following is the simplest version of the CLT and this is the version that we shall mostly deal with in this class.

Theorem 15.1 (Central Limit Theorem)**.** *Suppose $X_i, i = 1, 2, \dots$ are i.i.d with $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2 < \infty$. Then*

$$\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma}$$

converges in distribution to $N(0,1)$ where $\bar{X}_n = (X_1 + \dots + X_n)/n$.

We will discuss the following points about the CLT:

1. What does "convergence in distribution" mean?
2. How is the CLT proved?
3. Consequences and applications.

Informally, the CLT says that for i.i.d observations $X_1, \dots, X_n$ with finite mean μ and variance σ^2, the quantity $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is approximately (or asymptotically) $N(0,1)$. Informally, the CLT also implies that

1. $\sqrt{n}(\bar{X}_n - \mu)$ is approximately $N(0, \sigma^2)$.
2. $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$.
3. $S_n = X_1 + \dots + X_n$ is approximately $N(n\mu, n\sigma^2)$.
4. $S_n - n\mu$ is approximately $N(0, n\sigma^2)$.
5. $(S_n - n\mu)/(\sqrt{n}\sigma)$ is approximately $N(0,1)$.

It may be helpful here to note that

$$\mathbb{E}(\bar{X}_n) = \mu \quad \text{and} \quad var(\bar{X}_n) = \sigma^2/n$$

and also

$$\mathbb{E}(S_n) = n\mu \quad \text{and} \quad var(S_n) = n\sigma^2.$$

The most remarkable feature of the CLT is that it holds regardless of the distribution of X_i (as long as they are i.i.d from a distribution F that has a finite mean and variance). Therefore the CLT is, in this sense, distribution-free. This makes it possible to derive, using the CLT, statistical procedures which are asymptotically valid without specific distributional assumptions. To illustrate the fact that the distribution of X_i can be arbitrary, let us consider the following examples.

1. **Bernoulli**: Suppose X_i are i.i.d Bernoulli random variables with probability of success given by p. Then $\mathbb{E}X_i = p$ and $var(X_i) = p(1-p)$ so that the CLT implies that $\sqrt{n}(\bar{X}_n - p)/\sqrt{p(1-p)}$ is approximately $N(0,1)$. This is actually called De Moivre's theorem which was proved in 1733 before the general CLT. The general CLT stated above was proved by Laplace in 1810.

   The CLT also implies here that S_n is approximately $N(np, np(1-p))$. We know that S_n is exactly distributed according to the $Bin(n,p)$ distribution. We therefore have the following result: When p is fixed and n is large, the Binomial distribution $Bin(n,p)$ is approximately same as the normal distribution with mean np and variance $np(1-p)$.

2. **Poisson**: Suppose X_i are i.i.d $Poi(\lambda)$ random variables. Then $\mathbb{E}X_i = \lambda = var(X_i)$ so that the CLT says that $S_n = X_1 + \cdots + X_n$ is approximately Normal with mean $n\lambda$ and variance $n\lambda$. It is not hard to show here that S_n is exactly distributed as a $Poi(n\lambda)$ random variable (**prove this!**). We deduce therefore that when n is large and λ is held fixed, $Poi(n\lambda)$ is approximately same as the Normal distribution with mean $n\lambda$ and variance $n\lambda$.

3. **Gamma**: Suppose X_i are i.i.d random variables having the $Gamma(\alpha, \lambda)$ distribution. Check then that $\mathbb{E}X_i = \alpha/\lambda$ and $var(X_i) = \alpha/\lambda^2$. We deduce then, from the CLT, that $S_n = X_1 + \cdots + X_n$ is approximately normally distributed with mean $n\alpha/\lambda$ and variance $n\alpha/\lambda^2$. We derived in the last class that S_n is exactly distributed as $Gamma(n\alpha, \lambda)$. Thus when n is large and α and λ are held fixed, the $Gamma(n\alpha, \lambda)$ is approximately closely by the $N(n\alpha/\lambda, n\alpha/\lambda^2)$ distribution according to the CLT.

4. **Chi-squared**. Suppose X_i are i.i.d chi-squared random variables with 1 degree of freedom i.e., $X_i = Z_i^2$ for i.i.d standard normal random variables $Z_1, Z_2, \ldots$. It is easy to check then that X_i is a $Gamma(1/2, 1/2)$ random variable. This gives that $X_1 + \cdots + X_n$ is exactly $Gamma(n/2, 1/2)$. This exact distribution of $X_1 + \cdots + X_n$ is also called the chi-squared distribution with n degrees of freedom (denoted by χ_n^2). The CLT therefore implies that the χ_n^2 distribution is closely approximated by $N(n, 2n)$.

5. **Cauchy**. Suppose X_i are i.i.d standard Cauchy random variables. Then X_i's do not have finite mean and variance. Thus the CLT does not apply here. In fact, it can be proved here that $(X_1 + \cdots + X_n)/n$ has the Cauchy distribution for every n.

16 Convergence in Distribution

In order to understand the precise meaning of the CLT, we need to understand the notion of *convergence in distribution.*

Definition 16.1 (Convergence in Distribution)**.** *Suppose $Y_1, Y_2, \ldots$ are random variables and F is a cdf. We say that Y_n converges in distribution to F (or that Y_n converges in Law to F) as $n \to \infty$ if*

$$\mathbb{P}\{Y_n \le y\} \to F(y) \qquad \text{as } n \to \infty$$

for every y at which the cdf F is continuous. We denote this by $Y_n \xrightarrow{L} F$.

Put another way, if F_n denotes the cdf of Y_n, then $Y_n \xrightarrow{L} F$ if and only if

$$F_n(y) \to F(y) \qquad \text{as } n \to \infty$$

for every y that is a continuity point of F.

We shall use the following conventions when talking about convergence in distribution.

1. If F is the cdf of a standard distribution such as $N(0,1)$, then we shall take
$$Y_n \xrightarrow{L} N(0,1)$$
to mean that Y_n converges in distribution to the cdf of $N(0,1)$.

2. For a random variable Y, we shall take
$$Y_n \xrightarrow{L} Y$$
to mean that Y_n converges in distribution to the cdf of Y.

Note that convergence in distribution is defined in terms of cdfs which makes it possible to talk about a sequence of discrete random variables converging to a continuous distribution. For example, if Y_n has the discrete uniform distribution on the finite set $\{1/n, 2/n, \dots, 1\}$, then according to the above definition $Y_n \xrightarrow{L} Unif(0,1)$. Note however that Y_n is discrete but $U(0,1)$ is a continuous distribution.

Note that the definition of $Y_n \xrightarrow{L} Y$ only requires that $F_n(y)$ converges to $F(y)$ at every y which is a continuity point of F (here F_n and F are the cdfs of Y_n and Y respectively). If F is a continuous cdf (such as a normal or a uniform cdf), then every point is a continuity point and then $Y_n \xrightarrow{L} F$ is the same as saying that
$$\mathbb{P}\left\{Y_n \leq y\right\} \to F(y) \qquad \text{for every } y.$$
But when F is a discrete cdf, then, for $Y_n \xrightarrow{L} F$, we do not insist on $\mathbb{P}\{Y_n \leq y\}$ converging to $F(y)$ at points y where F is discontinuous. This is advantageous in a situation such as the following. Suppose that $Y_n = 1/n$ for every $n \geq 1$ and $Y = 0$. Then it is easy to see that
$$\mathbb{P}\{Y_n \leq y\} \to \mathbb{P}\{Y \leq y\} \qquad \text{for every } y \neq 0.$$
However, the convergence above does not hold for $y = 0$ as $\mathbb{P}\{Y_n \leq 0\} = 0$ for every n while $\mathbb{P}\{Y \leq 0\} = 1$. Thus if insisted on $\mathbb{P}\{Y_n \leq y\}$ to converge to $\mathbb{P}\{Y \leq y\}$ at all points y (as opposed to only continuity points), then $Y_n = 1/n$ will not converge in distribution to $Y = 0$ (which will be quite unnatural). This is one justification for including the restriction of continuity points of F in the definition of convergence of distribution.

Let us now isolate the following two special cases of $Y_n \xrightarrow{L} Y$.

1. Y **has a continuous cdf** F: In this case, $Y_n \xrightarrow{L} Y$ if $\mathbb{P}\{Y_n \leq y\}$ converges to $\mathbb{P}\{Y \leq y\}$ for every y. This actually implies that
$$\mathbb{P}\{Y_n < y\} \to \mathbb{P}\{Y \leq y\} \qquad \text{for every } y$$
as well as
$$\mathbb{P}\{a \leq Y_n \leq b\} \to \mathbb{P}\{a \leq Y \leq b\} \qquad \text{for every } a \text{ and } b.$$

2. Y **is equal to a constant**. Suppose that the limit random variable Y equals a constant c. The cdf $F(y)$ of Y is then easily seen to be equal to 0 for $y < c$ and 1 for $y > c$. The definition of $\xrightarrow{L}$ then implies that $Y_n \xrightarrow{L} c$ if and only if $\mathbb{P}\{Y_n \leq y\}$ converges to 0 for $y < c$ and converges to 1 for $y > c$. This then is easily seen to be equivalent to :
$$\mathbb{P}\{|Y_n - c| \geq \epsilon\} \to 0 \qquad \text{as } n \to \infty \tag{27}$$
for every $\epsilon > 0$. In this case when Y is a constant, we actually write $Y_n \xrightarrow{L} c$ as $Y_n \xrightarrow{P} c$ and say that Y_n converges in probability to c. Alternatively, you can take (27) as the definition of $Y_n \xrightarrow{P} c$.

The main goal for today is to introduce Moment Generating Functions and use them to prove the Central Limit Theorem. Let us first start by recapping the statement of the CLT.

Theorem 16.2 (Central Limit Theorem)**.** *Suppose $X_i, i = 1, 2, \dots$ are i.i.d with $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2 < \infty$. Then*

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

where $\bar{X}_n = (X_1 + \dots + X_n)/n$.

To understand the statement of the above theorem, we first need to know what the symbol $\xrightarrow{L}$ means. This is the notion of convergence of distribution which is defined as follows. We say that a sequence of random variables $Y_1, Y_2, \dots$ converges in distribution to a cdf F (written as $Y_n \xrightarrow{L} F$) as $n \to \infty$ if $\mathbb{P}\{Y_n \le y\}$ converges as $n \to \infty$ to $F(y)$ for every y at which the function F is continuous. Also, we say that $Y_n \xrightarrow{L} Y$ if $\mathbb{P}\{Y_n \le y\}$ converges (as $n \to \infty$) to $\mathbb{P}\{Y \le y\}$ at every y where the cdf of Y is continuous.

The statement $Y_n \xrightarrow{L} Y$ might suggest that Y_n is close to Y for large n. This is actually not true. $Y_n \xrightarrow{L} Y$ only says that the **distribution** of Y_n is close to that of Y. It is actually more appropriate to write $Y_n \xrightarrow{L} F$ where F is the cdf of Y. For example, suppose that $Y \sim Unif(0,1)$ and let Y_n be equal to Y for odd values of n and equal to $(1 - Y)$ for even values of n. Then, clearly each $Y_n \sim Unif(0,1)$ so that both $Y_n \xrightarrow{L} Y$ as well as $Y_n \xrightarrow{L} 1 - Y$ are true. But obviously Y_n is not close to Y for even n and Y_n is not close to $1 - Y$ for odd n.

When F is a continuous cdf (which is the case when F is, for example, the cdf of $N(0,1)$), the statement $Y_n \xrightarrow{L} F$ is equivalent to

$$\mathbb{P}\{Y_n \le y\} \to F(y) \qquad \text{for every } y.$$

In this case (i.e., when F is continuous), it also follows that

$$\mathbb{P}\{Y < y\} \to F(y) \qquad \text{for every } y$$

and also that

$$\mathbb{P}\{a \le Y_n \le b\} \to \mathbb{P}\{a \le Y \le b\} \qquad \text{for every } a \text{ and } b.$$

As a result, the precise implication of the CLT is that

$$\mathbb{P}\left\{\frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \le y\right\} \to \Phi(y) \qquad \text{as } n \to \infty$$

for every $y \in \mathbb{R}$. Here $\Phi(\cdot)$ is the cdf of $N(0,1)$. Equivalently, the CLT also implies that

$$\mathbb{P}\left\{a \le \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \le b\right\} \to \Phi(b) - \Phi(a) \qquad \text{as } n \to \infty$$

for every $a \le b$. This is same as

$$\mathbb{P}\left\{\bar{X}_n - \frac{b\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n - \frac{a\sigma}{\sqrt{n}}\right\} \to \Phi(b) - \Phi(a) \qquad \text{as } n \to \infty.$$

Suppose now that $z_{\alpha/2} > 0$ is the point on the real line such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ for $0 < \alpha < 1$. Then taking $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$, we deduce that

$$\mathbb{P}\left\{\bar{X}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \le \mu \le \bar{X}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right\} \to \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha \qquad \text{as } n \to \infty.$$

This means that

$$\left[\bar{X}_n - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X}_n + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right]$$

is an asymptotic $100(1 - \alpha)$ % confidence interval for μ (assuming that σ is known). The application of the CLT ensures that no specific distributional assumptions on $X_1, X_2, \dots$ are required for this result.

17 The Weak Law of Large Numbers

The CLT states that $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ converges in distribution to $N(0,1)$ which informally means that $\sqrt{n}(\bar{X}_n - \mu)/\sigma$ is approximately $N(0,1)$. This means that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$. Because the $N(\mu, \sigma^2/n)$ becomes more and more concentrated about the single point μ, it makes sense to conjecture that $\bar{X}_n$ converges to the single point μ as $n \to \infty$. This is made precise in the following result which is called the **Weak Law of Large Numbers**.

Theorem 17.1 (Weak Law of Large Numbers). *Suppose $X_1, X_2, \dots$ are independent and identically distributed random variables. Suppose that $\mathbb{E}|X_i| < \infty$ so that $\mathbb{E}X_i$ is well-defined. Let $\mathbb{E}X_i = \mu$. Then*

$$\bar{X}_n := \frac{X_1 + \dots + X_n}{n} \xrightarrow{P} \mu \qquad \text{as } n \to \infty.$$

Recall from last class that $Y_n \xrightarrow{P} c$ (here c is a constant) means that $\mathbb{P}\{|Y_n - c| > \epsilon\}$ converges to zero as $n \to \infty$ for every $\epsilon > 0$. Equivalently, if F is the cdf of the constant random variable which is always equal to c, then $Y_n \xrightarrow{P} c$ is the same as $Y_n \xrightarrow{L} F$.

The Weak Law of Large Numbers as stated above is non-trivial to prove. However an easy proof can be given if we make the additional assumption that the X_i's have a finite variance. In this case, we can simply use the Chebyshev inequality. Indeed, Chebyshev's inequality gives

$$\mathbb{P}\left\{|\bar{X}_n - \mu| > \epsilon\right\} \le \frac{var(\bar{X}_n)}{\epsilon^2} = \frac{var(X_1)}{n\epsilon^2}$$

which converges to zero as $n \to \infty$ (because of the n in the denominator). Note that this proof does not work when $var(X_1) = \infty$.

18 Moment Generating Functions

We shall next attempt to prove the CLT. Our main tool for the proof is the Moment Generating Function which is introduced now.

The Moment Generating Function (MGF) of a random variable X is defined as the function:

$$M_X(t) := \mathbb{E}\left(e^{tX}\right)$$

for all $t \in \mathbb{R}$ for which $\mathbb{E}(e^{tX}) < \infty$. Note that $M_X(0) = 1$. There exist random variables (such as those that are distributed according to the standard Caucy distribution) for which $M_X(t)$ is infinite for every $t \neq 0$.

Example 18.1 (MGF of Standard Gaussian). *If $X \sim N(0,1)$, then its MGF can be easily computed as follows:*

$$\mathbb{E}(e^{tX}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-x^2/2} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\frac{-(x-t)^2}{2}\right) \exp(t^2/2) dx = e^{t^2/2}.$$

Thus $M_X(t) = e^{t^2/2}$ for all $t \in \mathbb{R}$.

The basic properties of MGFs are summarized below.

1) Factorization for Sums of Independent Random Variables: Suppose $X_1, \dots, X_n$ are independent, then

$$M_{X_1 + \dots + X_n}(t) = M_{X_1}(t) M_{X_2}(t) \dots M_{X_n}(t).$$

This is a consequence of the fact that

$$\mathbb{E}e^{t(X_1 + \dots + X_n)} = \mathbb{E}\left(\prod_{i=1}^{n} e^{tX_i}\right) = \prod_{i=1}^{n} \mathbb{E}e^{tX_i},$$

the last equality being a consequence of independence.

2) Scaling: $M_{a+bX}(t) = e^{at} M_X(bt)$ for all t (a and b are constants here). This is easy to prove.

3) MGFs determine distributions: If two random variables have MGFs that are finite and equal in an open interval containing 0, then they have the same distribution (i.e., same cdf everywhere). An implication of this is that $N(0,1)$ is the same distribution which has MGF equal to $e^{t^2/2}$ for all t.

4) MGFs provide information on moments: For $k \geq 1$, the number $\mathbb{E}(X^k)$ is called the k^{th} moment of the random variable X. Knowledge of the MGF allows one to easily read off the moments of X. Indeed, the power series expansion of the MGF is:

$$M_X(t) = \mathbb{E}e^{tX} = \sum_{k=0}^{\infty} \frac{t^k}{k!} \mathbb{E}(X^k).$$

Therefore the k^{th} moment of X is simply the coefficient of t^k in the power series of expansion of $M_X(t)$ multiplied by $k!$.

Alternatively, one can derive the moments $\mathbb{E}(X^k)$ as derivatives of the MGF at 0. Indeed, it is easy to see that

$$M_X^{(k)}(t) = \frac{d^k}{dt^k} \mathbb{E}(e^{tX}) = \mathbb{E}\left(X^k e^{tX}\right)$$

so that

$$M_X^{(k)}(0) = \mathbb{E}(X^k).$$

In words, $\mathbb{E}(X^k)$ equals the k^{th} derivative of M_X at 0. Therefore

$$M_X'(0) = \mathbb{E}(X) \quad \text{and} \quad M_X''(0) = \mathbb{E}(X^2)$$

and so on.

As an example, we can deduce the moments of the standard normal distribution from the fact that its MGF equals $e^{t^2/2}$. Indeed, because

$$e^{t^2/2} = \sum_{i=0}^{\infty} \frac{t^{2i}}{2^i i!},$$

it immediately follows that the k^{th} moment of $N(0,1)$ equals 0 when k is odd and equals

$$\frac{(2j)!}{2^j j!} \qquad \text{when } k = 2j.$$

The final important property of the MGF is the following.

5) Connection between MGFs and Convergence in Distribution: Suppose $Y, Y_1, Y_2, \dots$ are random variables that have finite MGFs in an open interval containing zero. Suppose that $M_{Y_n}(t)$ converges to $M_Y(t)$ as $n \to \infty$ for every t in that open interval. Then $Y_n \xrightarrow{L} Y$.

19 Proof of the CLT using MGFs

Let us recall the basic setting. We have i.i.d random variables $X_1, X_2, \dots$ which have mean μ and finite variance σ^2.

Let $Y_n := \sqrt{n}(\bar{X}_n - \mu)/\sigma$. We need to show that $Y_n \xrightarrow{L} N(0,1)$. From the discussion on MGFs in the previous section, it is clear that it is enough to show that

$$M_{Y_n}(t) \to e^{t^2/2} \qquad \text{for every } t \in (-\infty, \infty).$$

Note that

$$Y_n = \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i - \mu}{\sigma}.$$

As a result,

$$M_{Y_n}(t) = M_{\sum_i (X_i-\mu)/(\sqrt{n}\sigma)}(t) = M_{\sum_i (X_i-\mu)/\sigma}(tn^{-1/2}) = \prod_{i=1}^{n} M_{(X_i-\mu)/\sigma}(tn^{-1/2}) = \left(M(tn^{-1/2})\right)^n$$

where $M(\cdot)$ is the MGF of $(X_1 - \mu)/\sigma$. We now use Taylor's theorem to expand $M(tn^{-1/2})$ up to a quadratic polynomial around 0.

Let us first quickly recap Taylor's theorem. This says that for a function f and two points x and p in the domain of f, we can write

$$f(x) = f(p) + f'(p)(x-p) + \frac{f''(p)}{2!}(x-p)^2 + \cdots + \frac{f^{(r)}(p)}{r!}(x-p)^r + \frac{f^{(r+1)}(\xi)}{(r+1)!}(x-p)^{r+1}$$

where ξ is some point that lies between x and p. This formula requires that f has $(r+1)$ derivatives in an open interval containing p and x.

Using Taylor's theorem with $r = 1$, $x = tn^{-1/2}$ and $p = 0$, we obtain

$$M(tn^{-1/2}) = M(0) + \frac{t}{\sqrt{n}}M'(0) + \frac{t^2}{2n}M''(s_n)$$

for some s_n that lies between 0 and $tn^{-1/2}$. This implies therefore that $s_n \to 0$ as $n \to \infty$. Note now that $M(0) = 0$ and $M'(0) = \mathbb{E}((X_1 - \mu)/\sigma) = 0$. We therefore deduce that

$$M_{Y_n}(t) = \left(1 + \frac{t^2}{2n}M''(s_n)\right)^n.$$

Note now that

$$M''(s_n) \to M''(0) = \mathbb{E}\left(\frac{X_1-\mu}{\sigma}\right)^2 = 1 \qquad \text{as } n \to \infty.$$

We therefore invoke the following fact:

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a \qquad \text{provided } \lim_{n\to\infty} a_n = a \tag{28}$$

to deduce that

$$M_{Y_n}(t) = \left(1 + \frac{t^2}{2n}M''(s_n)\right)^n \to e^{t^2/2} = M_{N(0,1)}(t).$$

This completes the proof of the CLT assuming the fact (28). It remains to prove (28). There exist many proofs for this. Here is one. Write

$$\left(1 + \frac{a_n}{n}\right)^n = \exp\left(n\log\left(1 + \frac{a_n}{n}\right)\right).$$

Let $\ell(x) := \log(1 + x)$. Taylor's theorem for ℓ for $r = 2$ and $p = 0$ gives

$$\ell(x) = \ell(0) + \ell'(0)x + \ell''(\xi)\frac{x^2}{2} = x - \frac{x^2}{2(1+\xi)^2}$$

for some ξ that lies between 0 and x. Taking $x = a_n/n$, we get

$$\ell(a_n/n) = \log(1 + (a_n/n)) = \frac{a_n}{n} - \frac{a_n^2}{2n^2(1+\xi_n)^2}$$

for some ξ_n that lies between 0 and a_n/n (and hence $\xi_n \to 0$ as $n \to \infty$). As a result,

$$\left(1+\frac{a_n}{n}\right)^n = \exp\left(n \log\left(1+\frac{a_n}{n}\right)\right) = \exp\left(a_n - \frac{a_n^2}{2n(1+\xi_n)^2}\right) \to e^a$$

as $n \to \infty$. This proves (28).

This completes the proof of the CLT. Note that we have tacitly assumed that the moment generating function of $X_1, \dots, X_n$ exists for all t. This is much stronger than the existence of the variance of X_i. This proof does not work if the MGF is not finite. There exist more advanced proofs (for example, which work with Characteristic functions as opposed to MGFs) which work only under the assumption of finite variance. These are beyond the scope of this class.

20 Two Remarks on the CLT

A natural question with respect to the CLT is: why is $N(0,1)$ (or $N(0,\sigma^2)$) arising as the limit for sums of independent random variables (and not some other distribution)?

This can be explained in many ways. I will mention two common explanations below.

1. The CLT computes the limiting distribution of

   $$Y_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sqrt{n}\sigma}.$$

   Consider now the following:

   $$Y_{2n} = \frac{\sum_{i=1}^{2n} (X_i - \mu)}{\sqrt{2n}\sigma} = \frac{\sum_{i=1}^{n} (X_i - \mu)}{\sqrt{2}\sqrt{n}\sigma} + \frac{\sum_{i=n+1}^{2n} (X_i - \mu)}{\sqrt{2}\sqrt{n}\sigma} = \frac{Y_n + Y_n'}{\sqrt{2}}$$

   where Y_n' is an independent copy of Y_n (independent copy means that Y_n' and Y_n are independent and have the same distribution). Thus if $Y_n \xrightarrow{L} Y$ for a random variable Y, then it must hold that

   $$Y \overset{d}{=} \frac{Y + Y'}{\sqrt{2}}$$

   where $\overset{d}{=}$ means "equality in distribution" meaning that Y and $(Y+Y')/\sqrt{2}$ have the same distribution. It is easy to see that if $Y \sim N(0,\tau^2)$, then Y and $(Y+Y')/\sqrt{2}$ have the same distribution. Conversely and remarkably, the $N(0,\tau^2)$ distribution is the only distribution which has this property (harder to prove). This, and the fact that $var(Y_n) = 1$ for all n, implies that $N(0,1)$ is the only possible limiting distribution of Y_n.

2. Another interesting interpretation and explanation for the CLT comes from information theoretic considerations. Note that the random variables Y_n have variance equal to 1 for each n. However, as n increases, more variables X_i are involved in the formula for Y_n. One can say therefore that the "entropy" of Y_n is increasing with n while the variance stays the same at 1. Now there is a way of formalizing this notion of entropy and it is possible to show that the $N(0,1)$ is the distribution that **maximizes** entropy subject to variance being equal to 1. This therefore says that the entropy of Y_n increases with n (as more variables X_i are involved in computing Y_n) and eventually as $n \to \infty$, one gets the maximally entropic distribution, $N(0,1)$, as the limit. There is a way of making these precise.

21 More on the Weak Law and Convergence in Probability

In the last couple of classes,we studied the Weak Law of Large Numbers and the Central Limit Theorem. The Weak Law of Large Numbers is the following:

Theorem 21.1 (Weak Law of Large Numbers)**.** *Suppose $X_1, X_2, \dots$ are independent and identically distributed random variables. Suppose that $\mathbb{E}|X_i| < \infty$ so that $\mathbb{E}X_i$ is well-defined. Let $\mathbb{E}X_i = \mu$. Then*

$$\bar{X}_n := \frac{X_1 + \dots + X_n}{n} \xrightarrow{P} \mu \qquad \textit{as } n \to \infty.$$

Recall, from last lecture, that $\xrightarrow{P}$ is defined as follows: $Y_n \xrightarrow{P} c$ if $\mathbb{P}\{|Y_n - c| > \epsilon\}$ converges to zero as $n \to \infty$ for every $\epsilon > 0$. The following result presents an intuitively obvious simple fact about convergence in probability. However, this result is slightly tricky to prove (you are welcome to try proving this; the result itself is useful for us but not the proof).

Lemma 21.2. *If $X_1, X_2, \dots$ and $Y_1, Y_2, \dots$ are two sequences of random variables satisfying $X_n \xrightarrow{P} c$ and $Y_n \xrightarrow{P} c$ for two constants c and d. Then*

1. $X_n + Y_n \xrightarrow{P} c + d$
2. $X_n - Y_n \xrightarrow{P} c - d$
3. $X_n Y_n \xrightarrow{P} cd$
4. $X_n / Y_n \xrightarrow{P} c/d$ *provided* $d \neq 0$.

Let us now get back to the Weak Law of Large Numbers. Note that (21.1) holds with any distributional assumptions on the random variables $X_1, X_2, \dots$ (only the assumptions of independence and having identical distributions and the existence of the expectations are sufficient). The weak law is easy to prove under the additional assumption that the random variables have finite variances. This proof, which we have already seen in the last class, is based on the Chebyshev inequality which says that

$$\mathbb{P}\left\{\left|\bar{X}_n - \mu\right| > \epsilon\right\} \leq \frac{var(\bar{X}_n)}{\epsilon^2}. \tag{29}$$

Because

$$var(\bar{X}_n) = var\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2} var(X_1 + \dots + X_n) = \frac{1}{n^2} n \times var(X_1) = \frac{\sigma^2}{n} \to 0$$

as $n \to \infty$. As a result, from (29), we have that the left hand side of (29) converges to 0 which means that $\bar{X}_n \xrightarrow{P} \mu$.

It follows more generally that if $Y_1, Y_2, \dots$ is a sequence of random variables for which $\mathbb{E}Y_n$ converges to some parameter θ and for which $var(Y_n)$ converges to zero, then $Y_n \xrightarrow{P} \theta$. This is given in the following result.

Lemma 21.3. *Suppose $Y_1, Y_2, \dots$ is a sequence of random variables such that*

1. $\mathbb{E}Y_n \to \theta$ *as* $n \to \infty$
2. $var(Y_n) \to 0$ *as* $n \to \infty$.

Then $Y_n \xrightarrow{P} \theta$ as $n \to \infty$.

