STAT 201A - Introduction to Probability at an advanced level
All Lectures

Fall 2022, UC Berkeley

Aditya Guntuboyina

December 2, 2022

Contents

1 Lecture One

1.1 What is Probability Theory?

Probability theory is what one should use when reasoning in the presence of uncertainty.

1.2 How does Probability Theory work?

Suppose we are interested in knowing whether a certain proposition is true. Suppose that we do not have access to full information that would allow us to conclusively determine whether the proposition is true or not. Probability theory allows us to determine a number between 0 and 1 representing how likely it is that the proposition is true based on the available information. This is achieved by the following two steps:

1. **Step One**: The available information that we either possess or that we assume for the sake of argument is converted into **numerical assignments** for the probabilities of certain basic or elementary propositions. This step is often referred to as the **modeling** step.

2. **Step Two**: Based on the probability model, we calculate probabilities of the propositions of interest using the **rules of probability**.

1.3 Rules of Probability

Probabilities are assigned to propositions (also known as events). Every probability is conditional on some information (this could be available information or some information that we assume for the sake of argument). We shall denote the probability of a proposition A conditioned on some information I by $\mathbb{P}(A \mid I)$. When the information I is clear from context, we sometimes omit it and write the probability $\mathbb{P}(A \mid I)$ as simply $\mathbb{P}(A)$. Even when we do this, it should always be kept in mind that probabilities are always conditioned on some information.

1. The probability of a proposition always lies between 0 and 1. The probability of an impossible proposition is 0 and the probability of a certain proposition is 1.
2. **Product Rule**: $\mathbb{P}(A \cap B \mid I) = \mathbb{P}(A \mid I)\mathbb{P}(B \mid A, I) = \mathbb{P}(B \mid I)\mathbb{P}(A \mid B, I)$. Here $A \cap B$ is the proposition: "both A and B are true". Also $\mathbb{P}(A \mid B, I)$ is the probability of A conditioned on the truth of the proposition B as well as the information I. A direct consequence of the product rule is:

$$\mathbb{P}(B \mid A, I) = \frac{\mathbb{P}(A \mid B, I)\mathbb{P}(B \mid I)}{\mathbb{P}(A \mid I)}.$$

   The above formula is known as the Bayes rule.
3. **Sum Rule**: $\mathbb{P}(A \cup B \mid I) = \mathbb{P}(A \mid I) + \mathbb{P}(B \mid I)$ for disjoint propositions A and B. Here $A \cup B$ denotes the proposition: "at least one of A and B is true".

We shall see some justification for these rules later.

1.4 Example 1: Testing and Covid

Problem 1.1. *Suppose I just tested positive for Covid. Do I really have Covid?*

This is a situation involving uncertainty mainly because the test may not be 100% accurate. In other words, my result could be a false positive. I need to calculate

$$\mathbb{P}\{\text{I have Covid} \mid \text{I tested positive+other background information}\}$$

which we abbreviate as $\mathbb{P}(C \mid +, B)$ for simplicity of notation. Here B denotes relevant background information that I may have. For example, B could include things like "I have been strictly quarantining for the past 3 weeks" etc.

We can attempt to calculate $\mathbb{P}(C \mid +, B)$ as:

$$\mathbb{P}(C \mid +, B) = \frac{\mathbb{P}(C, + \mid B)}{\mathbb{P}(+ \mid B)} = \frac{\mathbb{P}(+ \mid C, B)\mathbb{P}(C \mid B)}{\mathbb{P}(+ \mid C, B)\mathbb{P}(C \mid B) + \mathbb{P}(+ \mid C^c, B)\mathbb{P}(C^c \mid B)}$$

where we used the product rule and sum rule of probability. Here C^c denotes the proposition that I do not have Covid.

In order to proceed further, we need some probability assignment. Consider the following assignment:

$$\mathbb{P}(C \mid B) = 0.02 \quad \mathbb{P}(+ \mid C, B) = \mathbb{P}(+ \mid C) = 0.99 \quad \mathbb{P}(+ \mid C^c, B) = \mathbb{P}(+ \mid C^c) = 0.04. \quad (1)$$

$\mathbb{P}(C \mid B)$ represents the probability of Covid based on background information alone. The fact that it is low (0.02) is meaningful when I know that I have been largely isolating myself

for the past few weeks. With this assignment, we can calculate the required probability $\mathbb{P}(C \mid +, B)$ as follows:

$$\begin{aligned}\mathbb{P}(C \mid +, B) &= \frac{\mathbb{P}(+ \mid C, B)\mathbb{P}(C \mid B)}{\mathbb{P}(+ \mid C, B)\mathbb{P}(C \mid B) + \mathbb{P}(+ \mid C^c, B)\mathbb{P}(C^c \mid B)} \\ &= \frac{0.99 * 0.02}{0.99 * 0.02 + 0.04 * 0.98} = 0.3356.\end{aligned}$$

Note that 0.3356 (33.56%) is not very high even though the test has very good false positive and false negative rates. This is because $\mathbb{P}(C \mid B)$ (which can be interpreted as probability of having Covid without taking into the account the test result) is very low (0.02).