Proof. Write $Y_n = \mathbb{E}Y_n + (Y_n - \mathbb{E}Y_n)$. Chebyshev's inequality (and the fact that $var(Y_n) \to \infty$) gives

$$\mathbb{P}\left\{|Y_n - \mathbb{E}Y_n| > \epsilon\right\} \leq \frac{var(Y_n)}{\epsilon^2} \to 0$$

for every $\epsilon > 0$ so that $Y_n - \mathbb{E}Y_n \xrightarrow{P} 0$. This and $\mathbb{E}Y_n \to \theta$ implies (via the first assertion of Lemma 21.2) that $Y_n = \mathbb{E}Y_n + (Y_n - \mathbb{E}Y_n) \xrightarrow{P} \theta$. □

In mathematical statistics, when $Y_n \xrightarrow{P} \theta$, we say that Y_n is a consistent estimator for θ or simply that Y_n is consistent for θ. The Weak Law of Large Numbers simply says that $\bar{X}_n$ is consistent for $\mathbb{E}(X_1)$. More generally, Lemma 21.3 states that Y_n is consistent for θ if $\mathbb{E}(Y_n) \to 0$ and $var(Y_n) \to 0$. The following examples present two more situations where consistency holds.

Example 21.4. *Suppose $X_1, X_2, \dots$ are i.i.d having the uniform distribution on $(0, \theta)$ for a fixed $\theta > 0$. Then the maximum order statistic $X_{(n)} := \max(X_1, \dots, X_n)$ is a consistent estimator for θ i.e., $X_{(n)} \xrightarrow{P} \theta$. We can see this in two ways. The first way is to use the Result (Lemma 21.3) above and compute the mean and variance of $X_{(n)}$. $X_{(n)}/\theta$ is the largest order statistic from an i.i.d sample of size n from $Unif(0,1)$ and, as we have seen in the last class, $X_{(n)}/\theta$ has the $Beta(n, 1)$ distribution. Therefore, using the mean and variance formulae for the Beta distribution (see wikipedia for these formulae), we have*

$$\mathbb{E}\left(\frac{X_{(n)}}{\theta}\right) = \frac{n}{n+1}$$

and

$$var\left(\frac{X_{(n)}}{\theta}\right) = \frac{n}{(n+1)^2(n+2)}.$$

which gives

$$\mathbb{E}X_{(n)} = \frac{n\theta}{n+1}$$

and

$$var(X_{(n)}) = \frac{n\theta^2}{(n+1)^2(n+2)}.$$

It is clear from these that $\mathbb{E}X_{(n)}$ converges to θ and $var(X_{(n)})$ converges to 0 respectively as $n \to \infty$ which implies (via Lemma 21.3) that $X_{(n)}$ converges in probability to θ.

There is a second (more direct) way to see that $X_{(n)} \xrightarrow{P} \theta$. This involves writing

$$\mathbb{P}\{|X_{(n)} - \theta| \geq \epsilon\} = \mathbb{P}\{X_{(n)} \leq \theta - \epsilon\} = \mathbb{P}\{X_i \leq \theta - \epsilon \text{ for all } i\} = \left(1 - \frac{\epsilon}{\theta}\right)^n$$

which clearly goes to zero as $n \to \infty$ (note that ϵ and θ are fixed). This, by the definition of convergence in probability, shows that $X_{(n)} \xrightarrow{P} \theta$.

Example 21.5. *Suppose $X_1, X_2, \dots$ are i.i.d observations with mean μ and finite variance σ^2. Then*

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$$

is a consistent estimator for σ^2. To see this first note that

$$\tilde{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2$$

converges in probability to σ^2 as $n \to \infty$ by the Weak Law of Large Numbers. This is because $\tilde{\sigma}_n^2$ is the average of i.i.d random variables $Y_i = (X_i - \mu)^2$ for $i = 1, \dots, n$. The Weak Law therefore says that $\tilde{\sigma}_n^2$ converges in probability to $\mathbb{E}Y_1 = \mathbb{E}(X_1 - \mu)^2 = \sigma^2$.

Now to argue that $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$, the idea is simply to relate $\hat{\sigma}_n^2$ to $\tilde{\sigma}_n^2$. This can be done as follows:

$$\begin{aligned}\hat{\sigma}_n^2 &= \frac{1}{n}\sum_{i=1}^n \left(X_i - \mu - \bar{X}_n + \mu\right)^2 \\ &= \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 + (\bar{X}_n - \mu)^2 - 2\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)(\bar{X}_n - \mu) \\ &= \frac{1}{n}\sum_{i=1}^n (X_i-\mu)^2 - \left(\bar{X}_n - \mu\right)^2\end{aligned}$$

The first term on the right hand side above converges to σ^2 by the Weak Law of Large Numbers (note that $\sigma^2 = \mathbb{E}(X_1-\mu)^2$). The second term converges to zero because $\bar{X}_n \xrightarrow{P} \mu$ (and Lemma 21.2). We use Lemma 21.2 again to conclude that $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$. Note that we have not assumed any distributional assumptions on $X_1, X_2, \dots, X_n$ (the only requirement is they have mean zero and variance σ^2).

By the way, we could have also defined $\hat{\sigma}_n$ by

$$\hat{\sigma}_n^2 := \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2$$

with the factor of $1/(n-1)$ as opposed to $1/n$. This will also converge in probability to σ^2 simply because

$$\frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2 = \left(\frac{1}{n}\sum_{i=1}^n \left(X_i - \bar{X}_n\right)^2\right)\left(\frac{n}{n-1}\right).$$

Since the first term above converges in probability to σ^2 and the second term converges to one, the product converges in probability to σ^2 (by Lemma 21.2).

22 Slutsky's Theorem, Continuous Mapping Theorem and Applications

We also looked at the statement and proof of the Central Limit Theorem which is the following.

Theorem 22.1 (Central Limit Theorem)**.** *Suppose $X_i, i = 1, 2, \dots$ are i.i.d with $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2 < \infty$. Then*

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

where $\bar{X}_n = (X_1 + \dots + X_n)/n$.

Here convergence in distribution ($\xrightarrow{L}$) is defined as follows: A sequence $Y_1, Y_2, \dots$ of random variables is said to converge in distribution to F if $\mathbb{P}\{Y_n \le y\}$ converges to $F(y)$ for every y which is a point of continuity of F. Although convergence in distribution is defined in terms of cdfs, the CLT was proved via moment generating functions because cdfs of sums of independent random variables are not so easy to work with.

An important consequence of the CLT from the statistical point of view is that it gives asymptotically valid confidence intervals for μ. Indeed, as a consequence of the CLT, we have

$$\mathbb{P}\left\{a \le \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \le b\right\} \to \Phi(b) - \Phi(a) \qquad \text{as } n \to \infty$$

for every $a \leq b$. This is same as

$$\mathbb{P}\left\{\bar{X} - \frac{b\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - \frac{a\sigma}{\sqrt{n}}\right\} \to \Phi(b) - \Phi(a) \qquad \text{as } n \to \infty.$$

Suppose now that $z_{\alpha/2} > 0$ is the point on the real line such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$ for $0 < \alpha < 1$. Then taking $a = -z_{\alpha/2}$ and $b = z_{\alpha/2}$, we deduce that

$$\mathbb{P}\left\{\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right\} \to \Phi(z_{\alpha/2}) - \Phi(-z_{\alpha/2}) = 1 - \alpha \qquad \text{as } n \to \infty.$$

This means that

$$\left[\bar{X} - \frac{z_{\alpha/2}\sigma}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\sigma}{\sqrt{n}}\right] \tag{30}$$

is an asymptotic $100(1 - \alpha)$ % confidence interval for μ (assuming that σ is known). The application of the CLT ensures that no distributional assumptions on $X_1, X_2, \dots$ are required for this result.

The problem with the interval (30) is that it depends on σ which will be unknown in a statistical setting (the only available data will be $X_1, \dots, X_n$). A natural idea is to replace σ by a natural estimate such as $\hat{\sigma}_n$ defined in Example 21.5:

$$\hat{\sigma}_n^2 := \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 \tag{31}$$

This will result in the interval:

$$\left[\bar{X} - \frac{z_{\alpha/2}\hat{\sigma}_n}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\hat{\sigma}_n}{\sqrt{n}}\right] \tag{32}$$

Slutsky's theorem stated next will imply that

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n} \xrightarrow{L} N(0, 1) \tag{33}$$

which will mean that (32) is also an asymptotic $100(1 - \alpha)$% confidence interval for μ.

Theorem 22.2 (Slutsky's theorem). *If $Y_n \xrightarrow{L} Y$, $A_n \xrightarrow{P} a$ and $B_n \xrightarrow{P} b$, then*

$$A_n + B_n Y_n \xrightarrow{L} a + bY$$

.

Another useful result that we shall often use is the continuous mapping theorem:

Theorem 22.3 (Continuous Mapping Theorem). *1. Suppose $Y_n \xrightarrow{L} Y$ and f is a function that is continuous in the range of values of Y, then $f(Y_n) \xrightarrow{L} f(Y)$.*

2. Suppose $Y_n \xrightarrow{P} c$ and f is continuous at c, then $f(Y_n) \xrightarrow{P} f(c)$.

One immediate application of these two results is (33) as shown below.

Example 22.4. *Let $X_1, \dots, X_n$ be i.i.d observations with mean μ and variance σ^2. We need to look at the limiting distribution of:*

$$T_n := \frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n}. \tag{34}$$

where $\hat{\sigma}_n$ is as defined in (31). *Note that*

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\hat{\sigma}_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\frac{\sigma}{\hat{\sigma}_n} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}\sqrt{\frac{\sigma^2}{\hat{\sigma}_n^2}}.$$

The first term on the right hand side above converges in probability to $N(0,1)$ by the usual CLT. For the second term, note that $\sigma_n^2 \xrightarrow{P} \sigma^2$ (as proved in Example 21.5) and so applying the continuous mapping theorem with $f(x) = \sqrt{\sigma^2/x}$ implies that $f(\hat{\sigma}_n^2) \xrightarrow{P} 1$. This gives that the second term above converges in probabilty to 1. We can thus use Slutsky's theorem to observe that, since the first term above converges to $N(0,1)$ in distribution and the second term converges in probability to 1, the random variable T_n converges in distribution to $N(0,1)$. As a result,

$$\left[\bar{X} - \frac{z_{\alpha/2}\hat{\sigma}_n}{\sqrt{n}}, \bar{X} + \frac{z_{\alpha/2}\hat{\sigma}_n}{\sqrt{n}}\right]$$

is still a $100(1-\alpha)$ % asymptotically valid C.I for μ. Note that we have not assumed any distributional assumptions on $X_1, \dots, X_n$. In particular, the data can be non-Gaussian.

The random variable T_n in (34) is called the sample t-statistic. The name comes from the t-distrbution or t-density. For a given integer $k \geq 1$, the t-density with k degrees of freedom is the density of the random variable

$$\frac{Z}{\sqrt{A/k}}$$

where $Z \sim N(0,1)$, A has the chi-squared density with k degrees (i.e., $A \sim \chi_k^2$) and Z and A are independent random variables.

Now when $X_1, \dots, X_n$ are i.i.d $N(\mu, \sigma^2)$, it can be shown (we will see how to do this later) that

$$\frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} \sim N(0,1) \quad \textit{and} \quad \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi_{n-1}^2$$

and moreover the above two random variables are independent. As a result, the t-statistic T_n has the t-distribution with $n-1$ degrees of freedom when $X_1, \dots, X_n$ are i.i.d $N(\mu, \sigma^2)$.

Therefore

1. *When $X_1, \dots, X_n$ are i.i.d $N(\mu, \sigma^2)$, the sample t-statistic T_n has the t-distribution with $n-1$ degrees of freedom.*
2. *When $X_1, \dots, X_n$ are i.i.d with mean μ and finite variance σ^2 (no distributional assumption), the t-statistic, T_n converges in distribution to $N(0,1)$.*

It may be helpful to note in connection with the above that the t-distribution with k degrees of freedom itself converges in distribution to $N(0,1)$ as $k \to \infty$.

Example 22.5 (Bernoulli Parameter Estimation)**.** *Suppose $X_1, X_2, \dots, X_n$ are i.i.d having the Ber(p) distribution. The CLT then gives*

$$\frac{\sum_i X_i - np}{\sqrt{np(1-p)}} \xrightarrow{L} N(0,1)$$

which gives

$$\mathbb{P}\left\{-z_{\alpha/2} \leq \frac{\sum_i X_i - np}{\sqrt{np(1-p)}} \leq z_{\alpha/2}\right\} \to 1 - \alpha \qquad \textit{as } n \to \infty.$$

This will not directly lead to a C.I for p. To do this, it is natural to replace p in the denominator by $\bar{X}$. This can be done because

$$\frac{\sum_i X_i - np}{\sqrt{n\bar{X}_n(1-\bar{X}_n)}} = \frac{\sum_i X_i - np}{\sqrt{np(1-p)}}\sqrt{\frac{p(1-p)}{\bar{X}_n(1-\bar{X}_n)}}$$

and by Slutsky's theorem, the above random variables converge in distribution to $N(0,1)$. To give more details, we are using the fact that the first random variable above converges in distribution to $N(0,1)$ by the CLT *and the second random variable converges in probabiliity to* 1 *(basically* $\bar{X}_n \xrightarrow{P} \mu$ *and then use the continuous mapping theorem). This allows us to deduce that*

$$\mathbb{P}\left\{-z_{\alpha/2} \leq \frac{\sum_i X_i - np}{\sqrt{n\bar{X}_n(1-\bar{X}_n)}} \leq z_{\alpha/2}\right\} \to 1-\alpha \qquad as\ n \to \infty.$$

so that

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n(1-\bar{X}_n)}{n}}\right]$$

is an asymptotically valid $100(1-\alpha)$ *% C.I for p.*

Example 22.6 (Poisson Mean Estimation)**.** *Suppose* $X_1, X_2, \dots, X_n$ *are i.i.d having the Poi*(λ) *distribution. The CLT then gives*

$$\frac{\sum_i X_i - n\lambda}{\sqrt{n\lambda}} \xrightarrow{L} N(0,1)$$

which gives

$$\mathbb{P}\left\{-z_{\alpha/2} \leq \frac{\sum_i X_i - n\lambda}{\sqrt{n\lambda}} \leq z_{\alpha/2}\right\} \to 1-\alpha \qquad as\ n \to \infty.$$

It is not easy to convert this into a C.I for λ*. This will become much simpler if we can replace the* λ *in the denominator by* $\bar{X}$*. This can be done because*

$$\frac{\sum_i X_i - n\lambda}{\sqrt{n\bar{X}_n}} = \frac{\sum_i X_i - n\lambda}{\sqrt{n\lambda}}\sqrt{\frac{\lambda}{\bar{X}_n}}$$

and by Slutsky's theorem, the above random variables converge in distribution to $N(0,1)$ *(we are using here that* $\bar{X}_n \xrightarrow{P} \lambda$ *which is a consequence of the Weak Law of Large Numbers). This allows us to deduce that*

$$\mathbb{P}\left\{-z_{\alpha/2} \leq \frac{\sum_i X_i - n\lambda}{\sqrt{n\bar{X}_n}} \leq z_{\alpha/2}\right\} \to 1-\alpha \qquad as\ n \to \infty.$$

so that

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n}{n}}\right]$$

is an asymptotically valid $100(1-\alpha)$ *% C.I for* λ*.*

Example 22.7 (Asymptotic Distribution of sample variance)**.** *Let* $X_1, X_2, \dots$ *be i.i.d with mean* μ *and finite variance* σ^2*. Let*

$$\hat{\sigma}_n^2 := \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2.$$

We know that $\hat{\sigma}_n^2 \xrightarrow{P} \sigma^2$*. Can we also find the limiting distribution of* $\sqrt{n}\left(\hat{\sigma}_n^2 - \sigma^2\right)$*?*

To do this, write

$$\begin{aligned}\sqrt{n}\left(\hat{\sigma}^2 - \sigma^2\right) &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2 - \sigma^2\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu - \bar{X} + \mu)^2 - \sigma^2\right) \\ &= \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 - \sigma^2\right) - \sqrt{n}\left(\bar{X}_n - \mu\right)^2.\end{aligned}$$

Now by the CLT,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\sigma^2\right)\xrightarrow{L}N(0,\tau^2)$$

where $\tau^2 = var((X_1-\mu)^2)$ (we are assuming, of course, that $\tau^2<\infty$) and, by Slutsky's theorem,

$$\sqrt{n}(\bar{X}_n-\mu)^2=\left\{\sqrt{n}\left(\bar{X}_n-\mu\right)\right\}\left(\bar{X}_n-\mu\right)\xrightarrow{P}(N(0,1))\cdot(0)=0$$

Thus by Slutsky's theorem again, we obtain

$$\sqrt{n}\left(\hat{\sigma}^2-\sigma^2\right)\xrightarrow{L}N(0,\tau^2).$$

Another easy consequence of Slutsky's theorem is the following.

Fact: If $r_n(T_n-\theta)\xrightarrow{L}Y$ for some rate $r_n\to\infty$ (typically $r_n=\sqrt{n}$). Then $T_n\xrightarrow{P}\theta$.

This immediately follows from Slutsky's theorem because

$$T_n-\theta=\left(\frac{1}{r_n}\right)r_n(T_n-\theta)\xrightarrow{P}0.Y=0$$

as $1/r_n\to 0$ and $r_n(T_n-\theta)\xrightarrow{L}Y$.

Here is a simple consequence of the CLT and the continuous mapping theorem. Suppose $X_1, X_2, \dots$ are i.i.d random variables with mean μ and finite variance σ^2. Then the CLT says that

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

The continuous mapping theorem then gives

$$\frac{n(\bar{X}-\mu)^2}{\sigma^2}\xrightarrow{L}\chi_1^2.$$

23 Delta Method

Delta Method is another general statement about convergence in distribution that has interesting applications when used in conjunction with the CLT.

Theorem 23.1 (Delta Method)**.** *If $\sqrt{n}(T_n-\theta)\xrightarrow{L}N(0,\tau^2)$, then*

$$\sqrt{n}(g(T_n)-g(\theta))\xrightarrow{L}N(0,\tau^2(g'(\theta))^2)$$

provided $g'(\theta)$ exists and is non-zero.

Informally, the Delta method states that if T_n has a limiting Normal distribution, then $g(T_n)$ also has a limiting normal distribution and also gives an explicit formula for the asymptotic variance of $g(T_n)$. This is surprising because g can be linear or non-linear. In general, non-linear functions of normal random variables do not have a normal distribution. But the Delta method works because under the assumption that $\sqrt{n}(T_n-\theta)\xrightarrow{L}N(0,\tau^2)$, it follows that $T_n\xrightarrow{P}\theta$ so that T_n will be close to θ at least for large n. In a neighborhood of θ, the non-linear function g can be approximated by a linear function which means that g effectively behaves like a linear function. Indeed, the Delta method is a consequence of the approximation:

$$g(T_n)-g(\theta)\approx g'(\theta)\left(T_n-\theta\right).$$

Here is an application of the Delta method.

Example 23.2. *Suppose $0 \leq p \leq 1$ is a fixed parameter and suppose that we want to estimate p^2. Let us assume that we have two choices for estimating p^2:*

1. *We can estimate p^2 by X/n where X is the number of successes in n binomial trials with probability p^2 of success.*
2. *We can estimate p^2 by $(Y/n)^2$ where Y is the number of successes in n binomial trials with probability p of success.*

Which of the above is a better estimator of p^2 and why? The Delta method provides a simple answer to this question. Note that, by the CLT, we have

$$\sqrt{n}\left(\frac{X}{n} - p^2\right) \xrightarrow{L} N(0, p^2(1-p^2))$$

and that

$$\sqrt{n}\left(\frac{Y}{n} - p\right) \xrightarrow{L} N(0, p(1-p)).$$

The Delta method can now be used to convert the above limiting statement into an accuracy statement for $(Y/n)^2$ as:

$$\sqrt{n}\left(\left(\frac{Y}{n}\right)^2 - p^2\right) \xrightarrow{L} N(0, 4p(1-p)p^2).$$

We deduce therefore that (X/n) is a better estimator of p^2 compared to $(Y/n)^2$ provided

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

which is equivalent to $p > 1/3$. Thus when $p > 1/3$, X/n is a better estimator of p^2 compared to $(Y/n)^2$ and when $p < 1/3$, $(Y/n)^2$ is the better estimator.

24 Application of the Delta Method to Variance Stabilizing Transformations

24.1 Motivating Variance Stabilizing Transformations

The Delta method can be applied to variance stabilizing transformations. For example, consider the example where we observe data $X_1, X_2, \ldots, X_n$ that are i.i.d having the *Ber*(p) distribution. The CLT then states that

$$\sqrt{n}\left(\bar{X}_n - p\right) \xrightarrow{L} N(0, p(1-p)). \tag{35}$$

It is inconvenient that p also appears in the variance term. This presents an annoyance while finding confidence intervals for p. One way around this problem is to observe that, by Slutsky's theorem,

$$\frac{\sqrt{n}\left(\bar{X}_n - p\right)}{\sqrt{\bar{X}_n(1-\bar{X}_n)}} \xrightarrow{L} N(0, 1).$$

This was done in the last class. While this method is okay, one might still wonder if it is possible to obtain a function f having the property that

$$\sqrt{n}\left(f(\bar{X}_n) - f(p)\right) \xrightarrow{L} N(0, c^2)$$

where the variance c^2 **does not depend on** p. Such a function f would be called a variance stabilizing transformation.

For another example, consider the case where we observe data $X_1, \ldots, X_n$ that are i.i.d having the $Poi(\lambda)$ distribution. The CLT then states that

$$\sqrt{n}\left(\bar{X}_n - \lambda\right) \xrightarrow{L} N(0, \lambda). \tag{36}$$

The fact that λ appears in the variance term above presents an annoyance while finding confidence intervals for λ. As done in last class, we can get around this by observing (via Slutsky's theorem) that

$$\frac{\sqrt{n}(\bar{X}_n - \lambda)}{\sqrt{\bar{X}_n}} \xrightarrow{L} N(0, 1).$$

While this method is okay, one might still wonder if it is possible to obtain a function f having the property that

$$\sqrt{n}\left(f(\bar{X}_n) - f(\lambda)\right) \xrightarrow{L} N(0, c^2)$$

where the variance c^2 does not depend on λ. If one could indeed find such an f, it will be referred to as a variance stabilizing transformation.

24.2 Construction of the Variance Stabilizing Transformation

More generally, given the result:

$$\sqrt{n}\left(T_n - \theta\right) \xrightarrow{L} N(0, \tau^2(\theta)) \tag{37}$$

where the variance $\tau^2(\theta)$ depends on θ, is it possible to find a transformation f for which

$$\sqrt{n}\left(f(T_n) - f(\theta)\right) \xrightarrow{L} N(0, c^2) \tag{38}$$

where the variance c^2 does not depend on θ. We would then say that the function f is a *variance stabilizing transformation.*

This is possible to do via the Delta method. Indeed, Delta method states that

$$\sqrt{n}\left(f(T_n) - f(\theta)\right) \xrightarrow{L} N(0, (f'(\theta))^2 \tau^2(\theta))$$

and so, in order to guarantee (38), we only have to choose f so that

$$f'(\theta) = \frac{c}{\tau(\theta)} \tag{39}$$

which means that $f(\theta) = \int \frac{c}{\tau(\theta)} d\theta$ (indefinite integral).

24.3 Back to the Bernoulli Example

Here we have $X_1, \ldots, X_n$ which are i.i.d having the $Ber(p)$ distribution so that by CLT

$$\sqrt{n}(\bar{X}_n - p) \xrightarrow{L} N(0, p(1-p)).$$

Therefore (37) holds with $T_n = \bar{X}_n$, $\theta = p$ and $\tau^2(\theta) = \theta(1-\theta)$. The formula (38) says therefore that we choose f as

$$f'(\theta) = \frac{c}{\sqrt{\theta(1-\theta)}}$$

which means that $f(\theta) = 2c \arcsin(\sqrt{\theta})$. The Delta method then guarantees that

$$2\sqrt{n}\left(\arcsin(\sqrt{\bar{X}_n}) - \arcsin(\sqrt{p})\right) \xrightarrow{L} N(0, 1).$$

This implies that

$$\mathbb{P}\left\{\left|\arcsin(\sqrt{\bar{X}_n}) - \arcsin(\sqrt{p})\right| \leq \frac{z_{\alpha/2}}{2\sqrt{n}}\right\} \to 1-\alpha \qquad \text{as } n \to \infty$$

so that

$$\left[\sin\left(\arcsin(\sqrt{\bar{X}_n}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right), \sin\left(\arcsin(\sqrt{\bar{X}_n}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ C.I for $\sqrt{p}$. The lower end point of the above interval can be negative (note that $\arcsin(\sqrt{\bar{X}_n})$ takes values between 0 and $\pi/2$ but $\arcsin(\sqrt{\bar{X}_n}) - z_{\alpha/2}/(2\sqrt{n})$ can be negative) while $\sqrt{p}$ is always positive. So we can replace the lower end point by 0 if it turns out to be negative. Using the notation $x_+ = \max(x, 0)$, we see that

$$\left[\left(\sin\left(\arcsin(\sqrt{\bar{X}_n}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right)_+, \sin\left(\arcsin(\sqrt{\bar{X}_n}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ C.I for $\sqrt{p}$. To get a confidence interval for p, we can simply square the two end points of the above interval. This allows us to deduce that

$$\left[\sin^2\left(\arcsin(\sqrt{\bar{X}_n}) - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)_+, \sin^2\left(\arcsin(\sqrt{\bar{X}_n}) + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)\right]$$

is an approximate $100(1-\alpha)\%$ C.I for p.

24.4 Back to the Poisson Example

Let us now get back to the Poisson distribution where we have $X_1, \dots, X_n$ are i.i.d $Poi(\lambda)$ and CLT gives (36). Therefore $T_n = \bar{X}_n$, $\theta = \lambda$ and $\tau^2(\theta) = \theta$. The equation (39) suggests that we choose f as

$$f'(\theta) = \frac{c}{\sqrt{\theta}}$$

where means that $f(\theta) = 2c\sqrt{\theta}$. The Delta method then guarantees that

$$2\sqrt{n}\left(\sqrt{\bar{X}_n} - \sqrt{\lambda}\right) \xrightarrow{L} N(0,1). \tag{40}$$

Therefore the square-root transformation applied to $\bar{X}_n$ ensures that the resulting variance (of $\sqrt{\bar{X}_n}$) does not depend on λ (in a limiting sense).

The fact (40) will lead to approximate confidence intervals for λ. Indeed, (40) immediately implies that

$$\mathbb{P}\left\{\left|\sqrt{\bar{X}_n} - \sqrt{\lambda}\right| \leq \frac{z_{\alpha/2}}{2\sqrt{n}}\right\} \to 1-\alpha \qquad \text{as } n \to \infty$$

so that

$$\left[\sqrt{\bar{X}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}}, \sqrt{\bar{X}_n} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right]$$

is an approximate $100(1-\alpha)$ % C.I for $\sqrt{\lambda}$. Note that the lower end point of the above interval can be negative while λ is always positive. So we can replace the lower end point by 0 if it turns out to be negative. Again using the notation $x_+ := \max(x, 0)$, we see that

$$\left[\left(\sqrt{\bar{X}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)_+, \sqrt{\bar{X}_n} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right]$$

is an approximate $100(1-\alpha)$ % C.I for $\sqrt{\lambda}$. To get a confidence interval for λ, we can simply square the two end points of the above interval. This allows us to deduce that

$$\left[\left(\sqrt{\bar{X}_n} - \frac{z_{\alpha/2}}{2\sqrt{n}}\right)_+^2, \left(\sqrt{\bar{X}_n} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right)^2\right] \tag{41}$$

is an approximate $100(1-\alpha)$ % C.I for λ.

This interval can be compared with the interval that was obtained in the previous lecture using Slutsky's theorem. That interval was

$$\left[\bar{X}_n - z_{\alpha/2}\sqrt{\frac{\bar{X}_n}{n}}, \bar{X}_n + z_{\alpha/2}\sqrt{\frac{\bar{X}_n}{n}}\right]. \tag{42}$$

The intervals (41) and (42) may look different but they are actually quite close to each other for large n. To see this, note that the difference between the upper bounds of these two intervals is at most $z_{\alpha/2}^2/(4n)$ which is very small when n is large (the same is true of the lower bounds).

24.5 Chi-squared Example

Let us now look at another example where the variance stabilizing transformation is the log function.

Suppose $X_1, X_2, \dots$ are i.i.d such that X_i/σ^2 has the chi-squared distribution with one degree of freedom. In other words,

$$X_i \sim \sigma^2 \chi_1^2.$$

Because $\mathbb{E}(X_1) = \sigma^2$ and $var(X_1) = 2\sigma^4$, the CLT says that

$$\sqrt{n}\left(\bar{X}_n - \sigma^2\right) \xrightarrow{L} N(0, 2\sigma^4). \tag{43}$$

Can we now find a function f such that $f(\bar{X}_n)$ has a limiting variance that is independent of σ^2? Because (43) has the form (37) with $T_n = \bar{X}_n$, $\theta = \sigma^2$ and $\tau^2(\theta) = 2\theta^2$, we can use (39) which suggests taking f so that $f'(\theta) = c/\tau(\theta) = c/(\sqrt{2}\theta)$. This gives

$$f(\theta) = \frac{c}{\sqrt{2}}(\log\theta)$$

allowing us to conclude that

$$\sqrt{n}\left(\frac{1}{\sqrt{2}}\log\left(\bar{X}_n\right) - \frac{1}{\sqrt{2}}\log(\sigma^2)\right) \xrightarrow{L} N(0,1).$$

Square-roots and logarithms are common transformations that are applied to data when there is varying variance (see, for example, `https://en.wikipedia.org/wiki/Variance-stabilizing_transformation`).

24.6 Geometric Example

Suppose $X_1, X_2, \dots$ are i.i.d having the Geometric distribution with parameter p. Recall that X has the *Geo*(p) distribution if X takes the values $1, 2, \dots$ with the probabilities

$$\mathbb{P}\{X = k\} = (1-p)^{k-1}p \qquad \text{for } k = 1, 2, \dots.$$

The number of independent tosses (of a coin with probability of heads p) required to get the first head has the *Geo*(p) distribution.

I leave as an easy exercise to verify that for $X \sim Geo(p)$

$$\mathbb{E}X = \frac{1}{p} \quad \text{and} \quad var(X) = \frac{1-p}{p^2}.$$

The CLT therefore states that for i.i.d observations $X_1, X_2, \dots$ having the $Geo(p)$ distribution, we have

$$\sqrt{n}\left(\bar{X}_n - (1/p)\right) \xrightarrow{L} N(0, \frac{1-p}{p^2}).$$

What is the variance stabilizing transformation for $\bar{X}_n$ i.e., what is the transformation f for which $f(\bar{X}_n)$ has constant asymptotic variance? To answer this, note that the above displayed equation is of the same form as (37) with $T_n = \bar{X}_n$, $\theta = 1/p$ and $\tau^2(\theta) = (1-p)/p^2$. We then write $\tau(\theta)$ in terms of θ as (note that $p = 1/\theta$)

$$\tau(\theta) = \sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-(1/\theta)}{1/\theta^2}} = \sqrt{\theta(\theta-1)}.$$

The variance stabilizing transformation is therefore given by

$$f(\theta) = \int \frac{c}{\tau(\theta)} d\theta = \int \frac{c}{\sqrt{\theta(\theta-1)}} d\theta = 2c \log\left(\sqrt{\theta} + \sqrt{\theta - 1}\right)$$

Therfore $f(\theta) = 2c\log(\sqrt{\theta} + \sqrt{\theta-1})$ is the variance stabilizing transformation here and

$$\sqrt{n}\left(f(\bar{X}_n) - f(1/p)\right) \xrightarrow{L} N(0, c^2).$$

25 Delta Method when $g'(\theta) = 0$

Suppose that $\sqrt{n}(T_n - \theta) \xrightarrow{L} N(0, \tau^2)$. We are now interested in the asymptotic distribution of $g(T_n)$. The Delta method stated that when $g'(\theta) \neq 0$, we have

$$\sqrt{n}\left(g(T_n) - g(\theta)\right) \xrightarrow{L} N(0, \tau^2(g'(\theta))^2). \tag{44}$$

This is essentially a consequence of the Taylor approximation: $g(T_n)g(\theta) \approx g'(\theta)(T_n - \theta)$. What would happen if $g'(\theta) = 0$? In this case, the statement (44) will still be correct if the right hand side is interpreted as the constant 0 i.e., when $g'(\theta) = 0$, the following holds:

$$\sqrt{n}\left(g(T_n) - g(\theta)\right) \xrightarrow{P} 0.$$

However this only states that $g(T_n) - g(\theta)$ is of a smaller order compared to $n^{-1/2}$ but does not precisely say what the exact order is and what the limiting distribution is when scaled by the correct order. To figure out these, we need to consider the higher order terms in the Taylor expansion for $g(T_n)$ around θ. Assume, in the sequel, that $g'(\theta) = 0$ and that $g''(\theta) \neq 0$.

In this case, we do a two term Taylor approximation:

$$g(T_n) - g(\theta) \approx g'(\theta)(T_n - \theta) + \frac{1}{2}g''(\theta)(T_n - \theta)^2 = \frac{1}{2}g''(\theta)(T_n - \theta)^2$$

As a result, we have

$$n(g(T_n) - g(\theta)) \approx \frac{1}{2}g''(\theta)n(T_n - \theta)^2.$$

Now by the continuous mapping theorem:

$$n(T_n - \theta)^2 \xrightarrow{L} \tau^2\chi_1^2$$

and hence we have

$$n(g(T_n) - g(\theta)) \xrightarrow{L} \frac{1}{2} g''(\theta)\tau^2\chi_1^2. \tag{45}$$

Therefore when $g'(\theta) = 0$ and $g''(\theta) \neq 0$, the right scaling factor is n and the limiting distribution is a scaled multiple of χ_1^2 (note that the limit is not a normal distribution).

Example 25.1. *Suppose $X_1, X_2, \dots, X_n$ are i.i.d Ber(p) random variables. Suppose we want to estimate $p(1-p)$. The natural estimator is $\bar{X}_n(1-\bar{X}_n)$. What is the limiting behavior of this estimator?*

This can be answered by the Delta method by taking $g(\theta) = \theta(1-\theta)$. Note first that by the usual CLT,

$$\sqrt{n}\left(\bar{X}_n - p\right) \xrightarrow{L} N(0, p(1-p)).$$

For $g(\theta) = \theta(1-\theta)$, note that

$$g'(\theta) = 1 - 2\theta$$

so that $g'(p) \neq 0$ when $p \neq 1/2$. Thus when $p \neq 1/2$, the Delta method gives

$$\sqrt{n}\left(g(\bar{X}_n) - g(p)\right) \xrightarrow{L} N(0, \tau^2(g'(\theta))^2) = N(0, p(1-p)(1-2p)^2).$$

But when $p = 1/2$, we have to use (45) instead of (44) and this gives (note that $g''(p) = -2$)

$$n\left(g(\bar{X}_n) - g(1/2)\right) \xrightarrow{L} \frac{1}{2}(-2)p(1-p)\chi_1^2 = -\frac{1}{4}\chi_1^2.$$

26 Conditioning

Our next main topic is conditioning. This is a very important topic for statistics classes.

26.1 Basics

Let us first start by looking at the definition of conditional probability. Given two events A and B with $\mathbb{P}(A) > 0$, we define the conditional probability of B given A as

$$\mathbb{P}\left(B|A\right) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{46}$$

See Section 1.1 of Lecture 10 of Jim Pitman's 2016 notes for 201A to get some intuitive justification for this definition of conditional probability.

Using this definition of conditional probability, we can see that

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

Note here that A and A^c are disjoint events whose union is the entire space of outcomes Ω. More generally, if $A_1, A_2, \dots$ are disjoint events whose union is Ω, we have

$$\mathbb{P}(B) = \sum_{i \geq 1} \mathbb{P}(B|A_i)\mathbb{P}(A_i). \tag{47}$$

This is referred to as the **Law of total probability**.

Let us now come to **Bayes rule** which states that

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}.$$

26.2 Conditional Distributions, Law of Total Probability and Bayes Rule for Discrete Random Variables

Consider two random variables X and Θ. Assume that both are discrete random variables. One can then define the conditional distribution of X given $\Theta = \theta$ simply by defining the conditional probabilities:

$$\mathbb{P}\{X = x|\Theta = \theta\} = \frac{\mathbb{P}\{X = x, \Theta = \theta\}}{\mathbb{P}\{\Theta = \theta\}} \tag{48}$$

assuming that $\mathbb{P}\{\Theta = \theta\} > 0$. If $\mathbb{P}\{\Theta = \theta\} = 0$, we would not attempt to define $\mathbb{P}\{X = x|\Theta = \theta\}$.

As x varies over all values that the random variable X takes, the probabilities (51) determine the conditional distribution of X given $\Theta = \theta$. Note that the conditional probability $\mathbb{P}\{X = x|\Theta = \theta\}$ always lies between 0 and 1 and we have $\sum_x \mathbb{P}\{X = x|\Theta = \theta\} = 1$.

Example 26.1. *Suppose X and Y are independent random variables having the $Poi(\lambda_1)$ and $Poi(\lambda_2)$ distributions respectively. For $n \geq 0$, what is the conditional distribution of X given $X + Y = n$?*

We need to compute

$$\mathbb{P}\left\{X = i|X + Y = n\right\}$$

for various values of i. It is clear that the above probability is non-zero only when i is an integer between 0 and n. Let us therefore assume that i is an integer between 0 and n. By definition

$$\begin{aligned}\mathbb{P}\left\{X = i|X + Y = n\right\} &= \frac{\mathbb{P}\left\{X = i, X + Y = n\right\}}{P\{X + Y = n\}} \\ &= \frac{\mathbb{P}\left\{X = i, Y = n - i\right\}}{P\{X + Y = n\}} \\ &= \frac{\mathbb{P}\left\{X = i\right\}\mathbb{P}\left\{Y = n - i\right\}}{P\{X + Y = n\}}\end{aligned}$$

The numerator above can be evaluated directly as X and Y are independently distributed as $Poi(\lambda_1)$ and $Poi(\lambda_2)$ respectively. For the denominator, we use the fact that $X + Y$ is $Poi(\lambda_1 + \lambda_2)$ (the proof of this fact is left as exercise). We thus have

$$\begin{aligned}\mathbb{P}\left\{X = i|X + Y = n\right\} &= \frac{\mathbb{P}\left\{X = i\right\}\mathbb{P}\left\{Y = n - i\right\}}{P\{X + Y = n\}} \\ &= \frac{e^{-\lambda_1}\left(\lambda_1^i/i!\right)e^{-\lambda_2}\left(\lambda_2^{n-i}/(n-i)!\right)}{e^{-\lambda_1+\lambda_2}((\lambda_1 + \lambda_2)^n/n!)} \\ &= \frac{n!}{i!(n-i)!}\left(\frac{\lambda_1}{\lambda_1 + \lambda_2}\right)^i\left(\frac{\lambda_2}{\lambda_1 + \lambda_2}\right)^{n-i}\end{aligned}$$

which means that the conditional distribution of X given $X + Y = n$ is the Binomial distribution with parameters n and $p = \lambda_1/(\lambda_1 + \lambda_2)$.

Let us now look at the law of total probability and Bayes rule for discrete random variables X and Θ. As a consequence of (47), we have

$$\mathbb{P}\{X = x\} = \sum_\theta \mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\} \tag{49}$$

where the summation is over all values of θ that are taken by the random variable Θ. This formula allows one to calculate $\mathbb{P}\{X = x\}$ using knowledge of $\mathbb{P}\{X = x|\Theta = \theta\}$ and $\mathbb{P}\{\Theta = \theta\}$. We shall refer to (52) as the **Law of Total Probability** for discrete random variables.

The Bayes rule is

$$\mathbb{P}\{\Theta = \theta | X = x\} = \frac{\mathbb{P}\{X = x | \Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\mathbb{P}\{X = x\}} = \frac{\mathbb{P}\{X = x | \Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\sum_\theta \mathbb{P}\{X = x | \Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}. \tag{50}$$

The Bayes rule allows one to compute the conditional probabilities of Θ given X using knowledge of the conditional probabilities of X given Θ as well as the marginal probabilities of Θ. We shall refer to (53) as the **Bayes Rule** for discrete random variables.

Example 26.2. *Suppose N is a random variable having the $Poi(\lambda)$ distribution. Also suppose that, conditional on $N = n$, the random variable X has the $Bin(n, p)$ distribution. This setting is known as the **Poissonization of the Binomial**. Find the marginal distribution of X. Also what is the conditional distribution of N given $X = i$?*

To find the marginal distribution of X, we need to find $\mathbb{P}\{X = i\}$ for every integer $i \geq 0$. For this, we use the law of total probability which states that

$$\mathbb{P}\{X = i\} = \sum_{n=0}^{\infty} \mathbb{P}\{X = i | N = n\} \mathbb{P}\{N = n\}.$$

Because $X|N = n$ is $Bin(n, p)$, the probability $\mathbb{P}\{X = i | N = n\}$ is non-zero only when $0 \leq i \leq n$. Therefore the terms in the sum above are non-zero only when $n \geq i$ and we obtain

$$\begin{aligned}
\mathbb{P}\{X = i\} &= \sum_{n=i}^{\infty} \mathbb{P}\{X = i | N = n\} \mathbb{P}\{N = n\} \\
&= \sum_{n=i}^{\infty} \binom{n}{i} p^i (1-p)^{n-i} e^{-\lambda} \frac{\lambda^n}{n!} \\
&= \frac{e^{-\lambda} p^i}{i!} \sum_{n=i}^{\infty} \frac{(1-p)^{n-i}}{(n-i)!} \lambda^n \\
&= \frac{e^{-\lambda} (\lambda p)^i}{i!} \sum_{n=i}^{\infty} \frac{(1-p)^{n-i}}{(n-i)!} \lambda^{n-i} \\
&= \frac{e^{-\lambda} (\lambda p)^i}{i!} \sum_{n=i}^{\infty} \frac{(\lambda(1-p))^{n-i}}{(n-i)!} = \frac{e^{-\lambda} (\lambda p)^i}{i!} e^{\lambda(1-p)} = \frac{e^{-\lambda p} (\lambda p)^i}{i!}
\end{aligned}$$

This means that X has the $Poi(\lambda p)$ distribution.

To find the conditional distribution of Θ given $X = i$, we need to use Bayes rule which states that

$$\mathbb{P}\{N = n | X = i\} = \frac{\mathbb{P}\{X = i | N = n\}\mathbb{P}\{N = n\}}{\mathbb{P}\{X = i\}}$$

This is only nonzero when $n \geq i$ (otherwise $\mathbb{P}\{X = i | N = n\}$ will be zero). And when $n \geq i$, we have

$$\begin{aligned}
\mathbb{P}\{N = n | X = i\} &= \frac{\mathbb{P}\{X = i | N = n\}\mathbb{P}\{N = n\}}{\mathbb{P}\{X = i\}} \\
&= \frac{\binom{n}{i} p^i (1-p)^{n-i} e^{-\lambda} (\lambda^n / n!)}{e^{-\lambda p} ((\lambda p)^i / i!)} \\
&= e^{-\lambda(1-p)} \frac{[\lambda(1-p)]^{n-i}}{(n-i)!} \qquad \textit{for } n \geq i.
\end{aligned}$$

This means that conditional on $X = i$, the random variable N is distributed as $i + Poi(\lambda(1-p))$.

What is the joint distribution of X and $N-X$ in this example? To compute this, note that

$$\begin{aligned}\mathbb{P}\{X=i, N-X=j\} &= \mathbb{P}\{X=i, N=i+j\} \\ &= \mathbb{P}\{X=i|N=i+j\}\,\mathbb{P}\{N=i+j\} \\ &= \binom{i+j}{i} p^i(1-p)^j e^{-\lambda}\frac{\lambda^{i+j}}{(i+j)!} \\ &= e^{-\lambda}\frac{(\lambda p)^i}{i!}\frac{(\lambda(1-p))^j}{j!}\end{aligned}$$

Note that this factorizes into a term involving only i and a term involving only j. This means therefore that X and $N-X$ are independent. Also from the expression above, it is easy to deduce that the marginal distribution of X is $Poi(\lambda p)$ (which we have already derived via the law of total probability) and that $N-X$ is $Poi(\lambda(1-p))$.

The setting of this example arises when one tosses a coin with probability of heads p independently a $Poi(\lambda)$ number of times. Then N denotes the total number of tosses, X denotes the number of heads and $N-X$ denotes the number of tails. We have thus shown that X and $N-X$ are independent and are distributed according to $Poi(\lambda p)$ and $Poi(\lambda(1-p))$ respectively. Independence of X and $N-X$ here is especially interesting. When a coin is tossed a fixed number n of times, the number of heads and tails are obviously not independent (as they have to sum to n). But when the number of tosses is itself random and has the Poisson distribution, then the number of heads and tails become indepedent random variables.

27 Conditioning for Discrete Random Variables

In the last lecture, we studied conditioning for discrete random variables. Given two discrete random variables X and Θ, one can then define the conditional distribution of X given $\Theta=\theta$ simply by defining the conditional probabilities:

$$\mathbb{P}\{X=x|\Theta=\theta\} = \frac{\mathbb{P}\{X=x,\Theta=\theta\}}{\mathbb{P}\{\Theta=\theta\}} \tag{51}$$

assuming that $\mathbb{P}\{\Theta=\theta\}>0$. If $\mathbb{P}\{\Theta=\theta\}=0$, we would not attempt to define $\mathbb{P}\{X=x|\Theta=\theta\}$.

As x varies over all values that the random variable X takes, the probabilities (51) determine the conditional distribution of X given $\Theta=\theta$. Note that the conditional probability $\mathbb{P}\{X=x|\Theta=\theta\}$ always lies between 0 and 1 and we have $\sum_x \mathbb{P}\{X=x|\Theta=\theta\}=1$.

We also looked at the law of total probability and the Bayes rule. The Law of Total Probability is

$$\mathbb{P}\{X=x\} = \sum_\theta \mathbb{P}\{X=x|\Theta=\theta\}\mathbb{P}\{\Theta=\theta\} \tag{52}$$

where the summation is over all values of θ that are taken by the random variable Θ. This formula allows one to calculate $\mathbb{P}\{X=x\}$ using knowledge of $\mathbb{P}\{X=x|\Theta=\theta\}$ and $\mathbb{P}\{\Theta=\theta\}$.

The Bayes rule is

$$\mathbb{P}\{\Theta=\theta|X=x\} = \frac{\mathbb{P}\{X=x|\Theta=\theta\}\mathbb{P}\{\Theta=\theta\}}{\mathbb{P}\{X=x\}} = \frac{\mathbb{P}\{X=x|\Theta=\theta\}\mathbb{P}\{\Theta=\theta\}}{\sum_\theta \mathbb{P}\{X=x|\Theta=\theta\}\mathbb{P}\{\Theta=\theta\}}. \tag{53}$$

The Bayes rule allows one to compute the conditional probabilities of Θ given X using knowledge of the conditional probabilities of X given Θ as well as the marginal probabilities of Θ.

The goal of this lecture is to extend all of the above to the case when X and Θ are continuous random variables.

28 Conditional Densities for Continuous Random Variables

Consider now two continuous random variables X and Θ having a joint density $f_{X,\Theta}(x, \theta)$. Recall then that $f_{X,\Theta}(x, \theta) \geq 0$ for all x, θ and $\int \int f(x, \theta) dx d\theta = 1$. Also recall that the marginal densities of X and Θ are given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,\Theta}(x, \theta) d\theta \quad \text{and} \quad f_\Theta(\theta) = \int_{-\infty}^{\infty} f_{X,\Theta}(x, \theta) dx.$$

We shall now define the conditional density of X given $\Theta = \theta$ for a fixed value of θ. In order to define this conditional density at a point x, we need to consider

$$\mathbb{P}\{x \leq X \leq x + \delta | \Theta = \theta\} \tag{54}$$

for a small $\delta > 0$. Because $\mathbb{P}\{\Theta = \theta\} = 0$ (note that Θ is a continuous random variable), we cannot define this conditional probability using the definition $\mathbb{P}(B|A) := \mathbb{P}(B \cap A)/\mathbb{P}(A)$. But, intuitively, conditioning on $\Theta = \theta$ should be equivalent to conditioning on $\theta \leq \Theta \leq \theta + \epsilon$ for small ϵ. Therefore we can write

$$\mathbb{P}\{x \leq X \leq x + \delta | \Theta = \theta\} \approx \mathbb{P}\{x \leq X \leq x + \delta | \theta \leq \Theta \leq \theta + \epsilon\}$$

for small ϵ. For the probability on the right hand side above, we can use $\mathbb{P}(B|A) := \mathbb{P}(B \cap A)/\mathbb{P}(A)$ to obtain

$$\mathbb{P}\{x \leq X \leq x + \delta | \theta \leq \Theta \leq \theta + \epsilon\} = \frac{\mathbb{P}\{x \leq X \leq x + \delta, \theta \leq \Theta \leq \theta + \epsilon\}}{\mathbb{P}\{\theta \leq \Theta \leq \theta + \epsilon\}} \approx \frac{f_{X,\Theta}(x, \theta)\delta\epsilon}{f_\Theta(\theta)\epsilon} = \frac{f_{X,\Theta}(x, \theta)\delta}{f_\Theta(\theta)}$$

We have thus obtained that

$$\mathbb{P}\{x \leq X \leq x + \delta | \Theta = \theta\} \approx \frac{f_{X,\Theta}(x, \theta)\delta}{f_\Theta(\theta)}$$

for small δ. This suggests the definition

$$f_{X|\Theta=\theta}(x) := \frac{f_{X,\Theta}(x, \theta)}{f_\Theta(\theta)} \tag{55}$$

for the conditional density of X given $\Theta = \theta$. This definition makes sense as long as $f_\Theta(\theta) > 0$. If $f_\Theta(\theta) = 0$, we do not attempt to define $f_{X|\Theta=\theta}$.

Example 28.1. *Suppose X and Θ are independent random variables having the Gamma(α, λ) and Gamma(β, λ) distributions respectively. What then is the conditional density of X given $X + \Theta = 1$.*

The definition (55) *gives*

$$f_{X|X+\Theta=1}(x) = \frac{f_{X,X+\Theta}(x, 1)}{f_{X+\Theta}(1)}.$$

By the Jacobian formula for calculating densities of transformed random variables, it can be checked that

$$f_{X,X+\Theta}(x, 1) = f_{X,\Theta}(x, 1 - x) = f_X(x) f_\Theta(1 - x) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} e^{-\lambda}$$

for $0 < x < 1$. We have also seen previously that $X + \Theta$ is distributed as $\Gamma(\alpha + \beta, \lambda)$. Thus

$$f_{X+\Theta}(1) = \frac{\lambda^{\alpha+\beta}}{\Gamma(\alpha + \beta)} e^{-\lambda}.$$

Therefore

$$f_{X|X+\Theta=1}(x) = \frac{f_{X,X+\Theta}(x, 1)}{f_{X+\Theta}(1)} = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1 - x)^{\beta-1} \qquad \textit{for } 0 < x < 1.$$

This means therefore that

$$X|(X + \Theta = 1) \sim Beta(\alpha, \beta).$$

Example 28.2. *Suppose X and Y are independent $Unif(0,1)$ random variables. What is $f_{U|V=v}$ where $U = \min(X,Y)$ and $V = \max(X,Y)$ and $0 < v < 1$?*

Note first that

$$f_{U|V=v}(u) = \frac{f_{U,V}(u,v)}{f_V(v)}.$$

When $0 < u < v < 1$, we know that

$$f_{U,V}(u,v) = f_{X,Y}(u,v) + f_{X,Y}(v,u) = 2.$$

Also $V = \max(X,Y) \sim Beta(2,1)$ so that

$$f_V(v) = 2vI\{0 < v < 1\}.$$

We thus have

$$f_{U|V=v}(u) = \frac{2}{2v} = \frac{1}{v} \qquad \text{for } 0 < u < v.$$

In other words, $U|V = v$ is uniformly distributed on the interval $(0,v)$.

29 Conditional Density is Proportional to Joint Density

The conditional density

$$f_{X|\Theta=\theta}(x) := \frac{f_{X,\Theta}(x,\theta)}{f_\Theta(\theta)} \tag{56}$$

has the following important property. As a function of x (and keeping θ fixed), $f_{X|\Theta=\theta}(x)$ is a valid density i.e.,

$$f_{X|\Theta=\theta}(x) \geq 0 \text{ for every } x \quad \text{and} \quad \int_{-\infty}^{\infty} f_{X|\Theta=\theta}(x)dx = 1.$$

The integral above equals one because

$$\int_{-\infty}^{\infty} f_{X|\Theta=\theta}(x)dx = \int_{-\infty}^{\infty} \frac{f_{X,\Theta}(x,\theta)}{f_\Theta(\theta)}dx = \frac{\int_{-\infty}^{\infty} f_{X,\Theta}(x,\theta)dx}{f_\Theta(\theta)} = \frac{f_\Theta(\theta)}{f_\Theta(\theta)} = 1.$$

Because $f_{X|\Theta=\theta}(x)$ integrates to one as a function of x and because the denominator $f_\Theta(\theta)$ in the definition (56) does not depend on x, it is common to write

$$f_{X|\Theta=\theta}(x) \propto f_{X,\Theta}(x,\theta). \tag{57}$$

The symbol $\propto$ here stands for "proportional to" and the above statement means that $f_{X|\Theta=\theta}(x)$, as a function of x, is proportional to $f_{X,\Theta}(x,\theta)$. The proportionality constant then has to be $f_\Theta(\theta)$ because that is equal to the value of the integral of $f_{X,\Theta}(x,\theta)$ as x ranges over $(-\infty,\infty)$.

The proportionality statement (57) often makes calculations involving conditional densities much simpler. To illustrate this, let us revisit the calculations in Examples (28.1) and (28.2) respectively.

Example 29.1 (Example 28.1 revisited)**.** *Suppose X and Θ are independent random variables having the $Gamma(\alpha,\lambda)$ and $Gamma(\beta,\lambda)$ distributions respectively. What then is the conditional density of X given $X + \Theta = 1$? By (57),*

$$\begin{aligned}
f_{X|X+\Theta=1}(x) &\propto f_{X,X+\Theta}(x,1) \\
&= f_{X,\Theta}(x,1-x) \\
&= f_X(x)f_\Theta(1-x) \\
&\propto e^{-\lambda x}x^{\alpha-1}I\{x>0\}e^{-\lambda(1-x)}(1-x)^{\beta-1}I\{1-x>0\} \\
&\propto x^{\alpha-1}(1-x)^{\beta-1}I\{0<x<1\}
\end{aligned}$$

which immediately implies that $X|X + \Theta = 1$ has the Beta distribution with parameters α and β.

Example 29.2 (Example 28.2 revisited)**.** *Suppose X and Y are independent $Unif(0,1)$ random variables. What is $f_{U|V=v}$ where $U = \min(X,Y)$ and $V = \max(X,Y)$ and $0 < v < 1$?*

Write

$$\begin{aligned} f_{U|V=v}(u) &\propto f_{U,V}(u,v) \\ &= 2f_U(u)f_V(v)I\{u < v\} \\ &\propto f_U(u)I\{u < v\} \\ &= I\{0 < u < 1\}I\{u < v\} = I\{0 < u < \min(v,1)\} \end{aligned}$$

Thus for $v < 1$, the conditional density of U given $V = v$ is the uniform density on $[0, v]$. For $v > 1$, the conditional density of U given $V = 1$ is not defined as the density of V at $v > 1$ equals 0.

30 Conditional Densities and Independence

X and Θ are independent if and only if $f_{X|\Theta=\theta} = f_X$ for every value of θ. This latter statement is precisely equivalent to $f_{X,\Theta}(x,\theta) = f_X(x)f_\Theta(\theta)$. By switching roles of X and Θ, it also follows that X and Θ are independent if and only if $f_{\Theta|X=x} = f_\Theta$ for every x.

It is also not hard to see that X and Θ are independent if and only if the conditional density of X given $\Theta = \theta$ is the same for all values of θ for which $f_\Theta(\theta) > 0$.

Example 30.1 (Back to the Gamma example)**.** *We have previously seen that when $X \sim Gamma(\alpha, \lambda)$ and $Y \sim Gamma(\beta, \lambda)$, then*

$$X|(X + \Theta = 1) \sim Beta(\alpha, \beta).$$

This can be also be directly seen (using the observation that $X/(X + \Theta)$ is distributed as $Beta(\alpha, \beta)$ and that $X/(X + \Theta)$ is independent of $X + \Theta$) as follows:

$$X|(X + \Theta = 1) \overset{d}{=} \frac{X}{1}|(X + \Theta = 1) \overset{d}{=} \frac{X}{X + \Theta}|(X + \Theta = 1) \overset{d}{=} \frac{X}{X + \Theta} \sim Beta(\alpha, \beta).$$

Note that we removed the conditioning on $X + \Theta = 1$ in the last step above because $X/(X + \Theta)$ is independent of $X + \Theta$.

31 Law of Total Probability and Bayes Rule for Continuous Random Variables

Let us now get back to the properties of conditional densities.

1. From the definition of $f_{X|\Theta=\theta}(x)$, it directly follows that

   $$f_{X,\Theta}(x,\theta) = f_{X|\Theta=\theta}(x)f_\Theta(\theta).$$

   This tells us how to compute the joint density of X and Θ using knowledge of the marginal of Θ and the conditional density of X given Θ.

2. **Law of Total Probability for Continuous Random Variables**: Recall the law of total probability for discrete random variables in (52). The analogous statement for continuous random variables is

   $$f_X(x) = \int f_{X|\Theta=\theta}(x)f_\Theta(\theta)d\theta.$$

   This follows directly from the definition of $f_{X|\Theta=\theta}(x)$. This formula allows us to deduce the marginal density of X using knowledge of the conditional density of X given Θ and the marginal density of Θ.

3. **Bayes Rule for Continuous Random Variables**: Recall the Bayes rule for discrete random variables in (53). The analogous statement for continuous random variables is

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x) f_\Theta(\theta)}{\int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta}.$$

   This allows to deduce the conditional density of Θ given X using knowledge of the conditional density of X given Θ and the marginal density of Θ.

32 Law of Total Probability and Bayes Rule for Continuous Random Variables

Suppose X and Θ are continuous random variables having a joint density $f_{X,\Theta}$. We have then seen that

$$f_{X|\Theta=\theta}(x) = \frac{f_{X,\Theta}(x,\theta)}{f_\Theta(\theta)}.$$

Note that the denominator in the right hand side above does not involve x so we can write

$$f_{X|\Theta=\theta}(x) \propto f_{X,\Theta}(x,\theta).$$

Note that we also have

$$f_{\Theta|X=x}(\theta) \propto f_{X,\Theta}(x,\theta).$$

The following are immediate consequences of the definition of conditional density.

1. We have
$$f_{X,\Theta}(x,\theta) = f_{X|\Theta=\theta}(x) f_\Theta(\theta).$$

2. The Law of Total Probability states that
$$f_X(x) = \int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta$$

3. The Bayes rule states
$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x) f_\Theta(\theta)}{\int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta}$$

Here are two examples illustrating these.

Example 32.1. *Suppose $\Theta \sim N(\mu, \tau^2)$ and $X|\Theta = \theta \sim N(\theta, \sigma^2)$. What then is the marginal density of X and the conditional density of $\Theta|X = x$?*

To obtain the marginal density of X, we use the LTP which says

$$f_X(x) = \int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta$$

Now

$$f_{X|\Theta=\theta}(x) f_\Theta(\theta) = \frac{1}{2\pi\tau\sigma} \exp\left(-\frac{1}{2}\left\{\frac{(\theta-\mu)^2}{\tau^2} + \frac{(x-\theta)^2}{\sigma^2}\right\}\right)$$

The term in the exponent above can be simplified as

$$\begin{aligned}\frac{(\theta-\mu)^2}{\tau^2}+\frac{(x-\theta)^2}{\sigma^2} &= \frac{(\theta-x+x-\mu)^2}{\tau^2}+\frac{(x-\theta)^2}{\sigma^2}\\ &= (\theta-x)^2\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)+\frac{2(\theta-x)(x-\mu)}{\tau^2}+\frac{(x-\mu)^2}{\tau^2}\\ &= \left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\theta-\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2+\frac{(x-\mu)^2}{\tau^2+\sigma^2}\end{aligned}$$

where I skipped a few steps to get to the last equality (complete the square and simplify the resulting expressions).

As a result

$$f_{X|\Theta=\theta}(x)f_\Theta(\theta)=\frac{1}{2\pi\tau\sigma}\exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\theta-\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2\right)\exp\left(-\frac{(x-\mu)^2}{2(\tau^2+\sigma^2)}\right)$$

Consequently,

$$\begin{aligned}f_X(x) &= \int\frac{1}{2\pi\tau\sigma}\exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\theta-\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2\right)\exp\left(-\frac{(x-\mu)^2}{2(\tau^2+\sigma^2)}\right)d\theta\\ &= \frac{1}{2\pi\tau\sigma}\exp\left(-\frac{(x-\mu)^2}{2(\tau^2+\sigma^2)}\right)\int\exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\theta-\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2\right)d\theta\\ &= \frac{1}{2\pi\tau\sigma}\exp\left(-\frac{(x-\mu)^2}{2(\tau^2+\sigma^2)}\right)\sqrt{2\pi}\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)^{-1/2}\\ &= \frac{1}{\sqrt{2\pi(\tau^2+\sigma^2)}}\exp\left(-\frac{(x-\mu)^2}{2(\tau^2+\sigma^2)}\right)\end{aligned}$$

which gives

$$X\sim N(0,\tau^2+\sigma^2).$$

To obtain $f_{\Theta|X=x}(\theta)$, we use the Bayes rule as:

$$f_{\Theta|X=x}(\theta)=\frac{f_{X|\Theta=\theta}(x)f_\Theta(\theta)}{f_X(x)}=\frac{\sqrt{\tau^2+\sigma^2}}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\theta-\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2\right)$$

which means that

$$\Theta|X=x\sim N\left(\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2},\frac{1}{1/\sigma^2+1/\tau^2}\right).$$

For a normal density with mean m and variance v^2, the inverse of the variance $1/v^2$ is called the precision. The above calculation therefore reveals that the precision of the conditional distribution of Θ given X equals the sum of the precisions of the distribution of Θ and the distribution of X respectively.

In statistical terminology, it is common to call:

1. *the marginal distribution of Θ as the prior distribution of the unknown parameter θ.*
2. *the conditional distribution of $X|\Theta=\theta$ as the distribution of the data conditioned on the value of the true parameter.*
3. *the conditional distribution of $\Theta|X=x$ as the posterior distribution of Θ given the data.*

In this particular example, the mean of the posterior distribution is a weighted linear combination of the prior mean as well as the data where the weights are proportional to the precisions. Also, posterior precision equals the sum of the prior precision and the data precision which informally means, in particular, that the posterior is more precise than the prior.

Example 32.2. *Suppose $\Theta \sim Gamma(\alpha, \lambda)$ and $X|\Theta = \theta \sim Exp(\theta)$. What is the marginal density of X and the conditional density of Θ given $X = x$?*

For the marginal of X, use the LTP:

$$\begin{aligned} f_X(x) &= \int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta \\ &= \int_0^\infty \theta e^{-\theta x} \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda\theta} \theta^{\alpha-1} d\theta \\ &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \theta^\alpha e^{-(\lambda+x)\theta} d\theta = \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha+1)}{(\lambda+x)^{\alpha+1}} = \frac{\alpha\lambda^\alpha}{(\lambda+x)^{\alpha+1}}. \end{aligned}$$

This is called the Lomax distribution with shape parameter $\alpha > 0$ and scale/rate parameter $\lambda > 0$ (see `https://en.wikipedia.org/wiki/Lomax_distribution`).

For the conditional distribution of Θ given $X = x$, argue via proportionality that

$$\begin{aligned} f_{\Theta|X=x}(\theta) &\propto f_{X|\Theta=\theta}(x) f_\Theta(\theta) \\ &= \theta e^{-\theta x} \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda\theta} \theta^{\alpha-1} I\{\theta > 0\} \\ &= \theta^\alpha e^{-(\lambda+x)\theta} I\{\theta > 0\} \end{aligned}$$

which means that

$$\Theta|X = x \sim Gamma(\alpha + 1, \lambda + x).$$

33 LTP and Bayes Rule for general random variables

The LTP describes how to compute the distribution of X based on knowledge of the conditional distribution of X given $\Theta = \theta$ as well as the conditional distribution of Θ. The Bayes rule describes how to compute the conditional distribution of Θ given $X = x$ based on the same knowledge of the conditional distribution of X given $\Theta = \theta$ as well as the conditional distribution of Θ. We have so far looked at the LTP and Bayes rule when X and Θ are both discrete or when they are both continuous. Now we shall also consider the cases when one of them is discrete and the other is continuous.

33.1 X and Θ are both discrete

In this case, we have seen that the LTP is

$$\mathbb{P}\{X = x\} = \sum_\theta \mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}$$

and the Bayes rule is

$$\mathbb{P}\{\Theta = \theta|X = x\} = \frac{\mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\mathbb{P}\{X = x\}} = \frac{\mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\sum_\theta \mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}.$$

33.2 X and Θ are both continuous

Here LTP is

$$f_X(x) = \int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta$$

and Bayes rule is

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x) f_\Theta(\theta)}{f_X(x)} = \frac{f_{X|\Theta=\theta}(x) f_\Theta(\theta)}{\int f_{X|\Theta=\theta}(x) f_\Theta(\theta) dx}.$$

33.3 X is discrete while Θ is continuous

LTP is

$$\mathbb{P}\{X = x\} = \int \mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta) d\theta$$

and Bayes rule is

$$f_{\Theta|X=x}(\theta) = \frac{\mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta)}{\mathbb{P}\{X = x\}} = \frac{\mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta)}{\int \mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta) d\theta}.$$

33.4 X is continuous while Θ is discrete

LTP is

$$f_X(x) = \sum_\theta f_{X|\Theta=\theta}(x) \mathbb{P}\{\Theta = \theta\}$$

and Bayes rule is

$$\mathbb{P}\{\Theta = \theta|X = x\} = \frac{f_{X|\Theta=\theta}(x) \mathbb{P}\{\Theta = \theta\}}{f_X(x)} = \frac{f_{X|\Theta=\theta}(x) \mathbb{P}\{\Theta = \theta\}}{\sum_\theta f_{X|\Theta=\theta}(x) \mathbb{P}\{\Theta = \theta\}}$$

These formulae are useful when the conditional distribution of X given $\Theta = \theta$ as well as the marginal distribution of Θ are easy to determine (or are given as part of the model specification) and the goal is to determine the marginal distribution of X as well as the conditional distribution of Θ given $X = x$.

We shall now look at two applications of the LTP and Bayes Rule to when one of X and Θ is discrete and the other is continuous.

Example 33.1. *Suppose that Θ is the uniformly distributed on $(0, 1)$ and let $X|\Theta = \theta$ has the binomial distribution with parameters n and θ (i.e., conditioned on $\Theta = \theta$, the random variable X is distributed as the number of successes in n independent tosses of a coin with probability of success θ). What then is the marginal distribution of X as well as the conditional distribution of Θ given $X = x$?*

Note that this is a situation where X is discrete (taking values in $0, 1, \dots, n$) and Θ is continuous (taking values in the interval $(0, 1)$). To compute the marginal distribution of X, we use the appropriate LTP to write (for $x = 0, 1, \dots, n$)

$$\begin{aligned}
\mathbb{P}\{X = x\} &= \int \mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta) d\theta \\
&= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} d\theta \\
&= \binom{n}{x} Beta(x + 1, n - x + 1) \\
&= \binom{n}{x} \frac{\Gamma(x + 1)\Gamma(n - x + 1)}{\Gamma(n + 2)} = \frac{n!}{(n - x)!x!} \frac{x!(n - x)!}{(n + 1)!} = \frac{1}{n + 1}
\end{aligned}$$

which means that X is (discrete) uniformly distributed on the finite set $\{0, 1, \dots, n\}$.

Let us now calculate the posterior distribution of Θ given $X = x$. Using the Bayes rule, we obtain

$$\begin{aligned} f_{\Theta|X=x}(\theta) &= \frac{\mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta)}{\mathbb{P}\{X = x\}} \\ &= \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{1/(n+1)} \propto \theta^x(1-\theta)^{n-x} \end{aligned}$$

for $0 < \theta < 1$. From here, it immediately follows that

$$\Theta|X = x \sim Beta(x+1, n-x+1).$$

The mean of the $Beta(\alpha, \beta)$ distribution is $\alpha/(\alpha+\beta)$. Therefore the mean of the conditional distribution of Θ given $X = x$ (also known as the posterior mean) equals

$$\mathbb{E}(\Theta|X = x) = \frac{x+1}{n+2}.$$

As the prior mean equals $1/2$ and we can write

$$\frac{x+1}{n+2} = \left(\frac{n}{n+2}\right)\frac{x}{n} + \left(\frac{2}{n+2}\right)\frac{1}{2},$$

it follows that the posterior mean falls between the prior mean and x/n. As n becomes large, the posterior mean approaches x/n.

Example 33.2 (Statistical Classification)**.** *In a statistical classification problem, the random variable Θ is discrete and X is usually continuous. The simplest situation is when Θ is binary. Let us say that*

$$\mathbb{P}\{\Theta = 1\} = p \quad \text{and} \quad \mathbb{P}\{\Theta = 0\} = 1 - p.$$

Also assume that the conditional density of X given $\Theta = 0$ is f_0 and that the conditional density of X given $\Theta = 1$ is f_1 i.e.,

$$X|\Theta = 0 \sim f_0 \quad \text{and} \quad X|\Theta = 1 \sim f_1.$$

Using the LTP, we see that the marginal density of X equals

$$f_X = (1-p)f_0 + pf_1.$$

In other words, f_X is a ***mixture*** *of f_0 and f_1 with the mixing weights being equal to the marginal probabilities of Θ.*

According to the Bayes rule, the conditional distribution of Θ given $X = x$ is given by

$$\mathbb{P}\{\Theta = 0|X = x\} = \frac{f_{X|\Theta=0}(x)\mathbb{P}\{\Theta = 0\}}{f_X(x)} = \frac{(1-p)f_0(x)}{(1-p)f_0(x) + pf_1(x)}$$

and

$$\mathbb{P}\{\Theta = 1|X = x\} = \frac{pf_1(x)}{(1-p)f_0(x) + pf_1(x)}.$$

These are also referred to as the posterior probabilities of Θ given $X = x$.

34 Conditional Joint Densities

Given continuous random variables $X_1, \dots, X_m, Y_1, \dots, Y_k$, the conditional joint density of $Y_1, \dots, Y_k$ given $X_1 = x_1, X_2 = x_2, \dots, X_m = x_m$ is defined as

$$f_{Y_1,\dots,Y_k|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k) := \frac{f_{X_1,\dots,X_m,Y_1,\dots,Y_k}(x_1, \dots, x_m, y_1, \dots, y_k)}{f_{X_1,\dots,X_m}(x_1, \dots, x_m)}$$

provided $x_1, \dots, x_m$ are such that $f_{X_1,\dots,X_m}(x_1, \dots, x_m) > 0$.

Here are some simple but important properties of conditional joint densities.

1. For every $x_1, \dots, x_m, y_1, \dots, y_k$, we have

$$f_{X_1,\dots,X_m,Y_1,\dots,Y_k}(x_1, \dots, x_m, y_1, \dots, y_k) = f_{Y_1,\dots,Y_k|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k) f_{X_1,\dots,X_m}(x_1, \dots, x_m).$$

2. The joint density of every set of random variables $Y_1, \dots, Y_n$ satisfies the following:

$$f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = f_{Y_n|Y_1=y_1,\dots,Y_{n-1}=y_{n-1}}(y_n) f_{Y_{n-1}|Y_1=y_1,\dots,Y_{n-1}=y_{n-2}}(y_{n-1}) \dots f_{Y_2|Y_1=y_1}(y_2) f_{Y_1}(y_1).$$

3. This is a generalization of the previous fact. The conditional joint density

$$f_{Y_1,\dots,Y_n|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_n)$$

of $Y_1, \dots, Y_n$ given $X_1 = x_1, \dots, X_m = x_m$ equals the product

$$\prod_{i=1}^{n} f_{Y_i|Y_1=y_1,\dots,Y_{i-1}=y_{i-1},X_1=x_1,\dots,X_m=x_m}(y_i).$$

4. This can be viewed as a **law of total conditional probability**: For random variables $Y_1, \dots, Y_k, X_1, \dots, X_m$ and Θ, we have

$$f_{Y_1,\dots,Y_k|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k) = \int f_{Y_1,\dots,Y_k,\Theta|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k, \theta) f_{\Theta|X_1=x_1,\dots,X_m=x_m}(\theta) d\theta.$$

We shall look at some applications of the above facts in the next class.

35 Conditional Joint Densities

Given continuous random variables $X_1, \dots, X_m, Y_1, \dots, Y_k$, the conditional joint density of $Y_1, \dots, Y_k$ given $X_1 = x_1, X_2 = x_2, \dots, X_m = x_m$ is defined as

$$f_{Y_1,\dots,Y_k|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k) := \frac{f_{X_1,\dots,X_m,Y_1,\dots,Y_k}(x_1, \dots, x_m, y_1, \dots, y_k)}{f_{X_1,\dots,X_m}(x_1, \dots, x_m)}$$

provided $x_1, \dots, x_m$ are such that $f_{X_1,\dots,X_m}(x_1, \dots, x_m) > 0$.

Example 35.1. *Suppose $U_1, \dots, U_n$ are independent observations having the uniform density on $(0, 1)$. What is the conditional joint density of $U_{(1)}, \dots, U_{(n-1)}$ given $U_{(n)} = u$?*

By definition,

$$f_{U_{(1)},\dots,U_{(n-1)}|U_{(n)}=u}(u_1, \dots, u_{n-1}) = \frac{f_{U_{(1)},\dots,U_{(n)}}(u_1, \dots, u_{n-1}, u)}{f_{U_{(n)}}(u)}.$$

By the joint distribution of order statistics that we worked out previously, it follows first that the above quantity is non-zero only when $0 < u_1 < \dots < u_{n-1} < u < 1$ and it is then equal to

$$f_{U_{(1)},\dots,U_{(n-1)}|U_{(n)}=u}(u_1, \dots, u_{n-1}) = \frac{n!}{nu^{n-1}}.$$

For the denominator above, we used the fact that $U_{(n)} \sim Beta(n, 1)$. We have thus proved that

$$f_{U_{(1)},\dots,U_{(n-1)}|U_{(n)}=u}(u_1, \dots, u_{n-1}) = (n-1)! \left(\frac{1}{u}\right)^{n-1} \qquad \text{for } 0 < u_1 < \dots < u_{n-1} < u < 1.$$

Note that the right hand side above is the joint density of the order statistics of $(n-1)$ i.i.d observations drawn from the uniform distribution on the interval $(0, u)$. We have therefore proved that, conditioned on $U_{(n)} = u$, the joint density of $U_{(1)}, \dots, U_{(n-1)}$ is the same as the joint density of the order statistics of $(n-1)$ i.i.d observations drawn from the uniform distribution on $(0, u)$.

Here are some simple but important properties of conditional joint densities.

1. For every $x_1, \dots, x_m, y_1, \dots, y_k$, we have

$$f_{X_1,\dots,X_m,Y_1,\dots,Y_k}(x_1, \dots, x_m, y_1, \dots, y_k) = f_{Y_1,\dots,Y_k|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k) f_{X_1,\dots,X_m}(x_1, \dots, x_m).$$

2. The joint density of every set of random variables $Y_1, \dots, Y_n$ satisfies the following:

$$f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) = f_{Y_n|Y_1=y_1,\dots,Y_{n-1}=y_{n-1}}(y_n) f_{Y_{n-1}|Y_1=y_1,\dots,Y_{n-1}=y_{n-2}}(y_{n-1}) \dots f_{Y_2|Y_1=y_1}(y_2) f_{Y_1}(y_1).$$

3. This is a generalization of the previous fact. The conditional joint density

$$f_{Y_1,\dots,Y_n|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_n)$$

of $Y_1, \dots, Y_n$ given $X_1 = x_1, \dots, X_m = x_m$ equals the product

$$\prod_{i=1}^{n} f_{Y_i|Y_1=y_1,\dots,Y_{i-1}=y_{i-1},X_1=x_1,\dots,X_m=x_m}(y_i).$$

4. This can be viewed as a **law of total conditional probability**: For random variables $Y_1, \dots, Y_k, X_1, \dots, X_m$ and Θ, we have

$$f_{Y_1,\dots,Y_k|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k) = \int f_{Y_1,\dots,Y_k,\Theta|X_1=x_1,\dots,X_m=x_m}(y_1, \dots, y_k, \theta) f_{\Theta|X_1=x_1,\dots,X_m=x_m}(\theta) d\theta.$$

Here are some applications of the above facts.

Example 35.2 (Joint density of an autoregressive process)**.** *Suppose $X_1, Z_2, \dots, Z_n$ are independent random variables with $Z_2, \dots, Z_n$ being distributed as $N(0, \sigma^2)$. Define new random variables $X_2, \dots, X_n$ via*

$$X_i = \phi X_{i-1} + Z_i \qquad \text{for } i = 2, \dots, n$$

where ϕ is some real number. The process $X_1, \dots, X_n$ is called an autoregressive process of order 1. What is the conditional joint density of $X_2, \dots, X_n$ given $X_1 = x_1$? What is the joint density of $X_1, \dots, X_n$?

Let us first calculate the conditional joint density of $X_2, \dots, X_n$ given $X_1 = x_1$. For this, write

$$f_{X_2,\dots,X_n|X_1=x_1}(x_2, \dots, x_n) = \prod_{i=2}^{n} f_{X_i|X_1=x_1,\dots,X_{i-1}=x_{i-1}}(x_i) \tag{58}$$

Now for each $i = 2, \dots, n$, observe that

$$\begin{aligned} X_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1} &\overset{d}{=} (\phi X_{i-1} + Z_i) \,| X_1 = x_1, \dots, X_{i-1} = x_{i-1} \\ &\overset{d}{=} (\phi x_{i-1} + Z_i) \,| X_1 = x_1, \dots, X_{i-1} = x_{i-1} \\ &\overset{d}{=} \phi x_{i-1} + Z_i \sim N(\phi x_{i-1}, \sigma^2). \end{aligned}$$

We were able to remove conditioning on $X_1 = x_1, \dots, X_{i-1} = x_{i-1}$ above because $X_1, \dots, X_{i-1}$ only depend on $X_1, Z_2, \dots, Z_{i-1}$ and hence are independent of Z_i.

From the above chain of assertions, we deduce that

$$f_{X_i|X_1=x_1,\dots,X_{i-1}=x_{i-1}}(x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \phi x_{i-1})^2}{2\sigma^2}\right) \qquad \text{for } i = 2, \dots, n.$$

Combining with (58)*, we obtain*

$$\begin{aligned} f_{X_2,\dots,X_n|X_1=x_1}(x_2, \dots, x_n) &= \prod_{i=2}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x_i - \phi x_{i-1})^2}{2\sigma^2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=2}^{n}(x_i - \phi x_{i-1})^2\right). \end{aligned}$$

To obtain the joint density of $X_1, \dots, X_n$, write

$$\begin{aligned} f_{X_1,\dots,X_n}(x_1, \dots, x_n) &= f_{X_1}(x_1) f_{X_2,\dots,X_n|X_1=x_1}(x_2, \dots, x_n) \\ &= f_{X_1}(x_1)\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^{n-1} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=2}^{n}(x_i - \phi x_{i-1})^2\right). \end{aligned}$$

In a statistical setting, this joint density is used to estimate the parameters ϕ and σ^2 via maximum likelihood estimation. For this model however, it is easier to work with the conditional density of $X_2, \dots, X_n$ given $X_1 = x_1$ instead of the full joint density of $X_1, \dots, X_n$.

35.1 Application to the Normal prior-Normal data model

Let us now look at the application of the conditional density formulae for the normal prior-normal data model. Here we first have a random variable Θ that has the $N(\mu, \tau^2)$ distribution. We also have random variables $X_1, \dots, X_{n+1}$ such that

$$X_1, \dots, X_{n+1}|\Theta = \theta \sim^{i.i.d} N(\theta, \sigma^2).$$

In other words, conditional on $\Theta = \theta$, the random variables $X_1, \dots, X_{n+1}$ are i.i.d $N(\theta, \sigma^2)$.

Let us first find the conditional distribution of Θ given $X_1 = x_1, \dots, X_n = x_n$. The answer to this turns out to be

$$\Theta|X_1 = x_1, \dots, X_n = x_n \sim N\left(\frac{n\bar{x}_n/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \frac{1}{n/\sigma^2 + 1/\tau^2}\right) \tag{59}$$

where $\bar{x}_n := (x_1 + \dots + x_n)/n$. Let us see why this is true below. Note first that we had solved this problem for $n = 1$ in the last class where we proved the following:

$$\Theta \sim N(\mu, \tau^2), X|\Theta = \theta \sim N(\theta, \sigma^2) \implies \Theta|X = x \sim N\left(\frac{x/\sigma^2 + \mu/\tau^2}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\sigma^2 + 1/\tau^2}\right), X \sim N(\mu, \sigma^2 + \tau^2).$$

The result (59) for general $n \geq 1$ can actually be deduced from the above result for $n = 1$. There are two ways of seeing this.

Method One: We use mathematical induction on $n \geq 1$. We already know that (59) is true for $n = 1$. Assume that it is true for n and we shall try to prove it for $n + 1$. The key to this is to note that

$$(\Theta|X_1 = x_1, \dots, X_{n+1} = x_{n+1}) \overset{d}{=} \tilde{\Theta}|Y = x_{n+1} \tag{60}$$

where

$$Y|\tilde{\Theta} = \theta \sim N(\theta, \sigma^2) \quad \text{and} \quad \tilde{\Theta} \sim \Theta|X_1 = x_1, \dots, X_n = x_n.$$

In words, (60) states that the posterior of Θ after observing $(n+1)$ observations $X_1 = x_1, \dots, X_{n+1} = x_{n+1}$ is the same as the posterior after observing one observation $Y = x_{n+1}$ under the prior $\Theta|X_1 = x_1, \dots, X_n = x_n$.

To formally see why (60) is true, just note that

$$\begin{aligned} f_{\Theta|X_1=x_1,\dots,X_n=x_n,X_{n+1}=x_{n+1}}(\theta) &\propto f_{X_{n+1}|\Theta=\theta,X_1=x_1,\dots,X_n=x_n}(x_{n+1}) f_{\Theta|X_1=x_1,\dots,X_n=x_n}(\theta) \\ &= f_{X_{n+1}|\Theta=\theta}(x_{n+1}) f_{\Theta|X_1=x_1,\dots,X_n=x_n}(\theta). \end{aligned}$$

The first equality is a consequence of the properties of conditional densities. The second equality above is a consequence of the fact that X_{n+1} is independent of $X_1, \dots, X_n$ **conditional on** Θ.

The statement (60) allows us to use the result for $n = 1$ and the induction hypothesis that (59) holds for n. Indeed, using the $n = 1$ result for

$$\mu = \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2} \quad \text{and} \quad \tau^2 = \frac{1}{n/\sigma^2 + 1/\tau^2}$$

and $x = x_{n+1}$, we deduce that $\Theta | X_1 = x_1, \dots, X_{n+1} = x_{n+1}$ is a normal distribution with mean

$$\frac{x_{n+1}/\sigma^2 + \frac{n\bar{x}/\sigma^2 + \mu/\tau^2}{n/\sigma^2+1/\tau^2}\left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right)}{\frac{1}{\sigma^2} + \frac{n}{\sigma^2} + \frac{1}{\tau^2}} = \frac{\frac{x_1+\dots+x_{n+1}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n+1}{\sigma^2} + \frac{1}{\tau^2}} = \frac{\frac{(n+1)\bar{x}_{n+1}}{\sigma^2} + \frac{\mu}{\tau^2}}{\frac{n+1}{\sigma^2} + \frac{1}{\tau^2}}$$

and variance

$$\frac{1}{\frac{1}{\sigma^2} + \frac{n}{\sigma^2} + \frac{1}{\tau^2}} = \frac{1}{\frac{n+1}{\sigma^2} + \frac{1}{\tau^2}}.$$

This proves (59) for $n + 1$. The proof of (59) is complete by induction.

Method Two. The second method for proving (59) proceeds more directly by writing:

$$\begin{aligned} f_{\Theta|X_1=x_1,\dots,X_n=x_n}(\theta) &\propto f_{X_1,\dots,X_n|\Theta=\theta}(x_1, \dots, x_n) f_\Theta(\theta) \\ &= f_{X_1|\Theta=\theta}(x_1) \dots f_{X_n|\Theta=\theta}(x_n) f_\Theta(\theta) \\ &\propto \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \theta)^2\right) \exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right) \\ &= \exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n (x_i - \bar{x}_n)^2 + n(\bar{x}_n - \theta)^2\right]\right) \exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right) \\ &\propto \exp\left(-\frac{n}{2\sigma^2}(\bar{x}_n - \theta)^2\right) \exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right) \\ &= \exp\left(-\frac{1}{2(\sigma^2/n)}(\bar{x}_n - \theta)^2\right) \exp\left(-\frac{1}{2\tau^2}(\theta - \mu)^2\right). \end{aligned}$$

This now resembles the calculation we did previously for $n = 1$. The only difference being that x is now replaced by $\bar{x}_n$ and σ^2 is replaced by σ^2/n. Therefore the $n = 1$ result applied to $x \to \bar{x}_n$ and $\sigma^2 \to \sigma^2/n$ should yield (59). This proves (59).

Let us now compute the conditional density of X_{n+1} given $X_1 = x_1, \dots, X_n = x_n$. For this, we can use the law of total conditional probability to write

$$\begin{aligned} f_{X_{n+1}|X_1=x_1,\dots,X_n=x_n}(x) &= \int f_{X_{n+1}|\Theta=\theta,X_1=x_1,\dots,X_n=x_n}(x) f_{\Theta|X_1=x_1,\dots,X_n=x_n}(\theta) d\theta \\ &= \int f_{X_{n+1}|\Theta=\theta}(x) f_{\Theta|X_1=x_1,\dots,X_n=x_n}(\theta) d\theta \end{aligned}$$

This again resembles the calculation of the marginal density of X in the $n = 1$ problem (where the answer is $X \sim N(\mu, \tau^2 + \sigma^2)$). The only difference is that the prior $N(\mu, \tau^2)$ is now replaced by the posterior density which is given by (59). We therefore obtain that

$$(X_{n+1}|X_1 = x_1, \dots, X_n = x_n) \sim N\left(\frac{n\bar{x}_n/\sigma^2 + \mu/\tau^2}{n/\sigma^2 + 1/\tau^2}, \sigma^2 + \frac{1}{n/\sigma^2 + 1/\tau^2}\right)$$

36 Conditional Expectation

Given two random variables X and Y, the conditional expectation (or conditional mean) of Y given $X = X$ is denoted by

$$\mathbb{E}\left(Y|X=x\right)$$

and is defined as the expectation of the conditional distribution of Y given $X = x$.

We can write

$$\mathbb{E}\left(Y|X=x\right) = \begin{cases} \int y f_{Y|X=x}(y)dy & : \text{ if } Y \text{ is continuous} \\ \sum_y y\mathbb{P}\{Y=y|X=x\} & : \text{ if } Y \text{ is discrete} \end{cases}$$

More generally

$$\mathbb{E}\left(g(Y)|X=x\right) = \begin{cases} \int g(y) f_{Y|X=x}(y)dy & : \text{ if } Y \text{ is continuous} \\ \sum_y g(y)\mathbb{P}\{Y=y|X=x\} & : \text{ if } Y \text{ is discrete} \end{cases}$$

and also

$$\mathbb{E}\left(g(X,Y)|X=x\right) = \mathbb{E}\left(g(x,Y)|X=x\right) = \begin{cases} \int g(x,y) f_{Y|X=x}(y)dy & : \text{ if } Y \text{ is continuous} \\ \sum_y g(x,y)\mathbb{P}\{Y=y|X=x\} & : \text{ if } Y \text{ is discrete} \end{cases}$$

The most important fact about conditional expectation is the **Law of Iterated Expectation** (also known as the **Law of Total Expectation**). We shall see this next.

36.1 Law of Iterated/Total Expectation

The law of total expectation states that

$$\mathbb{E}(Y) = \begin{cases} \int \mathbb{E}\left(Y|X=x\right) f_X(x)dx & : \text{ if } X \text{ is continuous} \\ \sum_x \mathbb{E}\left(Y|X=x\right)\mathbb{P}\{X=x\} & : \text{ if } X \text{ is discrete} \end{cases}$$

Basically the law of total expectation tells us how to compute the expectation of $\mathbb{E}(Y)$ using knowledge of the conditional expectation of Y given $X = x$. Note the similarity to law of total probability which specifies how to compute the marginal distribution of Y using knowledge of the conditional distribution of Y given $X = x$.

The law of total expectation can be proved as a consequence of the law of total probability. The proof when Y and X are continuous is given below. The proof in other cases (when one or both of Y and X are discrete) is similar and left as an exercise.

Proof of the law of total expectation: Assume that Y and X are both continuous. Then

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

By the law of total probability, we have

$$\begin{aligned} \mathbb{E}(Y) &= \int y f_Y(y)dy \\ &= \int y \left(\int f_{Y|X=x}(y) f_X(x)dx \right) dy \\ &= \int \left(\int y f_{Y|X=x}(y)dy \right) f_X(x)dx = \int \mathbb{E}(Y|X=x) f_X(x)dx. \end{aligned}$$

which proves the law of total expectation.

There is an alternate more succinct form of stating the law of total expectation which justifies calling the law of **iterated** expectation. We shall see this next. Note that $\mathbb{E}(Y|X = x)$ depends on x. In other words, $\mathbb{E}(Y|X = x)$ is a function of x. Let us denote this function by $h(\cdot)$:

$$h(x) := \mathbb{E}(Y|X = x).$$

If we now apply this function to the random variable X, we obtain a new random variable $h(X)$. This random variable is denoted by simply $\mathbb{E}(Y|X)$ i.e.,

$$\mathbb{E}(Y|X) := h(X).$$

Note that when X is discrete, the expectation of this random variable $\mathbb{E}(Y|X)$ becomes

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(h(X)) = \sum_x h(x)\mathbb{P}\{X = x\} = \sum_x \mathbb{E}(Y|X = x)\mathbb{P}\{X = x\}.$$

And when X is continuous, the expectation of $\mathbb{E}(Y|X)$ is

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(h(X)) = \int h(x)f_X(x)dx = \int \mathbb{E}(Y|X = x)f_X(x)dx.$$

Observe that the right hand sides in these expectations are precisely the terms on the right hand side of the law of total expectation. Therefore the law of total expectation can be rephrased as

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)).$$

Because there are two expectations on the right hand side, the law of total expectation is also known as the Law of Iterated Expectation.

The law of iterated expection has many applications. A couple of simple examples are given below following which we shall explore applications to *risk minimization.*

Example 36.1. *Consider a stick of length ℓ. Break it at a random point X that is chosen uniformly across the length of the stick. Then break the stick again at a random point Y that is also chosen uniformly across the length of the stick. What is the expected length of the final piece?*

According to the description of the problem,

$$Y|X = x \sim Unif(0, x) \quad \text{and} \quad X \sim Unif(0, \ell)$$

and we are required to calculate $\mathbb{E}(Y)$. Note first that $\mathbb{E}(Y|X = x) = x/2$ for every x which means that $\mathbb{E}(Y|X) = X/2$. Hence by the Law of Iterated Expectation,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(X/2) = \ell/4.$$

36.2 Law of Iterated/Total Expectation

In the last class, we looked at the law of iterated expectation (or the Law of Total Expectation) which stated that

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)).$$

On the right hand side above, $\mathbb{E}(Y|X)$ is the random variable obtained by applying the function $h(x) := \mathbb{E}(Y|X = x)$ to the random variable X (i.e., $\mathbb{E}(Y|X) = h(X)$).

The law of iterated expection has many applications. A couple of simple examples are given below following which we shall explore applications to *risk minimization.*

Example 36.2. *Suppose X, Y, Z are i.i.d $Unif(0,1)$ random variables. Find the value of $\mathbb{P}\{X \leq YZ\}$?*

By the Law of Iterated Expectation,

$$\mathbb{P}\{X \leq YZ\} = \mathbb{E}\left(I\{X \leq YZ\}\right) = \mathbb{E}\left[\mathbb{E}\left(I\{X \leq YZ\}|YZ\right)\right] = \mathbb{E}(YZ) = \mathbb{E}(Y)\mathbb{E}(Z) = 1/4.$$

Example 36.3 (Sum of a random number of i.i.d random variables)**.** *Suppose $X_1, X_2, \dots$ are i.i.d random variables with $\mathbb{E}(X_i) = \mu$. Suppose also that N is a discrete random variable that takes values in $\{1, 2, \dots,\}$ and that is independent of $X_1, X_2, \dots$. Define*

$$S := X_1 + X_2 + \cdots + X_N.$$

In other words, S is the sum of a random number (N) of the random variables X_i. The law of iterated expectation can be used to compute the expectation of S as follows:

$$\mathbb{E}(S) = \mathbb{E}(\mathbb{E}(S|N)) = \mathbb{E}(N\mu) = (\mu)(\mathbb{E}N) = (\mathbb{E}N)(\mathbb{E}X_1).$$

This fact is actually a special case of a general result called ***Wald's identity****.*

36.3 Application of the Law of Total Expectation to Statistical Risk Minimization

The law of the iterated expectation has important applications to statistical risk minimization problems. The simplest of these problems is the following.

Problem 1: Given two random variables X and Y, what is the function $g^*(X)$ of X that minimizes

$$R(g) := \mathbb{E}\left(g(X) - Y\right)^2$$

over all functions g? The resulting random variable $g^*(X)$ can be called the Best Predictor of Y as a function of X in terms of expected squared error.

To find g^*, we use the law of iterated expectation to write

$$R(g) = \mathbb{E}\left(g(X) - Y\right)^2 = \mathbb{E}\left\{\mathbb{E}\left[\left(g(X) - Y\right)^2 |X\right]\right\}$$

The value $g^*(x)$ which minimizes the inner expectation:

$$\mathbb{E}\left[\left(Y - g(x)\right)^2 |X = x\right]$$

is simply

$$g^*(x) = \mathbb{E}(Y|X = x).$$

This is because $\mathbb{E}(Z - c)^2$ is minimized as c varies over $\mathbb{R}$ at $c^* = \mathbb{E}(Z)$. We have thus proved that the function $g^*(X)$ which minimizes $R(g)$ over all functions g is given by

$$g^*(X) = \mathbb{E}(Y|X).$$

Thus the function of X which is closest to Y in terms of *expected squared error* is given by the conditional mean $\mathbb{E}(Y|X)$.

Let us now consider a different risk minimization problem.

Problem 2: Given two random variables X and Y, what is the function $g^*(X)$ of X that minimizes

$$R(g) := \mathbb{E}\left|g(X) - Y\right|$$

over all functions g? The resulting random variable $g^*(X)$ can be called the Best Predictor of Y as a function of X in terms of expected absolute error.

To find g^* we use the law of iterated expectation to write

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

The value $g^*(x)$ which minimizes the inner expectation:

$$\mathbb{E}\left[\left|Y - g(x)\right| |X = x\right]$$

is simply given by any conditional median of Y given $X = x$. This is because $\mathbb{E}|Z - c|$ is minimized as c varies over $\mathbb{R}$ at any median of Z. To see this, assume that Z has a density f and write

$$\begin{aligned}
\mathbb{E}|Z - c| &= \int |z - c| f(z) dz \\
&= \int_{-\infty}^{c} (c - z) f(z) dz + \int_{c}^{\infty} (z - c) f(z) dz \\
&= c \int_{-\infty}^{c} f(z) dz - \int_{-\infty}^{c} z f(z) dz + \int_{c}^{\infty} z f(z) dz - c \int_{c}^{\infty} f(z) dz.
\end{aligned}$$

Differentiating with respect to c, we get

$$\frac{d}{dc}\mathbb{E}|Z - c| = \int_{-\infty}^{c} f(z) dz - \int_{c}^{\infty} f(z) dz$$

Therefore when c is a median, the derivative of $\mathbb{E}|Z - c|$ will equal zero. This shows that $c \mapsto \mathbb{E}|Z - c|$ is minimized when c is a median of Z.

We have thus shown that the function $g^*(x)$ which minimizes $R(g)$ over all functions g is given by any conditional mean of Y given $X = x$. Thus the conditional mean of Y given $X = x$ is the function of X that is closest to Y in terms of expected absolute error.

Problem 3: Suppose Y is a binary random variable taking the values 0 and 1 and let X be an arbitrary random variable. What is the function $g^*(X)$ of X that minimizes

$$R(g) := \mathbb{P}\{Y \neq g(X)\}$$

over all functions g? To solve this, again use the law of iterated expectation to write

$$R(g) = \mathbb{P}\{Y \neq g(X)\} = \mathbb{E}\left(\mathbb{P}\left\{Y \neq g(X) | X\right\}\right).$$

In the inner expectation above, we can treat X as a constant so that the problem is similar to minimizing $\mathbb{P}\{Z \neq c\}$ over $c \in \mathbb{R}$ for a binary random variable Z. It is easy to see that $\mathbb{P}\{Z \neq c\}$ is minimized at c^* where

$$c^* = \begin{cases} 1 & : \text{ if } \mathbb{P}\{Z = 1\} > \mathbb{P}\{Z = 0\} \\ 0 & : \text{ if } \mathbb{P}\{Z = 1\} < \mathbb{P}\{Z = 0\} \end{cases}$$

In case $\mathbb{P}\{Z = 1\} = \mathbb{P}\{Z = 0\}$, we can take c^* to be either 0 or 1. From here it can be deduced (via the law of iterated expectation) that the function $g^*(X)$ which minimizes $\mathbb{P}\{Y \neq g(X)\}$ is given by

$$g^*(x) = \begin{cases} 1 & : \text{ if } \mathbb{P}\{Y = 1 | X = x\} > \mathbb{P}\{Y = 0 | X = x\} \\ 0 & : \text{ if } \mathbb{P}\{Y = 1 | X = x\} < \mathbb{P}\{Y = 0 | X = x\} \end{cases}$$

Problem 4: Suppose again that Y is binary taking the values 0 and 1 and let X be an arbitrary random variable. What is the function $g^*(X)$ of X that minimizes

$$R(g) := W_0 \mathbb{P}\{Y \neq g(X), Y = 0\} + W_1 \mathbb{P}\{Y \neq g(X), Y = 1\}.$$

Using an argument similar to the previous problems, deduce that the following function minimizes $R(g)$:

$$g^*(x) = \begin{cases} 1 & : \text{ if } W_1\mathbb{P}\{Y=1|X=x\} > W_0\mathbb{P}\{Y=0|X=x\} \\ 0 & : \text{ if } W_1\mathbb{P}\{Y=1|X=x\} < W_0\mathbb{P}\{Y=0|X=x\} \end{cases}$$

The argument (via the law of iterated expectation) used in the above four problems can be summarized as follows. The function g^* which minimizes

$$R(g) := \mathbb{E}L(Y, g(X))$$

over all functions g is given by

$$g^*(x) = \text{minimizer of } \mathbb{E}(L(Y,c)|X=x) \ \text{ over } c \in \mathbb{R}.$$

37 Conditional Variance

Given two random variables Y and X, the conditional variance of Y given $X = x$ is defined as the variance of the conditional distribution of Y given $X = x$. More formally,

$$Var(Y|X=x) := \mathbb{E}\left[(Y - \mathbb{E}(Y|X=x))^2 \,|X=x\right] = \mathbb{E}\left(Y^2|X=x\right) - (\mathbb{E}(Y|X=x))^2 .$$

Like conditional expectation, the conditional variance $Var(Y|X=x)$ is also a function of x. We can apply this function to the random variable X to obtain a new random variable which we denote by $Var(Y|X)$. Note that

$$Var(Y|X) = \mathbb{E}(Y^2|X) - (\mathbb{E}(Y|X))^2 . \tag{61}$$

Analogous to the Law of Total Expectation, there is a Law of Total Variance as well. This formula says that

$$Var(Y) = \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X)).$$

To prove this formula, expand the right hand side as

$$\begin{aligned} \mathbb{E}(Var(Y|X)) + Var(\mathbb{E}(Y|X)) &= \mathbb{E}\left\{\mathbb{E}(Y^2|X) - (\mathbb{E}(Y|X))^2\right\} + \mathbb{E}\left(\mathbb{E}(Y|X)\right)^2 - (\mathbb{E}(\mathbb{E}(Y|X))^2 \\ &= \mathbb{E}(\mathbb{E}(Y^2|X)) - \mathbb{E}(\mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X))^2 - (\mathbb{E}(Y))^2 \\ &= \mathbb{E}(Y^2) - (\mathbb{E}Y)^2 = Var(Y). \end{aligned}$$

Example 37.1. *We have seen before that*

$$X|\Theta = \theta \sim N(\theta, \sigma^2) \quad \textit{and} \quad \Theta \sim N(\mu, \tau^2) \implies X \sim N(\mu, \sigma^2 + \tau^2).$$

This, of course, means that

$$\mathbb{E}(X) = \mu \quad \textit{and} \quad Var(X) = \sigma^2 + \tau^2.$$

Using the laws of total expectation and total variance, it is possible to prove these directly as follows.

$$\mathbb{E}(X) = \mathbb{E}(\mathbb{E}(X|\Theta)) = \mathbb{E}(\Theta) = \mu$$

and

$$Var(X) = \mathbb{E}(Var(X|\Theta)) + Var(\mathbb{E}(X|\Theta)) = \mathbb{E}(\sigma^2) + Var(\Theta) = \sigma^2 + \tau^2.$$

Example 37.2 (Sum of a random number of i.i.d random variables)**.** *Suppose $X_1, X_2, \dots$ are i.i.d random variables with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2 < \infty$. Suppose also that N is a discrete random variable that takes values in $\{1, 2, \dots, \}$ and that is independent of $X_1, X_2, \dots$. Define*

$$S := X_1 + X_2 + \cdots + X_N.$$

We have seen previously that

$$\mathbb{E}(S) = \mathbb{E}(\mathbb{E}(S|N)) = \mathbb{E}(N\mu) = (\mu)(\mathbb{E}N) = (\mathbb{E}N)(\mathbb{E}X_1).$$

Using the law of total variance, we can calculate $Var(X)$ as follows.

$$Var(S) = \mathbb{E}(Var(S|N)) + Var(\mathbb{E}(S|N)) = \mathbb{E}(N\sigma^2) + Var(N\mu) = \sigma^2(\mathbb{E}N) + \mu^2 Var(N).$$

.

38 Random Vectors

In this section, we view a finite number of random variables as a random vector and go over some basic formulae for the mean and covariance of random vectors.

A random vector is a vector whose entries are random variables. Let $Y = (Y_1, \dots, Y_n)^T$ be a random vector. Its expectation $\mathbb{E}Y$ is defined as a vector whose ith entry is the expectation of Y_i i.e., $\mathbb{E}Y = (\mathbb{E}Y_1, \mathbb{E}Y_2, \dots, \mathbb{E}Y_n)^T$.

The covariance matrix of Y, denoted by $Cov(Y)$, is an $n \times n$ matrix whose (i,j)th entry is the covariance between Y_i and Y_j.

Two important but easy facts about $Cov(Y)$ are:

1. The diagonal entries of $Cov(Y)$ are the variances of $Y_1, \dots, Y_n$. More specifically the (i,i)th entry of the matrix $Cov(Y)$ equals $var(Y_i)$.
2. $Cov(Y)$ is a symmetric matrix i.e., the (i,j)th entry of $Cov(Y)$ equals the (j,i) entry. This follows because $Cov(Y_i, Y_j) = Cov(Y_j, Y_i)$.

39 Random Vectors

In the last class, we started looking at random vectors. A $p \times 1$ random vector Y simply consists of p random variables $Y_1, \dots, Y_p$ i.e., $Y = (Y_1, \dots, Y_p)^T$.

The mean vector Y is given by $\mathbb{E}Y = (\mathbb{E}Y_1, \dots, \mathbb{E}Y_p)^T$ and the covariance matrix of Y is given by the $p \times p$ matrix whose (i,j)th entry equals the covariance between Y_i and Y_j. Note that the diagonal entries of $Cov(Y)$ are the variances of $Y_1, \dots, Y_p$. Also note that $Cov(Y)$ is a symmetric matrix.

We also looked at the following two important formulae in the last class:

1. $\mathbb{E}(AY + c) = A\mathbb{E}(Y) + c$ for every deterministic matrix A and every deterministic vector c.
2. $Cov(AY + c) = ACov(Y)A^T$ for every deterministic matrix A and every deterministic vector c.

Example 39.1 (White Noise)**.** *Random variables $Z_1, \dots, Z_p$ are said to form white noise if they have mean zero, variance one and if they are uncorrelated. Let Z be the random vector with components $Z_1, \dots, Z_p$. Then it is clear that the components of Z are white noise if and only if $\mathbb{E}Z = 0$ and $Cov(Z) = I_p$ (here I_p is the $p \times p$ identity matrix).*

As a consequence of the second formula above, we saw that

$$var(a^T Y) = a^T Cov(Y) a \qquad \text{for every } p \times 1 \text{ vector } a.$$

Because variance is always nonnegative, this means that $Cov(Y)$ satisfies the following property:

$$a^T Cov(Y) a = var(a^T Y) \geq 0 \qquad \text{for every } p \times 1 \text{ vector } a. \tag{62}$$

Now recall the following definition from linear algebra:

Definition 39.2. *Let Σ denote a $p \times p$ symmetric matrix.*

1. *Σ is said to be **positive semi-definite** if $a^T \Sigma a \geq 0$ for every $a \in \mathbb{R}^p$.*
2. *Σ is said to be **positive definite** if $a^T \Sigma a > 0$ for every $a \in \mathbb{R}^p$ with $a \neq 0$.*

From this definition and the fact (62), it follows that **the covariance matrix $Cov(Y)$ of every random vector Y is symmetric and positive semi-definite**.

However $Cov(Y)$ is not necessarily positive definite. To see this, just take $p = 2$ and $Y = (Y_1, -Y_1)^T$ for a random variable Y_1. Then with $a = (1, 1)$, it is easy to see that $a^T Cov(Y) a = Var(a^T Y) = Var(Y_1 + Y_2) = 0$.

But if $Cov(Y)$ is not positive definite, then there exists $a \neq 0$ such that $Var(a^T Y) = a^T Cov(Y) a = 0$. This must necessarily mean that $a^T(Y - \mu) = 0$ where $\mu = \mathbb{E}(Y)$. In other words, the random variables $Y_1, \dots, Y_n$ have to satisfy a linear equation. We can therefore say that: **$Cov(Y)$ is positive definite if and only if the random variables $Y_1, \dots, Y_n$ do not satisfy a linear equation**.

40 Detour – Spectral Theorem for Symmetric Matrices

Since $Cov(Y)$ is a symmetric and positive semi-definite matrix, some standard facts about such matrices are useful while working with covariance matrices. In particular, we shall make some use of the spectral theorem for symmetric matrices. Before looking at the spectral theorem, we need to recall the notion of an orthonormal basis.

40.1 Orthonormal Basis

Definition 40.1 (Orthonormal Basis)**.** *An orthonormal basis in $\mathbb{R}^p$ is a set of p vectors $u_1, \dots, u_p$ in $\mathbb{R}^p$ having the following properties:*

1. *$u_1, \dots, u_p$ are orthogonal i.e., $\langle u_i, u_j \rangle := u_i^T u_j = 0$ for $i \neq j$.*
2. *Each u_i has unit length i.e., $\|u_i\| = 1$ for each i.*

The simplest example of an orthonormal basis is $e_1, \dots, e_n$ where e_i is the vector that 1 in the i^{th} position and 0 at all other positions.

Every orthonormal basis $u_1, \dots, u_p$ satisfies the following properties:

1. $u_1, \dots, u_p$ are linearly independent and therefore form a basis of $\mathbb{R}^p$ (this explains the presence of the word "basis" in the definition of orthonormal basis).

   To see why this is true, suppose that

   $$\alpha_1 u_1 + \cdots + \alpha_p u_p = 0 \qquad \text{for some } \alpha_1, \dots, \alpha_p. \tag{63}$$

   Taking the dot product of both sides of the above equality with u_j (for a fixed j), we get

   $$0 = \left\langle u_j, \sum_{i=1}^{p} \alpha_i u_i \right\rangle = \sum_{i=1}^{p} \alpha_i \langle u_j, u_i \rangle = \alpha_j$$

because $\langle u_j, u_i \rangle$ is non-zero only when $i = j$ and $\langle u_j, u_j \rangle = 1$. Thus (63) implies that $\alpha_j = 0$ for every $j = 1, \dots, p$ and thus $u_1, \dots, u_p$ are linearly independent and consequently form a basis of $\mathbb{R}^p$.

2. The following formula holds for every vector $x \in \mathbb{R}^p$:

$$x = \sum_{i=1}^{p} \langle x, u_i \rangle \, u_i. \tag{64}$$

To see why this is true, note first that the previous property implies that $u_1, \dots, u_p$ form a basis of $\mathbb{R}^p$ so that every $x \in \mathbb{R}^p$ can be written as a linear combination

$$x = \beta_1 u_1 + \cdots + \beta_p u_p$$

of $u_1, \dots, u_p$. Now take dot product with u_j on both sides to prove that $\beta_j = \langle x, u_j \rangle$.

3. The formula

$$u_1 u_1^T + \cdots + u_p u_p^T = I_p \tag{65}$$

holds where I_p is the $p \times p$ identity matrix. To see why this is true, note that (64) can be rewritten as

$$x = \sum_{i=1}^{p} u_i \langle x, u_i \rangle = \sum_{i=1}^{p} u_i u_i^T x = \left(\sum_{i=1}^{p} u_i u_i^T \right) x.$$

Since both sides of the above identity are equal for every x, we must have (65).

4. Suppose U is the $p \times p$ matrix whose columns are the vectors $u_1, \dots, u_p$. Then

$$U^T U = U U^T = I_p.$$

To see why this is true, note that the (i, j)th entry of $U^T U$ equals $u_i^T u_j$ which (by definition of orthonormal basis) is zero when $i \neq j$ and 1 otherwise. On the other hand, the statement $UU^T = I_p$ is the same as (65).

5. For every vector $x \in \mathbb{R}^p$, the formula

$$\|x\|^2 = \sum_{i=1}^{p} \langle x, u_i \rangle^2$$

holds. To see why this is true, just write

$$\|x\|^2 = x^T x = x^T U U^T x = \|U^T x\|^2 = \sum_{i=1}^{p} (u_i^T x)^2 = \sum_{i=1}^{p} \langle x, u_i \rangle^2.$$

40.2 Spectral Theorem

Theorem 40.2 (Spectral Theorem)**.** *Suppose Σ is a $p \times p$ symmetric matrix. Then there exists an orthonormal basis $u_1, \dots, u_p$ and real numbers $\lambda_1, \dots, \lambda_p$ such that*

$$\Sigma = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_p u_p u_p^T. \tag{66}$$

The spectral theorem is also usually written in the following alternative form. Suppose U is the $p \times p$ matrix whose columns are the vectors $u_1, \dots, u_p$. Also suppose that Λ is the $p \times p$ diagonal matrix (a diagonal matrix is a matrix whose off-diagonal entries are all zero) whose diagonal entries are $\lambda_1, \dots, \lambda_p$. Then (66) is equivalent to

$$\Sigma = U \Lambda U^T \text{ and } U^T \Sigma U = \Lambda.$$

Here are some straightforward consequences of the spectral theorem:

1. For every $1 \leq j \leq p$, we have the identities
$$\Sigma u_j = \lambda_j u_j \quad \text{and} \quad u_j^T \Sigma u_j = \lambda_j.$$
These follow directly from (66). The first identity above implies that each λ_j is an eigenvalue of Σ with eigenvector u_j.

2. In the representation (66), the eigenvalues $\lambda_1, \dots, \lambda_p$ are unique while the eigenvectors $u_1, \dots, u_p$ are not necessarily unique (for every u_j can be replaced by $-u_j$ and if $\lambda_1 = \lambda_2$, then u_1 and u_2 can be replaced by any pair $\tilde{u}_1, \tilde{u}_2$ of orthogonal unit norm vectors in the span of u_1 and u_2).

3. The rank of Σ precisely equals the number of $\lambda_j' s$ that are non-zero.

4. If all of $\lambda_1, \dots, \lambda_p$ are non-zero, then Σ has full rank and is hence invertible. Moreover, we can then write
$$\Sigma^{-1} = \lambda_1^{-1} u_1 u_1^T + \lambda_2^{-1} u_2 u_2^T + \dots + \lambda_p^{-1} u_p u_p^T.$$

5. If Σ is positive semi-definite, then every λ_j in (66) is nonnegative (this is a consequence of $\lambda_j = u_j^T \Sigma u_j \geq 0$).

6. **Square Root of a Positive Semi-definite Matrix**: If Σ is positive semi-definite, then we can define a new matrix
$$\Sigma^{1/2} := \sqrt{\lambda_1} u_1 u_1^T + \dots + \sqrt{\lambda_p} u_p u_p^T.$$
It is easy to see that $\Sigma^{1/2}$ is symmetric, positive semi-definite and satisfies $(\Sigma^{1/2})(\Sigma^{1/2}) = \Sigma$. We shall refer to $\Sigma^{1/2}$ as the square root of Σ.

7. If Σ is positive definite, then every λ_j in (66) is strictly positive (this is a consequence of $\lambda_j = u_j^T \Sigma u_j > 0$).

40.3 Three Applications of the Spectral Theorem

40.3.1 Every symmetric positive semi-definite matrix is a Covariance Matrix

We have seen previously that the covariance matrix $Cov(Y)$ of every random vector Y is symmetric and positive semi-definite. It turns out that the converse of this statement is also true i.e., it is also true that every symmetric and positive semi-definite matrix equals $Cov(Y)$ for some random vector Y. To see why this is true, suppose that Σ is an arbitrary $p \times p$ symmetric and positive semi-definite matrix. Recall that, via the spectral theorem, we have defined $\Sigma^{1/2}$ (square-root of Σ) which is a symmetric and nonnegative definite matrix such that $\Sigma^{1/2}\Sigma^{1/2} = \Sigma$.

Now suppose that $Z_1, \dots Z_p$ are uncorrelated random variables all having unit variance and let $Z = (Z_1, \dots, Z_p)^T$ be the corresponding random vector. Because $Z_1, \dots, Z_p$ are uncorrelated and have unit variance, it is easy to see that $Cov(Z) = I_p$. Suppose now that $Y = \Sigma^{1/2} Z$. Then
$$Cov(Y) = Cov(\Sigma^{1/2} Z) = \Sigma^{1/2} Cov(Z) (\Sigma^{1/2})^T = \Sigma^{1/2} (I_n) \Sigma^{1/2} = \Sigma.$$
We have thus started with an arbitrary positive semi-definite matrix Σ and proved that it equals $Cov(Y)$ for some random vector Y.

We can thus summarize the following properties of a covariance matrix.

1. The covariance matrix of every random vector is positive semi-definite definite.

2. Every positive semi-definite matrix equals the covariance matrix of some random vector.

3. Unless the random variables $Y_1, \dots, Y_n$ satisfy an exact linear equation, their covariance matrix is positive definite.

40.3.2 Whitening

Given a $p \times 1$ random vector Y, how can we transform it into a $p \times 1$ white noise vector Z (recall that Z is white noise means that $\mathbb{E}Z = 0$ and $Cov(Z) = I_p$). This transformation is known as Whitening. Whitening can be done if $Cov(Y)$ is positive definite. Indeed suppose that $\Sigma := Cov(Y)$ is positive definite with spectral representation:

$$\Sigma = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_p u_p u_p^T.$$

The fact that Σ is positive definite implies that $\lambda_i > 0$ for every $i = 1, \dots, p$. In that case, it is easy to see that $\Sigma^{1/2}$ is invertible and

$$\Sigma^{-1/2} := (\Sigma^{1/2})^{-1} = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_p u_p u_p^T.$$

Moreover it is easy to check that $\Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_p$. Using this, it is straightforward to check that $Z = \Sigma^{-1/2}(Y - \mathbb{E}Y)$ is white noise. Indeed $\mathbb{E}Z = 0$ and

$$Cov(Z) = Cov\left(\Sigma^{-1/2}(Y - \mathbb{E}Y)\right) = \Sigma^{-1/2} Cov(Y) \Sigma^{-1/2} = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = I_p.$$

Therefore the spectral theorem is used to define the matrix $(Cov(Y))^{-1/2}$ which can be used to whiten the given random vector Y.

40.3.3 First Prinicipal Component of a Random Vector

Let Y be a $p \times 1$ vector. We say that a unit vector $a \in \mathbb{R}^p$ (unit vectors are vectors with norm equal to one) is a **first principal component** of Y if

$$var(a^T Y) \geq var(b^T Y) \qquad \text{for every unit vector } b.$$

In other words, the unit vector a maximizes the variance of $b^T Y$ over all unit vectors b.

Suppose that $\Sigma := Cov(Y)$ has the spectral representation (66). Assume, without loss of generality, that the eigenvalues $\lambda_1, \dots, \lambda_p$ appearing in (66) are arranged in decreasing order i.e.,

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0.$$

It then turns out that the vector u_1 is a first principal component of Y. To see this, simply note that

$$var(u_1^T Y) = u_1^T Cov(Y) u_1 = u_1^T \left(\lambda_1 u_1 u_1^T + \cdots + \lambda_p u_p u_p^T\right) u_1 = \lambda_1$$

and that for every unit vector b,

$$var(b^T Y) = b^T Cov(Y) b = b^T \left(\lambda_1 u_1 u_1^T + \cdots + \lambda_p u_p u_p^T\right) b = \sum_{i=1}^p \lambda_i \left\langle b, u_i \right\rangle^2 \leq \lambda_1 \sum_{i=1}^p \left\langle b, u_i \right\rangle^2 = \lambda_1 \|b\|^2 = \lambda_1.$$

Thus u_1 is a first principal component of Y. Note that first principal components are not unique. Indeed, $-u_1$ is also a first principal component and if $\lambda_1 = \lambda_2$, then u_2 and $(u_1 + u_2)/\sqrt{2}$ are also first principal components.

41 Best Linear Predictor

Consider random variables $Y, X_1, \dots, X_p$ that have finite variance. We want to predict Y on the basis of $X_1, \dots, X_p$. Given a predictor $g(X_1, \dots, X_p)$ of Y based on $X_1, \dots, X_p$, we measure the accuracy of prediction by

$$R(g) := \mathbb{E}\left(Y - g(X_1, \dots, X_p)\right)^2.$$

We have seen in the last class that the best predictor (i.e., the function g^* which minimizes $R(g)$) is given by the conditional expectation:

$$g^*(x_1, \dots, x_p) := \mathbb{E}\left(Y | X_1 = x_1, \dots, X_p = x_p\right).$$

This conditional expectation is often quite a complicated quantity. For example, in practice to estimate it, one would need quite a lot of data on the variables $X_1, \dots, X_p, Y$.

We now consider a related problem of predicting Y based only on **linear** functions of $X_1, \dots, X_p$. Specifically, we consider predictions of the form $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = \beta_0 + \beta^T X$ (where $\beta := (\beta_1, \dots, \beta_p)^T$ and $X = (X_1, \dots, X_p)^T$). The Best Linear Predictor (BLP) of Y in terms of $X_1, \dots, X_p$ is the linear function

$$\beta_0^* + \beta_1^* X_1 + \dots + \beta_p^* X_p = \beta_0^* + (\beta^*)^T X \qquad \text{with } \beta^* := (\beta_1^*, \dots, \beta_p^*)^T$$

where $\beta_0^*, \dots, \beta_p^*$ minimize

$$L(\beta_0, \dots, \beta_p) = \mathbb{E}\left(Y - \beta_0 - \beta_1 X_1 - \dots - \beta_p X_p\right)^2$$

over $\beta_0, \beta_1, \dots, \beta_p$.

One can get an explicit formula for β_0^* and β^* by minimizing L directly via calculus. Taking partial derivatives with respect to $\beta_0, \beta_1, \dots, \beta_p$ and setting them equal to zero, we obtain the following equations:

$$\mathbb{E}(Y - \beta_0^* - \beta_1^* X_1 - \dots - \beta_p^* X_p) = 0 \tag{67}$$

and

$$\mathbb{E}(Y - \beta_0^* - \beta_1^* X_1 - \dots - \beta_p^* X_p) X_i = 0 \qquad \text{for } i = 1, \dots, p. \tag{68}$$

The first equation above implies that $Y - \beta_0^* - \beta_1^* X_1 - \dots - \beta_p^* X_p$ is a mean zero random variable. Using this, we can rewrite the second equation as

$$Cov(Y - \beta_0^* - \beta_1^* X_1 - \dots - \beta_p^* X_p, X_i) = 0 \qquad \text{for } i = 1, \dots, p$$

which is same as

$$Cov(Y - \beta_1^* X_1 - \dots - \beta_p^* X_p, X_i) = 0 \qquad \text{for } i = 1, \dots, p. \tag{69}$$

Rearranging the above, we obtain

$$\sum_{j=1}^{p} \beta_i^* Cov(X_i, X_j) = Cov(Y, X_i) \qquad \text{for } i = 1, \dots, p.$$

In matrix notation, we can rewrite this as

$$Cov(X)\beta^* = Cov(X, Y) \qquad \text{with } \beta^* = (\beta_1^*, \dots, \beta_p^*)^T.$$

Here $Cov(X, Y)$ is the $p \times 1$ vector with entries $Cov(X_1, Y), \dots, Cov(X_p, Y)$. The above equation gives

$$\beta^* = (Cov(X))^{-1} Cov(X, Y)$$

assuming that $Cov(X)$ is invertible. This equation determines $\beta_1^*, \dots, \beta_p^*$. We can then use (72) to write β_0^* as

$$\beta_0^* = \mathbb{E}(Y) - Cov(Y, X)(Cov(X))^{-1}\mathbb{E}(X).$$

Note that the term $Cov(Y, X)$ appearing above is the transpose of $Cov(X, Y)$. More generally, given two random vectors $W = (W_1, \dots, W_p)$ and $Z = (Z_1, \dots, Z_q)$, we define $Cov(W, Z)$ to be the $p \times q$ matrix whose (i, j)th entry is the covariance between W_i and Z_j.

The Best Linear Predictor (BLP) of Y in terms of $X_1, \dots, X_p$ then equals

$$\begin{aligned} \beta_0^* + \beta_1^* X_1 + \dots \beta_p^* X_p &= \beta_0^* + (\beta^*)^T X \\ &= \mathbb{E}(Y) - Cov(Y, X)(Cov(X))^{-1}\mathbb{E}(X) + Cov(Y, X)(Cov(X))^{-1} X \\ &= \mathbb{E}(Y) + Cov(Y, X)(Cov(X))^{-1}(X - \mathbb{E}(X)). \end{aligned} \tag{70}$$

As mentioned previously, the Best Predictor (BP) of Y in terms of $X_1, \dots X_p$ is the function $g^*(X_1, \dots, X_p)$ of $X_1, \dots, X_p$ which minimizes

$$L(g) := \mathbb{E}(Y - g(X_1, \dots, X_p))^2$$

over all functions g and we have seen that

$$g^*(X_1, \dots, X_p) = \mathbb{E}(Y | X_1, \dots, X_p).$$

In other words, the best predictor is the conditional expectation. In general, the BP and BLP will differ and the BP will be a more accurate predictor of Y compared to BLP. Note that the BLP only depends on the variances and covariances between the random variables $Y, X_1, \dots, X_p$ while the BP depends potentially on the entire joint distribution. As a result, the BLP is usually much easier to estimate based on data compared to the BP.

In general, we shall refer to any quantity involving the distribution of $Y, X_1, \dots, X_p$ that depends only on the mean, variances and covariances of $Y, X_1, \dots, X_p$ as a second order property. Note that the BLP is a second order quantity while the BP is not.

42 Best Linear Predictor

In the last class, we looked at the problem of predicting a random variable Y based on **linear** functions of other random variables $X_1, \dots, X_p$. Specifically, we consider predictions of the form $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p = \beta_0 + \beta^T X$ (where $\beta := (\beta_1, \dots, \beta_p)^T$ and $X = (X_1, \dots, X_p)^T$). The Best Linear Predictor (BLP) of Y in terms of $X_1, \dots, X_p$ is the linear function

$$\beta_0^* + \beta_1^* X_1 + \dots + \beta_p^* X_p = \beta_0^* + (\beta^*)^T X \qquad \text{with } \beta^* := (\beta_1^*, \dots, \beta_p^*)^T$$

where $\beta_0^*, \dots, \beta_p^*$ minimize

$$L(\beta_0, \dots, \beta_p) = \mathbb{E}\left(Y - \beta_0 - \beta_1 X_1 - \dots - \beta_p X_p\right)^2 \tag{71}$$

over $\beta_0, \beta_1, \dots, \beta_p$.

Setting derivatives of (71) with respect to $\beta_0, \dots, \beta_p$ and setting them equal to zero, we observed that $\beta_0^*, \dots, \beta_p^*$ satisfy the equations:

$$\mathbb{E}(Y - \beta_0^* - \beta_1^* X_1 - \dots - \beta_p^* X_p) = 0 \tag{72}$$

and

$$Cov(Y - \beta_1^* X_1 - \dots - \beta_p^* X_p, X_i) = 0 \qquad \text{for } i = 1, \dots, p. \tag{73}$$

The equations in (73) can be written succinctly in matrix notation as:

$$Cov(X)\beta^* = Cov(X, Y) \qquad \text{with } \beta^* = (\beta_1^*, \dots, \beta_p^*)^T.$$

Here $Cov(X, Y)$ is the $p \times 1$ vector with entries $Cov(X_1, Y), \dots, Cov(X_p, Y)$. The above equation gives

$$\beta^* = (Cov(X))^{-1} Cov(X, Y).$$

This determines $\beta_1^*, \dots, \beta_p^*$. We can then use (72) to write β_0^* as

$$\beta_0^* = \mathbb{E}(Y) - Cov(Y, X)(Cov(X))^{-1}\mathbb{E}(X).$$

Note that the term $Cov(Y, X)$ appearing above is the transpose of $Cov(X, Y)$. More generally, given two random vectors $W = (W_1, \dots, W_p)$ and $Z = (Z_1, \dots, Z_q)$, we define $Cov(W, Z)$ to be the $p \times q$ matrix whose (i, j)th entry is the covariance between W_i and Z_j.

The Best Linear Predictor (BLP) of Y in terms of $X_1, \dots, X_p$ then equals

$$\begin{aligned}\beta_0^* + \beta_1^* X_1 + \dots \beta_p^* X_p &= \beta_0^* + (\beta^*)^T X \\ &= \mathbb{E}(Y) - Cov(Y, X)(Cov(X))^{-1}\mathbb{E}(X) + Cov(Y, X)(Cov(X))^{-1} X \\ &= \mathbb{E}(Y) + Cov(Y, X)(Cov(X))^{-1}(X - \mathbb{E}(X)). \end{aligned} \tag{74}$$

Here are some important properties of the BLP:

1. The BLP solves equations (72) and (73). These equations are called **normal equations**.
2. If $Cov(X)$ is invertible (equivalently, positive definite), then the BLP is uniquely given by (74).
3. $Y - BLP$ has mean zero (because of (72)) and $Y - BLP$ is uncorrelated with each $X_i, i = 1, \dots, p$ (because of (73)). In fact, this property characterizes the BLP (see next).
4. If $Cov(X)$ is invertible, then it is clear from the form of the normal equations that the BLP is the unique linear combination of $X_1, \dots, X_p$ such that $Y - BLP$ has mean zero and is uncorrelated with $X_1, \dots, X_p$.

Example 42.1 (The case $p = 1$)**.** *When $p = 1$, the random vector X has only element X_1 so that $Cov(X)$ is just equal to the number $Var(X_1)$. In that case, the BLP of Y in terms of X_1 is given by*

$$BLP = \mathbb{E}(Y) + \frac{Cov(Y, X_1)}{Var(X_1)} (X_1 - \mathbb{E}(X_1)).$$

In other words, when $p = 1$,

$$\beta_1^* = \frac{Cov(Y, X_1)}{Var(X_1)} = Corr(Y, X_1)\sqrt{\frac{Var(Y)}{Var(X_1)}} = \rho_{Y,X_1}\sqrt{\frac{Var(Y)}{Var(X_1)}}.$$

In the further special case when $Var(Y) = Var(X_1)$ and $\mathbb{E}(Y) = \mathbb{E}(X_1) = 0$, we have

$$\beta_1^* = \rho_{Y,X_1}$$

so that the BLP is simply given by $\rho_{Y,X_1} X_1$.

Example 42.2. *Suppose $X_1, X_2, Z_3, \dots, Z_n, Z_{n+1}$ are uncorrelated random variables and mean zero random variables. Define random variables $X_3, \dots, X_{n+1}$ as*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t \qquad \text{for } t = 3, \dots, n+1.$$

What is the BLP of X_{n+1} in terms of $X_1, \dots, X_n$ for $n \geq 2$?

By definition,

$$X_{n+1} = \phi_1 X_n + \phi_2 X_{n-1} + Z_{n+1}$$

which means that $X_{n+1} - \phi_1 X_n - \phi_2 X_{n-1} = Z_{n+1}$. It is now easy to see that each X_t depends only on $X_1, X_2, Z_3, \dots, Z_t$ for $t \geq 3$ which implies that Z_{n+1} is uncorrelated with all of $X_1, \dots, X_n$.

Therefore $\phi_1 X_n + \phi_2 X_{n-1}$ is a linear combination of $X_1, \dots, X_n$ such that $X_{n+1} - \phi_1 X_n - \phi_2 X_{n-1}$ is uncorrelated with each of $X_1, \dots, X_n$ (it also has mean zero). We deduce therefore that the BLP of X_{n+1} in terms of $X_1, \dots, X_n$ equals $\phi_1 X_n + \phi_2 X_{n-1}$.

As mentioned previously, the Best Predictor (BP) of Y in terms of $X_1, \dots X_p$ is the function $g^*(X_1, \dots, X_p)$ of $X_1, \dots, X_p$ which minimizes

$$L(g) := \mathbb{E}(Y - g(X_1, \dots, X_p))^2$$

over all functions g and we have seen that

$$g^*(X_1, \dots, X_p) = \mathbb{E}(Y|X_1, \dots, X_p).$$

In other words, the best predictor is the conditional expectation. In general, the BP and BLP will differ and the BP will be a more accurate predictor of Y compared to BLP. Note that the BLP only depends on the variances and covariances between the random variables $Y, X_1, \dots, X_p$ while the BP depends potentially on the entire joint distribution. As a result, the BLP is usually much easier to estimate based on data compared to the BP.

In general, we shall refer to any quantity involving the distribution of $Y, X_1, \dots, X_p$ that depends only on the mean, variances and covariances of $Y, X_1, \dots, X_p$ as a second order property. Note that the BLP is a second order quantity while the BP is not.

43 Residual

The residual of a random variable Y in terms of $X_1, \dots, X_p$ will be denoted by $r_{Y|X_1,\dots,X_p}$ and defined as the difference between Y and the BLP of Y in terms of $X_1, \dots, X_p$. In other words,

$$r_{Y|X_1,\dots,X_p} = Y - BLP.$$

Using the formula for the BLP, we can write down the following formula for the residual:

$$r_{Y|X_1,\dots,X_p} = Y - \mathbb{E}(Y) - Cov(Y, X)(CovX)^{-1}(X - \mathbb{E}(X)) \tag{75}$$

where X is the $p \times 1$ random vector with components $X_1, \dots, X_p$.

The residual has mean zero and is uncorrelated with each of $X_1, \dots, X_p$. This can be proved either directly from the formula (75) or from the properties of the BLP.

The variance of the residual can be calculated from the formula (75) as follows:

$$\begin{aligned} Var(r_{Y|X_1,\dots,X_p}) &= Var\left(Y - \mathbb{E}(Y) - Cov(Y, X)(CovX)^{-1}(X - \mathbb{E}(X))\right) \\ &= Var(Y) - 2Cov(Y, Cov(Y, X)(CovX)^{-1}X) + Var(Cov(Y, X)(CovX)^{-1}(X - \mathbb{E}(X))) \\ &= Var(Y) - 2Cov(Y, X)(CovX)^{-1}Cov(X, Y) + Cov(Y, X)(CovX)^{-1}Cov(X, Y) \\ &= Var(Y) - Cov(Y, X)(CovX)^{-1}Cov(X, Y). \end{aligned}$$

In other words, $Var(r_{Y|X_1,\dots,X_p})$ equals the **Schur complement** (recalled next) of $Var(Y)$ in the covariance matrix:

$$\begin{pmatrix} Cov(X) & Cov(X, Y) \\ Cov(Y, X) & Var(Y) \end{pmatrix}$$

of the $(n + 1) \times 1$ random vector $(X_1, \dots, X_p, Y)^T$.

Note that the residual is also a second order quantity.

44 Detour: Schur Complements

Consider an $n \times n$ matrix A that is partitioned into four blocks as

$$A = \begin{pmatrix} E & F \\ G & H \end{pmatrix}$$

where E is $p \times p$, F is $p \times q$, G is $q \times p$ and H is $q \times q$ (p and q are such that $p + q = n$).

We define

$$E^S := E - FH^{-1}G \quad \text{and} \quad H^S := H - GE^{-1}F$$

assuming that H^{-1} and E^{-1} exist. We shall refer to E^S and H^S as the *Schur complements* of E and H respectively (**Warning**: This is not standard terminology; it is more common to refer to E^S as the Schur complement of H and to H^S as the Schur complement of E. I find it more natural to think of E^S as the Schur complement of E and H^S as the Schur complement of H).

Note that both E and E^S are $p \times p$ while both H and H^S are $q \times q$.

Schur complements have many interesting properties such as:

1. $det(A) = det(E)det(H^S) = det(H)det(E^S)$.
2. If A is positive definite, then E, E^S, H, H^S are all positive definite.

and many others. Feel free to see the monograph titled *Schur Complements and Statistics* by Diane Ouellette for proofs and exposition of these facts (this is not really necessary for this course).

But one very important property of Schur Complements for our purpose is the fact that they arise naturally in inverses of partitioned matrices. A standard formula for the inverse of a partitioned matrix (see, for example, `https://en.wikipedia.org/wiki/Block_matrix#Block_matrix_inversion`) is

$$A^{-1} = \begin{pmatrix} (E^S)^{-1} & -E^{-1}F(H^S)^{-1} \\ -(H^S)^{-1}GE^{-1} & (H^S)^{-1} \end{pmatrix} \tag{76}$$

It must be noted from this formula that **the first (or $(1,1)^{th}$) block of A^{-1} equals the inverse of the Schur complement of the first block of A. Similarly, the last (or $(2,2)^{th}$) block of A^{-1} equals the inverse of the Schur complement of the last block of A.**

We shall use the expression (76) for the inverse of the partitioned matrix A but we will not see how to prove (76). You can find many proofs of this fact elsewhere (just google something like "inverse of partitioned matrices").

45 Partial Correlation

Given random variables Y_1, Y_2 and $X_1, \dots, X_p$, the partial correlation between Y_1 and Y_2 given $X_1, \dots, X_p$ is denoted by $\rho_{Y_1,Y_2|X_1,\dots,X_p}$ and defined as

$$\rho_{Y_1,Y_2|X_1,\dots,X_p} := Corr\left(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}\right).$$

In other words, $\rho_{Y_1,Y_2|X_1,\dots,X_p}$ is defined as the correlation between the residual of Y_1 given $X_1, \dots, X_p$ and the residual of Y_2 given $X_1, \dots, X_p$.

$\rho_{Y_1,Y_2|X_1,\dots,X_p}$ is also termed the partial correlation of Y_1 and Y_2 after controlling for $X_1, \dots, X_p$. Since residuals are second order quantities, it follows that the partial correlation is a second order quantity as well. We shall now see how to explicitly write the partial correlation in terms of the covariances of Y_1, Y_2 and X.

As

$$r_{Y_1|X_1,\dots,X_p} = Y_1 - \mathbb{E}(Y_1) - Cov(Y_1, X)(CovX)^{-1}(X - \mathbb{E}(X))$$

and

$$r_{Y_2|X_1,\dots,X_p} = Y_2 - \mathbb{E}(Y_2) - Cov(Y_2, X)(CovX)^{-1}(X - \mathbb{E}(X)),$$

it can be checked (left as an exercise) that

$$Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}) = Cov(Y_1,Y_2) - Cov(Y_1,X)(CovX)^{-1}Cov(X,Y_2).$$

This, along with the formula for the variance of the residuals from the previous subsections, gives the following formula for the partial correlation $\rho_{Y_1,Y_2|X_1,\dots,X_p}$:

$$\frac{Cov(Y_1,Y_2) - Cov(Y_1,X)(CovX)^{-1}Cov(X,Y_2)}{\sqrt{Var(Y_1) - Cov(Y_1,X)(CovX)^{-1}Cov(X,Y_1)}\sqrt{Var(Y_2) - Cov(Y_2,X)(CovX)^{-1}Cov(X,Y_2)}}.$$

When $p=1$ so that X equals the scalar random variable X_1, the above formula simplifies to (check this):

$$\rho_{Y_1,Y_2|X_1} = \frac{\rho_{Y_1,Y_2} - \rho_{Y_1,X_1}\rho_{Y_2,X_1}}{\sqrt{1-\rho^2_{Y_1,X_1}}\sqrt{1-\rho^2_{Y_2,X_1}}}.$$

It is instructive to put the variances of the residuals $r_{Y_1|X_1,\dots,X_p}$ and $r_{y_2|X_1,\dots,X_p}$ and their covariance in a matrix. Recall first that:

$$Var(r_{Y_1|X_1,\dots,X_p}) = Var(Y_1) - Cov(Y_1,X)(CovX)^{-1}Cov(X,Y_1),$$

$$Var(r_{Y_2|X_1,\dots,X_p}) = Var(Y_2) - Cov(Y_2,X)(CovX)^{-1}Cov(X,Y_2)$$

and

$$Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}) = Cov(Y_1,Y_2) - Cov(Y_1,X)(CovX)^{-1}Cov(X,Y_2).$$

Let $R_{Y_1,Y_2|X_1,\dots,X_p}$ denote the 2×1 random vector consisting of the residuals $r_{Y_1|X_1,\dots,X_p}$ and $r_{Y_2|X_1,\dots,X_p}$. The formulae for the variances and covariances of the residuals allows us then to write the 2×2 covariance matrix of $R_{Y_1,Y_2|X_1,\dots,X_p}$ as

$$\begin{aligned} Cov(R_{Y_1,Y_2|X_1,\dots,X_p}) &= Cov\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - \begin{pmatrix} Cov(Y_1,X) \\ Cov(Y_2,X) \end{pmatrix} (CovX)^{-1} \begin{pmatrix} Cov(X,Y_1) & Cov(X,Y_2) \end{pmatrix} \\ &= Cov(Y) - Cov(Y,X)(CovX)^{-1}Cov(X,Y) \end{aligned}$$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_p \end{pmatrix}.$$

The right hand side in the formula for $Cov(R_{Y_1,Y_2|X_1,\dots,X_p})$ equals precisely the Schur complement of $Cov(Y)$ in the matrix

$$\begin{pmatrix} Cov(X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y) \end{pmatrix} = Cov\begin{pmatrix} X \\ Y \end{pmatrix} =: \Sigma.$$

Thus if Σ denotes the covariance matrix of the $(p+2)\times 1$ random vector $(X_1,\dots,X_p,Y_1,Y_2)^T$, then $Cov(R_{Y_1,Y_2|X_1,\dots,X_p})$ equals precisely the Schur complement of $Cov(Y)$ in Σ. We shall come back to this fact in the next class and use it to describe an expression for the partial correlation $\rho_{Y_1,Y_2|X_1,\dots,X_p}$ involving Σ^{-1}.

46 Partial Correlation and Inverse Covariance

We defined partial correlation in the last lecture. Given random variables Y_1, Y_2 and $X_1,\dots,X_p$, the partial correlation between Y_1 and Y_2 given $X_1,\dots,X_p$ is denoted by $\rho_{Y_1,Y_2|X_1,\dots,X_p}$ and defined as

$$\rho_{Y_1,Y_2|X_1,\dots,X_p} := Corr\left(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}\right).$$

In other words, $\rho_{Y_1,Y_2|X_1,\dots,X_p}$ is defined as the correlation between the residual of Y_1 given $X_1, \dots, X_p$ and the residual of Y_2 given $X_1, \dots, X_p$.

Recall that the residuals $r_{Y_1|X_1,\dots,X_p}$ and $r_{Y_2|X_1,\dots,X_p}$ have the following expressions:

$$r_{Y_1|X_1,\dots,X_p} = Y_1 - \mathbb{E}(Y_1) - Cov(Y_1, X)(CovX)^{-1}(X - \mathbb{E}(X))$$

and

$$r_{Y_2|X_1,\dots,X_p} = Y_2 - \mathbb{E}(Y_2) - Cov(Y_2, X)(CovX)^{-1}(X - \mathbb{E}(X)),$$

In the last class, we computed the variances of $r_{Y_1|X_1,\dots,X_p}$ and $r_{Y_2|X_1,\dots,X_p}$ as well as the covariance between them. This gave us the formulae:

$$Var(r_{Y_1|X_1,\dots,X_p}) = Var(Y_1) - Cov(Y_1, X)(CovX)^{-1}Cov(X, Y_1),$$

$$Var(r_{Y_2|X_1,\dots,X_p}) = Var(Y_2) - Cov(Y_2, X)(CovX)^{-1}Cov(X, Y_2)$$

and

$$Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}) = Cov(Y_1, Y_2) - Cov(Y_1, X)(CovX)^{-1}Cov(X, Y_2).$$

We can put these expressions together to get the following formula for the partial correlation between Y_1 and Y_2 given $X_1, \dots, X_p$:

$$\frac{Cov(Y_1, Y_2) - Cov(Y_1, X)(CovX)^{-1}Cov(X, Y_2)}{\sqrt{Var(Y_1) - Cov(Y_1, X)(CovX)^{-1}Cov(X, Y_1)}\sqrt{Var(Y_2) - Cov(Y_2, X)(CovX)^{-1}Cov(X, Y_2)}}.$$

We shall now describe the connection between partial correlations and the inverse of the Covariance matrix. Let $R_{Y_1,Y_2|X_1,\dots,X_p}$ denote the 2×1 random vector consisting of the residuals $r_{Y_1|X_1,\dots,X_p}$ and $r_{Y_2|X_1,\dots,X_p}$. The formulae for the variances and covariances of the residuals allows us then to write the 2×2 covariance matrix of $R_{Y_1,Y_2|X_1,\dots,X_p}$ as

$$\begin{aligned} Cov(R_{Y_1,Y_2|X_1,\dots,X_p}) &= Cov \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} - \begin{pmatrix} Cov(Y_1, X) \\ Cov(Y_2, X) \end{pmatrix} (CovX)^{-1} \begin{pmatrix} Cov(X, Y_1) & Cov(X, Y_2) \end{pmatrix} \\ &= Cov(Y) - Cov(Y, X)(CovX)^{-1}Cov(X, Y) \end{aligned}$$

where

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} X_1 \\ X_2 \\ \cdot \\ \cdot \\ \cdot \\ X_p \end{pmatrix}.$$

The right hand side in the formula for $Cov(R_{Y_1,Y_2|X_1,\dots,X_p})$ equals precisely the Schur complement of $Cov(Y)$ in the matrix

$$\begin{pmatrix} Cov(X) & Cov(X, Y) \\ Cov(Y, X) & Cov(Y) \end{pmatrix} = Cov \begin{pmatrix} X \\ Y \end{pmatrix} =: \Sigma.$$

Thus if Σ denotes the covariance matrix of the $(p + 2) \times 1$ random vector $(X_1, \dots, X_p, Y_1, Y_2)^T$, then $Cov(R_{Y_1,Y_2|X_1,\dots,X_p})$ equals precisely the Schur complement of $Cov(Y)$ in Σ.

But we know if we invert Σ, then the last diagonal block (or the $(2, 2)^{th}$ block) of Σ^{-1} equals the inverse of the Schur complement of the $(2, 2)^{th}$ block of Σ. This and the above connection between Schur complement and the covariance of $R_{Y_1,Y_2|X_1,\dots,X_p}$ allows us to deduce that if

$$\Sigma^{-1} = \begin{pmatrix} (\Sigma^{-1})_{11} & (\Sigma^{-1})_{12} \\ (\Sigma^{-1})_{21} & (\Sigma^{-1})_{22}, \end{pmatrix}$$

then

$$(\Sigma^{-1})_{22} = \left(Cov(R_{Y_1,Y_2|X_1,\dots,X_p})\right)^{-1} = \begin{pmatrix} Var(r_{Y_1|X_1,\dots,X_p}) & Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}) \\ Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}) & Var(r_{Y_2|X_1,\dots,X_p}) \end{pmatrix}^{-1}$$

The usual formula for the inverse of a 2 × 2 matrix then gives

$$(\Sigma^{-1})_{22} = \frac{1}{D} \begin{pmatrix} Var(r_{Y_2|X_1,\dots,X_p}) & -Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}) \\ -Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p}) & Var(r_{Y_1|X_1,\dots,X_p}) \end{pmatrix}$$

where D is the determinant of $Cov(R_{Y_1,Y_2|X_1,\dots,X_p})$.

From here it follows that the partial correlation $\rho_{Y_1,Y_2|X_1,\dots,X_p}$ has the alternative expression:

$$\rho_{Y_1,Y_2|X_1,\dots,X_p} = \frac{Cov(r_{Y_1|X_1,\dots,X_p}, r_{Y_2|X_1,\dots,X_p})}{\sqrt{Var(r_{Y_1|X_1,\dots,X_p})Var(r_{Y_2|X_1,\dots,X_p})}} = \frac{-(\Sigma^{-1})(n-1,n)}{\sqrt{(\Sigma^{-1})(n-1,n-1)\Sigma^{-1}(n,n)}}.$$

This shows the connection between partial correlation and inverse covariance matrices.

More generally, if $Y_1, \dots, Y_n$ are random variables (no distributional assumptions are needed here) with covariance matrix given by Σ. Then the partial correlation between Y_i and Y_j given $Y_k, k \neq i, k \neq j$ can be written in terms of Σ^{-1} as

$$\rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j} = \frac{-(\Sigma^{-1})(i,j)}{\sqrt{(\Sigma^{-1})(i,i)(\Sigma^{-1})(j,j)}}.$$

This implies, in particular, that

$$(\Sigma^{-1})(i,j) = 0 \iff \rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j} = 0$$

Therefore $(\Sigma^{-1})(i,j) = 0$ is equivalent to the partial correlation between Y_i and Y_j given $Y_k, k \neq i, k \neq j$ being zero.

Also

$$(\Sigma^{-1})(i,j) \leq 0 \iff \rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j} \geq 0 \quad \text{and} \quad (\Sigma^{-1})(i,j) \geq 0 \iff \rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j} \leq 0$$

In other words, $\Sigma^{-1}(i,j)$ being nonpositive is equivalent to the partial correlation between Y_i and Y_j given $Y_k, k \neq i, k \neq j$ being nonnegative. Similarly, $\Sigma^{-1}(i,j)$ being nonnegative is equivalent to the partial correlation between Y_i and Y_j given $Y_k, k \neq i, k \neq j$ being nonpositive.

47 Partial Correlation and Best Linear Predictor

Consider random variables Y and $X_1, \dots, X_p$. Let $\beta_0^* + \beta_1^* X_1 + \dots + \beta_p^* X_p$ denote the BLP of Y in terms of $X_1, \dots, X_p$.

We have seen before that If $p = 1$, then X is equal to the scalar random variable X_1 and the BLP then has the expression:

$$BLP = \mathbb{E}(Y) + \frac{Cov(Y, X_1)}{Var(X_1)}(X_1 - \mathbb{E}(X_1)).$$

In other words, when $p = 1$, the slope coefficient of the BLP is given by

$$\beta_1^* = \frac{Cov(Y, X_1)}{Var(X_1)} = \rho_{Y,X_1}\sqrt{\frac{Var(Y)}{Var(X_1)}}. \tag{77}$$

When $p \geq 1$, we would have p "slope" coefficients $X_1, \dots, X_p$. In this case, one can write a formula analogous to (77) as follows:

$$\beta_i^* = \rho_{Y,X_i|X_k,k\neq i}\sqrt{\frac{Var(r_{Y|X_k,k\neq i})}{Var(r_{X_i|X_k,k\neq i})}} \tag{78}$$

In other words β_i^* equals the slope coefficient of BLP of $r_{Y|X_k,k\neq i}$ in terms of $r_{X_i|X_k,k\neq i}$.

We shall prove this fact now. We can assume without loss of generality $i = p$. The proof for other i can be completed by rearranging $X_1, \dots, X_p$ so that X_i appears at the last position. The formula for $\beta^* = (\beta_1, \dots, \beta_p)^*$ is

$$\beta^* = (CovX)^{-1}Cov(X, Y).$$

Let us write

$$X = \begin{pmatrix} X_{-p} \\ X_p \end{pmatrix}$$

where $X_{-p} := (X_1, \dots, X_{p-1})^T$ consists of all the X's except X_p. We can partition $Cov(X)$ as

$$Cov(X) = Cov\begin{pmatrix} X_{-p} \\ X_p \end{pmatrix} = \begin{pmatrix} Cov(X_{-p}) & Cov(X_{-p}, X_p) \\ Cov(X_p, X_{-p}) & Var(X_p) \end{pmatrix}.$$

The formula for β^* then becomes

$$\beta^* = (CovX)^{-1}Cov(X, Y) = \begin{pmatrix} Cov(X_{-p}) & Cov(X_{-p}, X_p) \\ Cov(X_p, X_{-p}) & Var(X_p) \end{pmatrix}^{-1}\begin{pmatrix} Cov(X_{-p}, Y) \\ Cov(X_p, Y) \end{pmatrix}$$

In order to derive an explicit formula for β_p^* from this expression, we need to figure out the last row of $(CovX)^{-1}$. A standard formula for the inverses of partitioned matrices states that

$$A = \begin{pmatrix} E & F \\ G & H \end{pmatrix} \implies A^{-1} = \begin{pmatrix} \text{something} & \text{something} \\ -(H^S)^{-1}GE^{-1} & (H^S)^{-1} \end{pmatrix}$$

where $H^S := H - GE^{-1}F$ is the Schur complement of H in A. We shall apply this formula to

$$E = Cov(X_{-p}), \quad F = Cov(X_{-p}, X_p), \quad G = Cov(X_p, X_{-p}), \quad \text{and } H = Var(X_p)$$

so that A equals $Cov(X)$. In this case,

$$H^S = Var(X_p) - Cov(X_p, X_{-p})(Cov(X_{-p}))^{-1}Cov(X_{-p}, X_p) = Var(r_{X_p|X_k,k\neq p})$$

so that

$$(H^S)^{-1} = \frac{1}{Var(r_{X_p|X_k,k\neq p})}.$$

We thus obtain

$$\begin{aligned}\beta_p^* &= -(H^S)^{-1}GE^{-1}Cov(X_{-p}, Y) + (H^S)^{-1}Cov(X_p, Y) \\ &= \frac{Cov(X_p, Y) - Cov(X_p, X_{-p})(CovX_{-p})^{-1}Cov(X_{-p}, Y)}{Var(r_{X_p|X_k,k\neq p})} \\ &= \frac{Cov(r_{Y|X_k,k\neq p}, r_{X_p|X_k,k\neq p})}{Var(r_{X_p|X_k,k\neq p})} = \rho_{Y,X_p|X_k,k\neq p}\sqrt{\frac{Var(r_{Y|X_k,k\neq p})}{Var(r_{X_p|X_k,k\neq p})}}.\end{aligned}$$

which proves the result for $i = p$. One can prove it for other i by simply rearranging $X_1, \dots, X_p$ so that X_i appears as the last variable.

An important consequence of (78) is:

$$\beta_i^* = 0 \iff \rho_{Y,X_i|X_k,k\neq i} = 0 \tag{79}$$

In other words, the coefficient of X_i in the BLP of Y based on $X_1, \dots, X_p$ equals zero if and only if the partial correlation between Y and X_i given $X_k, k \neq i$ equals 0.

Example 47.1. *Suppose $X_1, X_2, Z_3, \dots, Z_n, Z_{n+1}$ with $n \geq 2$ are uncorrelated random variables having mean zero. Define random variables $X_3, \dots, X_{n+1}$ as*

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + Z_t \qquad \text{for } t = 3, \dots, n+1.$$

We have seen in the last class that the BLP of X_{n+1} in terms of $X_1, \dots, X_n$ equals $\phi_1 X_n + \phi_2 X_{n-1}$. This means that the coefficient of X_i in the BLP of X_{n+1} in terms of $X_1, \dots, X_n$ equals 0 for $i = 1, \dots, n-2$. As a consequence of (79), we then deduce that

$$\rho_{X_{n+1}, X_i | X_k, k \neq i, 1 \leq k \leq n} = 0 \qquad \text{for } i = 1, \dots, n-2.$$

Using the connection between partial correlation and inverse covariance, we can further deduce that if Σ denotes the $(n+1) \times (n+1)$ covariance matrix of $X_1, \dots, X_{n+1}$, then

$$\Sigma^{-1}(i, n+1) = 0 \qquad \text{for } i = 1, \dots, n-2.$$

48 BLP when Y is a random vector

Let us first quickly recap the BLP. Given random variables Y and $X_1, \dots, X_p$, a linear predictor of Y in terms of $X_1, \dots, X_p$ is a random variable of the form $\beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$. The BLP is then given by $\beta_0^* + \beta_1^* X_1 + \dots + \beta_p^* X_p$ where $\beta_0^*, \dots, \beta_p^*$ minimize:

$$L(\beta_0, \beta_1, \dots, \beta_p) := \mathbb{E} \left(Y - \beta_0 - \beta_1 X_1 - \dots - \beta_p X_p \right)^2$$

over $\beta_0, \dots, \beta_p$. We have seen that $\beta_0^*, \dots, \beta_p^*$ can be figured out using calculus and this gives the formula:

$$BLP = \mathbb{E}Y + Cov(Y, X)(Cov X)^{-1}(X - \mathbb{E}X)$$

where X stands for the $p \times 1$ random vector with components $X_1, \dots, X_p$. The residual $r_{Y|X_1,\dots,X_p}$ simply equals $Y - BLP$ and we have seen that the variance of $r_{Y|X_1,\dots,X_p}$ equals:

$$var(r_{Y|X_1,\dots,X_p}) = var(Y) - Cov(Y, X)(Cov X)^{-1} Cov(X, Y).$$

Note that this is the Schur complement of $var(Y)$ in $Cov \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} Cov(X) & Cov(X, Y) \\ Cov(Y, X) & var(Y) \end{pmatrix}$

Now suppose that we have two random variables Y_1 and Y_2 with Y denoting the 2×1 random vector with components Y_1 and Y_2. Consider the problem of finding the BLP of Y in terms of X (where X, as before, is the $p \times 1$ random vector with components $X_1, \dots, X_p$). To formalize this, we first need to define what a linear predictor is (note that Y is a 2×1 random vector and not a scalar random variable). A linear predictor for Y in terms of X is given by

$$AX + c$$

where A is a $2 \times p$ matrix and c is a 2×1 vector. The accuracy of this linear predictor for predicting Y can be measured by

$$L(A, c) := \mathbb{E} \left\| Y - AX - c \right\|^2.$$

The BLP is then given by $A^* Y + c^*$ where A^* and c^* minimize $L(A, c)$ over all A and c. To solve this minimization, let us first write A and c as

$$A := \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1p} \\ a_{21} & a_{22} & \dots & a_{2p} \end{pmatrix} \quad \text{and} \quad c := \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

so that

$$AX + c = \begin{pmatrix} a_{11} X_1 + a_{12} X_2 + \dots + a_{1p} X_p + c_1 \\ a_{21} X_1 + a_{22} X_2 + \dots + a_{2p} X_p + c_2 \end{pmatrix}$$

and

$$L(A,c) = \mathbb{E}\left\|Y - AX - c\right\|^2 = \mathbb{E}\left(Y_1 - a_{11}X_1 - a_{12}X_2 - \cdots - a_{1p}X_p - c_1\right)^2 + \mathbb{E}\left(Y_2 - a_{21}X_1 - a_{22}X_2 - \cdots - a_{2p}X_p - c_2\right)^2$$

From here it is clear that to minimize the above with respect to A and c, we can minimize the first term over $a_{11}, a_{12}, \dots, a_{1p}, c_1$ and then minimize the second term over $a_{21}, a_{22}, \dots, a_{2p}, c_2$. From here, it is easy to see that the BLP of $Y = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ in terms of X is given by

$$\begin{aligned}\begin{pmatrix} BLP \text{ of } Y_1 \text{ in terms of } X_1, \dots, X_p \\ BLP \text{ of } Y_2 \text{ in terms of } X_1, \dots, X_p \end{pmatrix} &= \begin{pmatrix} \mathbb{E}Y_1 + Cov(Y_1, X)(CovX)^{-1}(X - \mathbb{E}X) \\ \mathbb{E}Y_2 + Cov(Y_2, X)(CovX)^{-1}(X - \mathbb{E}X) \end{pmatrix} \\ &= \begin{pmatrix} \mathbb{E}Y_1 \\ \mathbb{E}Y_2 \end{pmatrix} + \begin{pmatrix} Cov(Y_1, X) \\ Cov(Y_2, X) \end{pmatrix} (CovX)^{-1}(X - \mathbb{E}X) \\ &= \mathbb{E}Y + Cov(Y, X)(CovX)^{-1}(X - \mathbb{E}X).\end{aligned}$$

Thus the same formula $\mathbb{E}Y + Cov(Y, X)(CovX)^{-1}(X - \mathbb{E}X)$ gives the BLP for Y in terms of X even when Y is a 2×1 random vector. It is straightforward now to see that this holds when Y is a $k \times 1$ random vector for every $k \geq 1$ (not just $k = 1$ or $k = 2$). One can define the residual of Y in terms of $X_1, \dots, X_p$ as

$$R_{Y|X} := Y - \mathbb{E}Y - Cov(Y, X)(CovX)^{-1}(X - \mathbb{E}X)$$

and this is exactly the vector whose i^{th} component is the residual of Y_i in terms of $X_1, \dots, X_p$. It is also straightforward to check that that covariance matrix of $R_{Y|X}$ is given by

$$Cov(R_{Y|X}) = Cov(Y) - Cov(Y, X)(CovX)^{-1}Cov(X, Y)$$

which is exactly the Schur complement of $Cov(Y)$ in the matrix $\begin{pmatrix} Cov(X) & Cov(X, Y) \\ Cov(Y, X) & Cov(Y) \end{pmatrix}$

49 Moment Generating Functions of Random Vectors

We shall next move to the last topic of the class: the multivariate normal distribution. For this, it is helpful to know about moment generating functions of random vectors.

The Moment Generating Function of an $n \times 1$ random vector Y is defined as

$$M_Y(a) := \mathbb{E}e^{a^T Y}$$

for every $a \in \mathbb{R}^n$ for which the expectation exists. Note that when $a = (0, \dots, 0)^T$ is the zero vector, it is easy to see that $M_Y(a) = 1$.

Just like in the univariate case, Moment Generating Functions determine distributions when they exist in a neighbourhood of $a = 0$.

Moment Generating Functions behave very nicely in the presence of independence. Suppose $Y_{(1)}$ and $Y_{(2)}$ are two random vectors and let $Y = (Y_{(1)}^T, Y_{(2)}^T)^T$ be the vector obtained by putting $Y_{(1)}$ and $Y_{(2)}$ together in a single column vector. Then **$Y_{(1)}$ and $Y_{(2)}$ are independent if and only if**

$$M_Y(a) = M_{Y_{(1)}}(a_{(1)}) M_{Y_{(2)}}(a_{(2)}) \qquad \text{for every } a = (a_{(1)}^T, a_{(2)}^T)^T \in \mathbb{R}^n.$$

Thus under independence, the MGF factorizes and conversely, when the MGF factorizes, we have independence.

50 The Multivariate Normal Distribution

The multivariate normal distribution is defined in the following way.

Definition 50.1. *A random vector $Y = (Y_1, \dots, Y_n)^T$ is said to have the multivariate normal distribution if every linear function $a^T Y$ of Y has the univariate normal distribution.*

Remark 50.1. *It is important to emphasize that for $Y = (Y_1, \dots, Y_n)^T$ to be multivariate normal,* ***every*** *linear function $a^T Y = a_1 Y_1 + \dots a_n Y_n$ needs to be univariate normal. It is not enough for example to just have each Y_i to be univariate normal. It is very easy to construct examples where each Y_i is univariate normal but $a_1 Y_1 + \dots + a_n Y_n$ is not univariate normal for many vectors $(a_1, \dots, a_n)^T$. For example, suppose that $Y_1 \sim N(0,1)$ and that $Y_2 = \xi Y_1$ where ξ is a discrete random variable taking the two values 1 and -1 with probability $1/2$ and ξ is independent of Y_1. Then it is easy to see that*

$$Y_2 | \xi = 1 \sim N(0,1) \quad \textit{and} \quad Y_2 | \xi = -1 \sim N(0,1).$$

This means therefore that $Y_2 \sim N(0,1)$ (and that Y_2 is independent of ξ). Note however that $Y_1 + Y_2$ is not normal as

$$\mathbb{P}\{Y_1 + Y_2 = 0\} = \mathbb{P}\{\xi = 1\} = \frac{1}{2}.$$

Thus, in this example, even though Y_1 and Y_2 are both $N(0,1)$, the vector $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ is not multivariate normal.

Example 50.2. *We have seen earlier in the class that if $Z_1, \dots, Z_n$ are* ***independent*** *and univariate normal, then $a_1 Z_1 + \dots a_n Z_n$ is normal for every $a_1, \dots, a_n$. Therefore a random vector $Z = (Z_1, \dots, Z_n)^T$ that is made up of* ***independent*** *Normal random variables has the multivariate normal distribution. In fact, we shall show below that if Y has a multivariate normal distribution, then it should necessarily be the case that Y is a linear function of a random vector Z that is made of independent univariate normal random variables.*

50.1 Moment Generating Function of a Multivariate Normal

Suppose $Y = (Y_1, \dots, Y_n)^T$ is multivariate normal. Let $\mu = \mathbb{E}(Y)$ and $\Sigma = Cov(Y)$ be the mean vector and covariance matrix of Y respectively. Then, as a direct consequence of the definition of multivariate normality, it follows that the MGF of Y equals

$$M_Y(a) = \mathbb{E}(e^{a^T Y}) = \exp\left(a^T \mu + \frac{1}{2} a^T \Sigma a\right). \tag{80}$$

To see why this is true, note that by definition of multivariate normality, $a^T Y$ is univariate normal. Now the mean and variance of $a^T Y$ are given by

$$\mathbb{E}(a^T Y) = a^T \mu \quad \text{and} \quad Var(a^T Y) = a^T Cov(Y) a = a^T \Sigma a$$

so that

$$a^T Y \sim N(a^T \mu, a^T \Sigma a) \qquad \text{for every } a \in \mathbb{R}^n.$$

Then (80) directly follows from the formula for the MGF of a univariate normal.

Note that the MGF of Y (given by (80)) only depends on the mean vector μ and the covariance matrix Σ of Y. Thus the distribution of every multivariate normal vector Y is characterized by the mean vector μ and covariance Σ. We therefore use the notation $N_n(\mu, \Sigma)$ for the multivariate normal distribution with mean μ and covariance Σ.

50.2 Connection to i.i.d $N(0,1)$ random variables

Suppose that the covariance matrix Σ of Y is positive definite so that $\Sigma^{-1/2}$ is well-defined. Let $Z := \Sigma^{-1/2}(Y-\mu)$. The formula (80) allows the computation of the MGF of Z as follows:

$$\begin{aligned}
M_Z(a) &= \mathbb{E}e^{a^TZ} \\
&= \mathbb{E}\exp\left(a^T\Sigma^{-1/2}(Y-\mu)\right) \\
&= \exp(a^T\Sigma^{-1/2}\mu)\mathbb{E}\exp\left(a^T\Sigma^{-1/2}Y\right) \\
&= \exp(a^T\Sigma^{-1/2}\mu)M_Y(\Sigma^{-1/2}a) \\
&= \exp(a^T\Sigma^{-1/2}\mu)\exp\left(a^T\Sigma^{-1/2}\mu + \frac{1}{2}(a^T\Sigma^{-1/2})\Sigma(\Sigma^{-1/2}a)\right) = \exp\left(\frac{1}{2}a^Ta\right) = \prod_{i=1}^n \exp(a_i^2/2).
\end{aligned}$$

The right hand side above is clearly the MGF of a random vector having n i.i.d standard normal random variables. Thus because MGFs uniquely determine distributions, we conclude that $Z = (Z_1, \dots, Z_n)^T$ has independent standard normal random variables. We have thus proved that: **If $Y \sim N_n(\mu, \Sigma)$ and Σ is p.d, then the components $Z_1, \dots Z_n$ of $Z = \Sigma^{-1/2}(Y-\mu)$ are independent standard normal random variables**.

51 Joint Density of the Multivariate Normal Distribution

Suppose $Y = (Y_1, \dots, Y_n)^T$ is a random vector that has the multivariate normal distribution. What then is the joint density of $Y_1, \dots, Y_n$?

Let $\mu = \mathbb{E}(Y)$ and $\Sigma = Cov(Y)$ be the mean vector and covariance matrix of Y respectively. For Y to have a joint density, we need to assume that Σ is positive definite. We have then seen in the previous section that the components $Z_1, \dots, Z_n$ of Z are independent standard normal random variables where

$$Z = \Sigma^{-1/2}(Y-\mu).$$

Because $Z_1, \dots, Z_n$ are independent standard normals, their joint density equals

$$f_{Z_1,\dots,Z_n}(z_1, \dots, z_n) = (2\pi)^{-n/2}\prod_{i=1}^n e^{-z_i^2/2} = (2\pi)^{-n/2}\exp\left(-\frac{1}{2}z^Tz\right)$$

where $z = (z_1, \dots, z_n)^T$.

Using the above formula and the fact that $Y = \mu + \Sigma^{1/2}Z$, we can compute the joint density of $Y_1, \dots, Y_n$ via the Jacobian formula. This gives

$$\begin{aligned}
f_{Y_1,\dots,Y_n}(y_1, \dots, y_n) &= f_{Z_1,\dots,Z_n}(\Sigma^{-1/2}(y-\mu))\det(\Sigma^{-1/2}) \\
&= \frac{1}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}\exp\left(\frac{-1}{2}(y-\mu)^T\Sigma^{-1}(y-\mu)\right)
\end{aligned}$$

where $y = (y_1, \dots, y_n)^T$.

52 Properties of Multivariate Normal Random Variables

Suppose $Y = (Y_1, \dots, Y_n)^T \sim N_n(\mu, \Sigma)$. Note then that μ is the mean vector $\mathbb{E}(Y)$ of Y and Σ is the covariance matrix $Cov(Y)$. The following properties are very important.

1. **Linear Functions of Y are also multivariate normal**: If A is an $m \times n$ deterministic matrix and c is an $m \times 1$ deterministic vector, then $AY + c \sim N_m(A\mu + c, A\Sigma A^T)$.

   Reason: Every linear function of $AY + c$ is obviously also a linear function of Y and, thus, this fact follows from the definition of the multivariate normal distribution.

2. If Y is multivariate normal, then every random vector formed by taking a subset of the components of Y is also multivariate normal.

   Reason: Follows from the previous fact.

3. **Independence is the same as Uncorrelatedness**: If $Y_{(1)}$ and $Y_{(2)}$ are two random vectors such that $Y = (Y_{(1)}^T, Y_{(2)}^T)^T$ is multivariate normal. Then $Y_{(1)}$ and $Y_{(2)}$ are independent if and only if $Cov(Y_{(1)}, Y_{(2)}) = 0$.

   Reason: The fact that independence implies $Cov(Y_{(1)}, Y_{(2)}) = 0$ is obvious and does not require any normality. The key is the other implication that zero covariance implies independence. For this, it is enough to show that the MGF of Y equals the product of the MGFs of $Y_{(1)}$ and $Y_{(2)}$. The MGF of Y equals

   $$M_Y(a) = \exp\left(a^T\mu + \frac{1}{2}a^T\Sigma a\right)$$

   where $\Sigma = Cov(Y)$.

   Note that $Y_{(1)}$ and $Y_{(2)}$ are also multivariate normal so that

   $$M_{Y_{(i)}}(a_{(i)}) = \exp\left(a_{(i)}^T\mu_{(i)} + \frac{1}{2}a_{(i)}^T\Sigma_{ii}a_{(i)}\right) \qquad \text{for } i = 1, 2$$

   where

   $$\mu_{(i)} := \mathbb{E}(Y_{(i)}) \quad \text{and} \quad \Sigma_{ii} := Cov(Y_{(i)}).$$

   Now if $\Sigma_{12} := Cov(Y_{(1)}, Y_{(2)})$ and $\Sigma_{21} = Cov(Y_{(2)}, Y_{(1)}) = \Sigma_{12}^T$, then observe that

   $$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix}$$

   As a result, if $a = (a_{(1)}, a_{(2)})^T$, then

   $$a^T\Sigma a = a_{(1)}^T\Sigma_{11}a_{(1)} + a_{(2)}^T\Sigma_{22}a_{(2)} + 2a_{(1)}^T\Sigma_{12}a_{(2)}.$$

   Under the assumption that $\Sigma_{12} = 0$, we can therefore write

   $$a^T\Sigma a = a_{(1)}^T\Sigma_{11}a_{(1)} + a_{(2)}^T\Sigma_{22}a_{(2)}$$

   from which it follows that

   $$M_Y(a) = M_{Y_{(1)}}(a_{(1)})M_{Y_{(2)}}(a_{(2)}).$$

   Because the MGF of $Y = (Y_{(1)}, Y_{(2)})^T$ factorizes into the product of the MGF of $Y_{(1)}$ and the MGF of $Y_{(2)}$, it follows that $Y_{(1)}$ and $Y_{(2)}$ are independent. Thus under the assumption of multivariate normality of $(Y_{(1)}, Y_{(2)})^T$, uncorrelatedness is the same as independence.

4. Suppose $Y = (Y_1, \dots, Y_n)^T$ is a multivariate normal random vector. Then two components Y_i and Y_j are independent if and only if $\Sigma_{ij} = 0$ where $\Sigma = Cov(Y)$.

   Reason: Follows directly from the previous three facts.

5. **Independence of linear functions can be checked by multiplying matrices**: Suppose Y is multivariate normal. Then AY and BY are independent if and only if $A\Sigma B^T = 0$.

   Reason: Note first that

   $$\begin{pmatrix} AY \\ BY \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix} Y$$

   is multivariate normal. Therefore AY and BY are independent if and only if $Cov(AY, BY) = 0$. The claimed assertion then follows from the observation that $Cov(AY, BY) = A\Sigma B^T$.

53 Properties of Multivariate Normal Random Variables

Suppose $Y = (Y_1, \dots, Y_n)^T \sim N_n(\mu, \Sigma)$. Note then that μ is the mean vector $\mathbb{E}(Y)$ of Y and Σ is the covariance matrix $Cov(Y)$. In the last class, we looked at the following properties.

1. **Linear Functions of Y are also multivariate normal**: If A is an $m \times n$ deterministic matrix and c is an $m \times 1$ deterministic vector, then $AY + c \sim N_m(A\mu + c, A\Sigma A^T)$.
2. Every random vector formed by taking a subset of the components of Y is also multivariate normal.
3. **Independence is the same as Uncorrelatedness**: If $Y_{(1)}$ and $Y_{(2)}$ are two sub-vectors obtained from Y, then $Y_{(1)}$ and $Y_{(2)}$ are independent if and only if $Cov(Y_{(1)}, Y_{(2)}) = 0$.
4. Two components Y_i and Y_j are independent if and only if $\Sigma_{ij} = 0$ where $\Sigma = Cov(Y)$.
5. **Independence of linear functions can be checked by multiplying matrices**: AY and BY are independent if and only if $A\Sigma B^T = 0$.

54 Idempotent Matrices and Chi-Squared distributions

We shall next prove that quadratic forms of multivariate normal random variables with identity covariance have chi-squared distributions provided the symmetric matrix defining the quadratic form is **idempotent**. A square matrix A is said to be idempotent if $A^2 = A$. An important fact about idempotent matrices is the following.

Fact: If A is an $n \times n$ symmetric and idempotent matrix of rank r if and only if

$$A = u_1 u_1^T + \cdots + u_r u_r^T \tag{81}$$

for r orthogonal and unit length vectors $u_1, \dots, u_r$.

To prove this fact, note first that if A is symmetric, then by the spectral theorem

$$A = \lambda_1 u_1 u_1^T + \lambda_2 u_2 u_2^T + \cdots + \lambda_n u_n u_n^T$$

for an orthonormal basis $u_1, \dots, u_n$ and real numbers $\lambda_1, \dots, \lambda_n$. The rank of A precisely equals the number of λ_i's that are non-zero. If r is the rank of A, we can therefore write (assuming without loss of generality that $\lambda_1, \dots, \lambda_r$ are non-zero and $\lambda_{r+1} = \cdots = \lambda_n = 0$)

$$A = \lambda_1 u_1 u_1^T + \cdots + \lambda_r u_r u_r^T.$$

It then follows that

$$A^2 = \lambda_1^2 u_1 u_1^T + \cdots + \lambda_r^2 u_r u_r^T.$$

Therefore if A is idempotent, then $A^2 = A$ so that

$$\lambda_1 u_1 u_1^T + \cdots + \lambda_r u_r u_r^T = \lambda_1^2 u_1 u_1^T + \cdots + \lambda_r^2 u_r u_r^T$$

which implies that $\lambda_i^2 = \lambda_i$ which gives $\lambda_i = 1$ (note that we have assumed that $\lambda_i \neq 0$). This proves (81).

The following result states that quadratic forms of multivariate normal random vectors with identity covariance are chi-squared provided the underlying matrix is idempotent.

Theorem 54.1. *Suppose $Y \sim N_n(\mu, I_n)$ and let A is an $n \times n$ symmetric and idempotent matrix with rank r. Then*

$$(Y - \mu)^T A (Y - \mu) \sim \chi_r^2.$$

Proof. Because A is symmetric and idempotent and has rank r, we can write A as

$$A = u_1 u_1^T + \cdots + u_r u_r^T$$

for some orthogonal and unit norm vectors $u_1, \dots, u_r$. Then

$$(Y-\mu)^T A (Y-\mu) = \sum_{i=1}^r \left(u_i^T (Y-\mu)\right)^2 = \sum_{i=1}^r V_i^2.$$

where $V_i := u_i^T(Y-\mu)$. Note now that each $V_i \sim N(0,1)$ and that $V_1, \dots, V_r$ are uncorrelated and hence independent (because of normality). This proves that $(Y-\mu)^T A (Y-\mu) \sim \chi_r^2$. □

Example 54.2. *Suppose $X_1, \dots, X_n$ are i.i.d $N(0,1)$. Then $\bar{X} \sim N(0, 1/n)$ and $S \sim \chi_{n-1}^2$ where*

$$S := \sum_{i=1}^n \left(X_i - \bar{X}\right)^2.$$

Moreover $\bar{X}$ and S are independent.

The fact that $\bar{X} \sim N(0,1/n)$ is easy. To prove that $S \sim \chi_{n-1}^2$ and that S and $\bar{X}$ are independent, we shall show two methods.

Method One*: To prove $S \sim \chi_{n-1}^2$, the key is to note that*

$$S = \left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)X\right)^T \left(\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)X\right) = X^T \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)^T \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X = X^T \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X$$

where $X = (X_1, \dots, X_n)^T$ and $\mathbf{1} = (1, \dots, 1)^T$. In the last step above, we used the fact that $I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is symmetric. For the first step, we used the fact that

$$\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X = (X_1 - \bar{X}, \dots, X_n - \bar{X})^T.$$

Now if

$$A = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T,$$

then clearly A is symmetric and idempotent as

$$A^2 = \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) = I - 2\frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{\mathbf{1}^T\mathbf{1}}{n^2}\mathbf{1}\mathbf{1}^T = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T = A.$$

Also the rank of A equals $n-1$. Thus by Theorem 54.1 (note that $X = (X_1, \dots, X_n)^T \sim N_n(0, I_n)$), we have

$$S = X^T A X \sim \chi_{n-1}^2.$$

In order to prove that S and $\bar{X}$ are independent, we only need to observe that

$$\bar{X} = \frac{1}{n}\mathbf{1}^T X \quad \text{and} \quad \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X \tag{82}$$

are independent because S is a function of

$$\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X.$$

The independence of the random variables in (82) follows because

$$\frac{1}{n}\mathbf{1}^T \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X = 0$$

***Method Two**: Let $u_1, \dots, u_n$ be an orthonormal basis for $\mathbb{R}^n$ with $u_1 = \mathbf{1}/\sqrt{n}$ (check that u_1 has unit norm). Let U be the matrix with columns $u_1, \dots, u_n$ i.e.,*

$$U = [u_1 : \cdots : u_n].$$

Note that $UU^T = U^T U = I_n$ (by the properties of an orthonormal basis). Now let $Y = U^T X$. Then Y is a linear function of X (and $X \sim N_n(0, I_n)$) so that

$$Y \sim N_n(U^T(0), U^T I_n U) = N_n(0, U^T U) = N_n(0, I_n).$$

Further note that

$$\sum_{i=1}^n Y_i^2 = Y^T Y = X^T U U^T X = X^T X = \sum_{i=1}^n X_i^2 \tag{83}$$

and that $Y_1 = u_1^T X = (X_1 + \cdots + X_n)/\sqrt{n} = \sqrt{n}\bar{X}$. Thus, (83) gives

$$\sum_{i=1}^n X_i^2 = Y_1^2 + \sum_{i=2}^n Y_i^2 = n\bar{X}^2 + \sum_{i=2}^n Y_i^2$$

so that

$$\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = S.$$

This and the fact that $Y \sim N_n(0, I_n)$ (which is same as saying that $Y_1, \dots, Y_n$ are i.i.d $N(0,1)$) imply that $S \sim \chi^2_{n-1}$. Also note that S depends only on $Y_2, \dots, Y_n$ so that it is independent of Y_1 and thus S and $\bar{X}$ are independent (note that $\bar{X} = Y_1/\sqrt{n}$).

Example 54.3. *Suppose that $X \sim N_n(0, \Sigma)$ where Σ is an $n \times n$ matrix with 1 on the diagonal and ρ on the off-diagonal. Σ can also be represented as*

$$\Sigma = (1-\rho) I_n + \rho \mathbf{1}\mathbf{1}^T \qquad \text{where } \mathbf{1} := (1, \dots, 1)^T.$$

In other words $X_1, \dots, X_n$ are multivariate normal, have mean zero, unit variance and the correlation between every pair equals ρ. Find the distribution of $\bar{X}$ and $S := \sum_{i=1}^n (X_i - \bar{X})^2$ and argue that they are independent.

$\bar{X}$ is a linear function of X and so it will be normal. We then just have to find its mean and variance. Clearly $\mathbb{E}\bar{X} = 0$ (as each $\mathbb{E}X_i = 0$) and

$$var(\bar{X}) = \frac{1}{n^2} var(X_1 + \cdots + X_n) = \frac{1}{n^2}\left(\sum_i var(X_i) + \sum_{i \neq j} Cov(X_i, X_j)\right) = \frac{1}{n^2}\left(n + n(n-1)\rho\right) = \frac{1 + (n-1)\rho}{n}.$$

Thus

$$\bar{X} \sim N\left(0, \frac{1 + (n-1)\rho}{n}\right).$$

Observe that this implies that $1 + (n-1)\rho \geq 0$ or $\rho \geq -1/(n-1)$. In other words, if $\rho < -1/(n-1)$, then Σ will not be positive semi-definite.

To find the distribution of S, we can, as in the previous example, write

$$S = X^T \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X$$

but we cannot unfortunately use Theorem 54.1 as X does not have identity covariance (Theorem 54.1) only applies to multivariate normal random vectors with identity covariance. It turns out that here the second method (described in the previous example) works here and gives the distribution of S. This is explained below.

Let $u_1, \dots, u_n$ be an orthonormal basis for $\mathbb{R}^n$ with $u_1 = \mathbf{1}/\sqrt{n}$ and let U be the matrix with columns $u_1, \dots, u_n$ so that $U^TU = UU^T = I_n$. Let $Y = U^TX$ and note (as in the previous example) that

$$Y_1 = \sqrt{n}\bar{X} \quad \text{and} \quad S = \sum_{i=1}^{n}(X_i - \bar{X})^2 = Y_2^2 + \dots + Y_n^2.$$

The distribution of Y is now given by $Y \sim N_n(0, U^T\Sigma U)$ and

$$U^T\Sigma U = U^T\left((1-\rho)I_n + \rho\mathbf{1}\mathbf{1}^T\right)U = (1-\rho)U^TU + \rho(U^T\mathbf{1})(\mathbf{1}^TU) = (1-\rho)I_n + \rho(U^T\mathbf{1})(\mathbf{1}^TU).$$

To calculate $\mathbf{1}^TU$, note that

$$\mathbf{1}^TU = (\mathbf{1}^Tu_1, \mathbf{1}^Tu_2, \dots, \mathbf{1}^Tu_n) = (\sqrt{n}, 0, \dots, 0)$$

where we used that $\mathbf{1}^Tu_1 = \mathbf{1}^T\mathbf{1}/\sqrt{n} = \sqrt{n}$ and the fact that $\mathbf{1}$ is orthogonal to $u_2, \dots, u_n$ (this is because $\langle \mathbf{1}, u_i\rangle = \sqrt{n}\langle u_1, u_i\rangle = 0$ for $i > 1$). We have thus obtained

$$U^T\Sigma U = (1-\rho)I_n + \rho(\sqrt{n}, 0, \dots, 0)^T(\sqrt{n}, 0, \dots, 0).$$

This means that $U^T\Sigma U$ is a diagonal matrix with diagonal entries $(1-\rho+n\rho), 1-\rho, 1-\rho, \dots, 1-\rho$. Therefore $Y \sim N_n(0, U^T\Sigma U)$ implies that $Y_1, \dots, Y_n$ are independent with

$$Y_1 \sim N(0, 1 + (n-1)\rho) \quad \text{and} \quad Y_i \sim N(0, 1-\rho) \;\; \text{for } i > 1.$$

Thus

$$\frac{S}{1-\rho} = \sum_{i=2}^{n}\left(\frac{Y_i}{\sqrt{1-\rho}}\right)^2 \sim \chi^2_{n-1}$$

or $S \sim (1-\rho)\chi^2_{n-1}$. Also because $\bar{X}$ only depends on Y_1 and S depends only on $Y_2, \dots, Y_n$, we have that S and $\bar{X}$ are independent.

55 Conditional Distributions of Multivariate Normals

Suppose $Y \sim N_n(\mu, \Sigma)$. Let us partition Y into two parts $Y_{(1)}$ and $Y_{(2)}$ where $Y_{(1)} = (Y_1, \dots, Y_p)^T$ consists of the first p components of Y and $Y_{(2)} = (Y_{p+1}, \dots, Y_n)$ consists of the last $q := n - p$ components of Y.

We can then partition the mean vector μ analogously

$$\mu = \begin{pmatrix}\mu_{(1)} \\ \mu_{(2)}\end{pmatrix} = \begin{pmatrix}\mathbb{E}(Y_{(1)}) \\ \mathbb{E}(Y_{(2)})\end{pmatrix}$$

and the covariance matrix Σ as

$$\Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix} = \begin{pmatrix}Cov(Y_{(1)}) & Cov(Y_{(1)}, Y_{(2)}) \\ Cov(Y_{(2)}, Y_{(1)}) & Cov(Y_{(2)})\end{pmatrix}$$

The question we address now is the following: What is the conditional distribution of $Y_{(2)}$ given $Y_{(1)} = y_1$? The answer is given below.

Fact: Under the assumption that $Y \sim N_n(\mu, \Sigma)$, we have

$$Y_{(2)}|Y_{(1)} = y_1 \sim N_p\left(\mu_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_{(1)}), \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right). \tag{84}$$

In words, the conditional distribution of $Y_{(2)}$ given $Y_{(1)} = y_1$ is also multivariate normal with mean vector given by:

$$\mathbb{E}(Y_{(2)}|Y_{(1)} = y_1) = \mu_{(2)} + \Sigma_{21}\Sigma_{11}^{-1}\left(y_1 - \mu_{(1)}\right) = \mathbb{E}(Y_{(2)}) + Cov(Y_{(2)}, Y_{(1)})Cov(Y_{(1)}, Y_{(1)})^{-1}\left(y_1 - \mathbb{E}(Y_{(1)})\right)$$

and covariance matrix given by

$$Cov(Y_{(2)}|Y_{(1)} = y_1) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

We shall go over the proof of (90) below. Before that, let us make a few quick remarks on the form of the conditional distribution:

1. The conditional distributions are also multivariate normal.
2. The conditional expectation $\mathbb{E}(Y_{(2)}|Y_{(1)} = y_1)$ is a linear function of y_1.
3. $\mathbb{E}(Y_{(2)}|Y_{(1)})$ is exactly equal to the BLP of $Y_{(2)}$ in terms of $Y_{(1)}$. Thus the BP and BLP coincide.
4. The conditional covariance matrix $Cov(Y_{(2)}|Y_{(1)} = y_1)$ does not depend on y_1 (this can be viewed as some kind of *homoscedasticity*).
5. The conditional covariance matrix $Cov(Y_{(2)}|Y_{(1)} = y_1)$ equals Σ_{22}^S (the Schur complement of Σ_{22} in Σ). In other words, the conditional covariance matrix $Cov(Y_{(2)}|Y_{(1)} = y_1)$ is precisely equal to the Covariance Matrix of the Residual of $Y_{(2)}$ given $Y_{(1)}$.

Proof of Fact (90): It is easy to see that (90) is equivalent to:

$$\left\{Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(y_1 - \mu_{(1)})\right\}|Y_{(1)} = y_1 \sim N_p\left(0, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right).$$

which is further equivalent to

$$\left\{Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)})\right\}|Y_{(1)} = y_1 \sim N_p\left(0, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right). \tag{85}$$

Note that the distribution on the right hand above does not depend on y_1. Therefore (85) is equivalent to

$$Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)}) \sim N_p\left(0, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right) \quad \text{and } Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)}) \perp\!\!\!\perp Y_1$$

where $\perp\!\!\!\perp$ denotes independence. Because Y is multivariate normal, we know that linear functions of Y are also multivariate normal and that linear functions of Y are independent if and only if they are uncorrelated. The above displayed assertion is therefore equivalent to the following three equations:

1. $\mathbb{E}\left(Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)})\right) = 0$
2. $Cov\left(Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)})\right) = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$.
3. $Cov\left(Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)}), Y_{(1)}\right) = 0$.

In other words, we have noted that proving the above three equations is equivalent to proving (90) . We now complete the proof of (90) by proving the three equations above. We actually have already proved these three facts. The first fact simply says that the residual of $Y_{(2)}$ given $Y_{(1)}$ has zero expectation. The second fact says that the covariance matrix of the residual equals the Schur complement. The third fact says that the residual of $Y_{(2)}$ given $Y_{(1)}$ is uncorrelated with $Y_{(1)}$. For completeness, let us rederive these quickly as follows.

$$\mathbb{E}\left(Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)})\right) = \mu_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(\mu_{(1)} - \mu_{(1)}) = 0, \tag{86}$$

$$\begin{aligned} Cov\left(Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)})\right) &= Cov\left(\begin{pmatrix}-\Sigma_{21}\Sigma_{11}^{-1} & I\end{pmatrix}\begin{pmatrix}Y_{(1)} - \mu_{(1)} \\ Y_{(2)} - \mu_{(2)}\end{pmatrix}\right) \\ &= \begin{pmatrix}-\Sigma_{21}\Sigma_{11}^{-1} & I\end{pmatrix} Cov\begin{pmatrix}Y_{(1)} - \mu_{(1)} \\ Y_{(2)} - \mu_{(2)}\end{pmatrix}\begin{pmatrix}-\Sigma_{11}^{-1}\Sigma_{12} \\ I\end{pmatrix} \\ &= \begin{pmatrix}-\Sigma_{21}\Sigma_{11}^{-1} & I\end{pmatrix}\begin{pmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}\begin{pmatrix}-\Sigma_{11}^{-1}\Sigma_{12} \\ I\end{pmatrix} \\ &= \begin{pmatrix}0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\end{pmatrix}\begin{pmatrix}-\Sigma_{11}^{-1}\Sigma_{12} \\ I\end{pmatrix} \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{aligned} \tag{87}$$

and finally

$$
\begin{aligned}
Cov\left(Y_{(2)} - \mu_{(2)} - \Sigma_{21}\Sigma_{11}^{-1}(Y_{(1)} - \mu_{(1)}), Y_{(1)}\right) &= Cov(Y_{(2)}, Y_{(1)}) - \Sigma_{21}\Sigma_{11}^{-1}Cov(Y_{(1)}, Y_{(1)}) \\
&= \Sigma_{21} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{11} = 0 \qquad (88)
\end{aligned}
$$

This completes the proof of (90).

Let us reiterate that all the three calculations (86), (87) and (88) do not require any distributional assumptions on $Y_{(1)}$ and $Y_{(2)}$. They hold for all random variables $Y_{(1)}$ and $Y_{(2)}$. The multivariate normality assumption allows us to deduce (90) from these second order (i.e., mean and covariance) calculations.

56 Some Remarks on Multivariate Normals and Chi-Squared Distributions

In the last class, we saw the following result.

Theorem 56.1. *If $Z \sim N_n(0, I_n)$ and A is a symmetric and idempotent matrix, then $Z^T AZ \sim \chi^2_r$ where r is the rank of A.*

It turns out that the converse of this result is also true and we have

Theorem 56.2. *Suppose $Z \sim N_n(0, I_n)$ and A is a symmetric matrix. Then $Z^T AZ \sim \chi^2_r$ if and only if A is an idempotent matrix with rank r.*

In other words, the only way in which $Z^T AZ$ has the chi-squared distribution is when A is idempotent. Thus being idempotent is both necessary and sufficient for $Z^T AZ$ to be distributed as chi-squared. The fact that $Z^T AZ \sim \chi^2_r$ implies the idempotence of A can be proved via moment generating functions but we shall skip this argument.

When the covariance is not the identity matrix, Theorem 56.1 needs to be modified as demonstrated below. Suppose now that $Y \sim N_n(0, \Sigma)$ and we are interested in seeing when $Q := Y^T AY$ is idempotent (here A is a symmetric matrix). We know that $Z := \Sigma^{-1/2}Y \sim N_n(0, I_n)$ so we can write (as $Y = \Sigma^{1/2}Z$)

$$Q = Y^T AY = Z^T \Sigma^{1/2} A \Sigma^{1/2} Z.$$

Thus Q is chi-squared distributed if and only if $\Sigma^{1/2}A\Sigma^{1/2}$ is idempotent which is equivalent to

$$\Sigma^{1/2}A\Sigma^{1/2}\Sigma^{1/2}A\Sigma^{1/2} = \Sigma^{1/2}A\Sigma^{1/2} \iff A\Sigma A = A.$$

We thus have

Theorem 56.3. *Suppose $Y \sim N_n(0, \Sigma)$ and A is a symmetric matrix. Then $Y^T AY \sim \chi^2_r$ if and only if $A\Sigma A = A$ and $r = rank(\Sigma^{1/2}A\Sigma^{1/2}) = rank(A)$.*

Let us look at some examples of this result.

Example 56.4. *Suppose $Y \sim N_n(0, \sigma^2 I_n)$ and let A be an $n \times n$ symmetric idempotent matrix with rank r. Then it turns out that*

$$Q := \frac{1}{\sigma^2} Y^T AY \sim \chi^2_r.$$

This can be proved as a consequence of Theorem 56.3 because

$$Q = Y^T \frac{A}{\sigma^2} Y$$

and

$$\frac{A}{\sigma^2}(\sigma^2 I_n)\frac{A}{\sigma^2} = \frac{A}{\sigma^2}.$$

Example 56.5. *Suppose that $X \sim N_n(0, \Sigma)$ where Σ is an $n \times n$ matrix with 1 on the diagonal and ρ on the off-diagonal. Σ can also be represented as*

$$\Sigma = (1-\rho)I_n + \rho \mathbf{1}\mathbf{1}^T \qquad \text{where } \mathbf{1} := (1, \dots, 1)^T.$$

In the last class, we showed that $S := \sum_{i=1}^n (X_i - \bar{X})^2$ satisfies

$$\frac{S}{1-\rho} \sim \chi^2_{n-1}. \tag{89}$$

We shall show this here using Theorem 56.3. Note first that

$$S = X^T \left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right) X$$

so that

$$\frac{S}{1-\rho} = X^T A X \qquad \text{where } A := \frac{1}{1-\rho}\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)$$

Thus to show (89)*, we only need to prove that $A\Sigma A = A$. For this, see that*

$$\begin{aligned}\Sigma A &= I - \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{\rho}{1-\rho}\mathbf{1}\mathbf{1}^T\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\\ &= I - \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{\rho}{1-\rho}\mathbf{1}\left(\mathbf{1}^T - \frac{1}{n}\mathbf{1}^T\mathbf{1}\mathbf{1}^T\right)\\ &= I - \frac{1}{n}\mathbf{1}\mathbf{1}^T + \frac{\rho}{1-\rho}\mathbf{1}\left(\mathbf{1}^T - \frac{1}{n}n\mathbf{1}^T\right) = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T\end{aligned}$$

so that $A\Sigma A = A$. Thus Theorem 56.3 immediately gives (89)*.*

Example 56.6. *Suppose $Y \sim N_n(0, \Sigma)$. Then $Y^T \Sigma^{-1} Y \sim \chi^2_n$. This follows directly from Theorem 56.3 by taking $A = \Sigma^{-1}$.*

Finally let us mention that when $Z \sim N_n(\mu, I_n)$ and A is idempotent, then $Z^T A Z$ will be a non-central chi-squared distribution. We will not study these in this class.

57 Conditional Distributions of Multivariate Normals

Suppose that $Y_{(1)}$ and $Y_{(2)}$ are two random vectors (with possibly different lengths) such that the vector $(Y_{(1)}^T, Y_{(2)}^T)^T$ is multivariate normal. Then, as we saw in the last class, we have

$$\begin{aligned} Y_{(2)}|Y_{(1)} = y_1 \sim N_p &\left(\mathbb{E}(Y_{(2)}) + Cov(Y_{(2)}, Y_{(1)})(Cov Y_{(1)})^{-1}(y_1 - \mathbb{E}(Y_{(1)}))\right. \qquad (90)\\ &\left., Cov(Y_{(2)}) - Cov(Y_{(2)}, Y_{(1)})(Cov Y_{(1)})^{-1} Cov(Y_{(1)}, Y_{(2)})\right). \qquad (91)\end{aligned}$$

In words, the conditional distribution of $Y_{(2)}$ given $Y_{(1)} = y_1$ is also multivariate normal with mean vector given by:

$$\mathbb{E}(Y_{(2)}|Y_{(1)} = y_1) = \mathbb{E}(Y_{(2)}) + Cov(Y_{(2)}, Y_{(1)}) Cov(Y_{(1)})^{-1}\left(y_1 - \mathbb{E}(Y_{(1)})\right)$$

and covariance matrix given by

$$Cov(Y_{(2)}|Y_{(1)} = y_1) = Cov(Y_{(2)}) - Cov(Y_{(2)}, Y_{(1)})(Cov Y_{(1)})^{-1} Cov(Y_{(1)}, Y_{(2)})$$

It is important to note that

1. $\mathbb{E}(Y_{(2)}|Y_{(1)})$ is exactly equal to the BLP of $Y_{(2)}$ in terms of $Y_{(1)}$. Thus the BP and BLP coincide.

2. The conditional covariance matrix $Cov(Y_{(2)}|Y_{(1)} = y_1)$ does not depend on y_1 (this can be viewed as some kind of *homoscedasticity*).

3. The conditional covariance matrix $Cov(Y_{(2)}|Y_{(1)} = y_1)$ is precisely equal to the Covariance Matrix of the Residual of $Y_{(2)}$ given $Y_{(1)}$.

We can look at the following special case of this result. Fix two components Y_i and Y_j of Y. Let $Y_{(2)} := (Y_i, Y_j)^T$ and let $Y_{(1)}$ denote the vector obtained from all the other components $Y_k, k \neq i, k \neq j$. Then

$$Cov(Y_{(2)}|Y_{(1)} = y_1) = Cov(Y_{(2)}) - Cov(Y_{(2)}, Y_{(1)})(CovY_{(1)})^{-1}Cov(Y_{(1)}, Y_{(2)})$$

Now let $r_{Y_i|Y_k,k\neq i,k\neq j}$ and $r_{Y_j|Y_k,k\neq i,k\neq j}$ denote the residuals of Y_i in terms of $Y_k, k \neq i, k \neq j$ and Y_j in terms of $Y_k, k \neq i, k \neq j$, then we have seen previously that

$$Cov\begin{pmatrix} r_{Y_i|Y_k,k\neq i,k\neq j} \\ r_{Y_j|Y_k,k\neq i,k\neq j} \end{pmatrix} = Cov(Y_{(2)}) - Cov(Y_{(2)}, Y_{(1)})(CovY_{(1)})^{-1}Cov(Y_{(1)}, Y_{(2)}).$$

We thus have

$$Cov(\begin{pmatrix} Y_i \\ Y_j \end{pmatrix} \Bigg| Y_k = y_k, k \neq i, k \neq j) = Cov\begin{pmatrix} r_{Y_i|Y_k,k\neq i,k\neq j} \\ r_{Y_j|Y_k,k\neq i,k\neq j} \end{pmatrix} \qquad \text{for every } y_k, k \neq i, k \neq j.$$

This means, in particular, that the conditional correlation between Y_i and Y_j given $Y_k = y_k, k \neq i, k \neq j$ is precisely equal to the partial correlation $\rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j}$ (recall that $\rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j}$ is the correlation between $r_{Y_i|Y_k,k\neq i,k\neq j}$ and $r_{Y_j|Y_k,k\neq i,k\neq j}$).

Now recall the following connection between partial correlation and entries of Σ^{-1} that we have seen earlier:

$$\rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j} = \frac{-\Sigma^{-1}(i,j)}{\sqrt{\Sigma^{-1}(i,i)\Sigma^{-1}(j,j)}}.$$

Putting the above observations together, we can deduce that the following are **equivalent** when Y is multivariate normal with covariance matrix Σ:

1. $\Sigma^{-1}(i,j) = 0$

2. $\rho_{Y_i,Y_j|Y_k,k\neq i,k\neq j} = 0$

3. The conditional correlation between Y_i and Y_j given $Y_k = y_k, k \neq i, k \neq j$ equals 0 for every choice of $y_k, k \neq i, k \neq j$.

4. Y_i and Y_j are conditionally independent given $Y_k = y_k, k \neq i, k \neq j$ equals 0 for every choice of $y_k, k \neq i, k \neq j$.