Here is an alternative method of reasoning in this problem. We formulate this as a hypothesis testing problem with

$$H_0 : \text{I do not have Covid} \quad \text{versus} \quad H_1 : \text{I have Covid}$$

The p-value in the above testing problem equals:

$$\mathbb{P}\{+|H_0\} = \mathbb{P}(+|C^c) = 0.04.$$

Usage of the naive cutoff 0.05 on the p-value would now lead to rejection of the null hypothesis and declaring that I have Covid. On the other hand, the previous argument (based on probability theory) gave a much higher probability to me not having Covid. This p-value based method does not even make use of the information given on $\mathbb{P}(C \mid B)$ and $\mathbb{P}(+|C, B)$. It only makes use of $\mathbb{P}(+|C^c)$. Note that what we are after is $\mathbb{P}(C^c|+)$ (or $\mathbb{P}(C|+)$). In general, $\mathbb{P}(A|B)$ and $\mathbb{P}(B|A)$ can be quite different. Consider, for example, the case where A represents the event that a person is dead and B represents the event that they were hanged. It is therefore quite problematic that one can say something about $C|+$ or $C^c|+$ from knowledge of $\mathbb{P}(+|C^c)$ alone.

Methods such as testing based on p-values (and putting arbitrary cutoffs on them) are not based on probability theory. The use of p-values has been linked to serious issues such as lack of reproducibility. In this context, we can calculate the probability of reproducibility of the positive test:

$$\mathbb{P}(+_2|+_1, B) = \mathbb{P}(+_2|C, +_1, B)\mathbb{P}(C|+_1, B) + \mathbb{P}(+_2|C^c, +_1, B)\mathbb{P}(C^c|+_1, B).$$

Here $+_2$ denotes the proposition that the second test results in a positive (and $+_1$ denotes the proposition that the first test resulted in a positive). We now make the following probability assignment:

$$\mathbb{P}(+_2|C, +_1, B) = \mathbb{P}(+_2|C) = 0.99 \quad \text{and} \quad \mathbb{P}(+_2|C^c, +_1, B) = \mathbb{P}(+_2 \mid C^c) = 0.04.$$

This assumption means that conditional on my Covid status, the two tests are independent. Using this assignment, it is straightforward to calculate the reproducibility probability as follows (note that we already calculated $\mathbb{P}(C \mid +_1, B) = 1 - \mathbb{P}(C^c \mid +_1, B) = 0.3356$)

$$\mathbb{P}(+_2|+_1, B) = 0.99 * 0.3356 + 0.04 * (1 - 0.3356) = 0.35882.$$

Thus this positive test wil be reproducible with probability only 35.88%.

1.5 Example 2: Spots on a patient

Problem 1.2 (From the book "A tutorial introduction to Bayesian Analysis" by James Stone)**.** *Suppose you are a doctor confronted with a patient who is covered in spots. Ths*

patient's symptoms are consistent with chickenpox but they are also consistent with another, more dangerous, disease, smallpox. How would you decide if they have chickenpox or smallpox?

This is again a situation involving uncertainty as the doctor does not know which disease the patient has. The doctor needs to calculate the probability:

$$\mathbb{P}\{\text{smallpox} \mid \text{spots} + B\} \tag{2}$$

where B again represents background information. For example, B could represent any other symptoms that the patient has such as fever. Here is one probability assignment which allows us to calculate this probability:

$$\mathbb{P}\{\text{spots} \mid \text{smallpox}, B\} = 0.9 \quad \mathbb{P}\{\text{spots} \mid \text{chickenpox}, B\} = 0.8 \quad \mathbb{P}\{\text{spots} \mid \text{neither}, B\} = 0 \tag{3}$$

and

$$\mathbb{P}\{\text{smallpox} \mid B\} = 0.001 \quad \mathbb{P}\{\text{chickenpox} \mid B\} = 0.1 \quad \mathbb{P}\{\text{neither} \mid B\} = 0.899. \tag{4}$$

Here "neither" refers to an underlying cause for the patient's condition that is neither smallpox nor chickenpox.

Using this assignment, the required probability (2) can be calculated via Bayes rule and this leads to

$$\mathbb{P}\{\text{smallpox} \mid \text{spots}, B\} \approx 0.011 \quad \mathbb{P}\{\text{chickenpox} \mid \text{spots}, B\} \approx 0.988 \quad \mathbb{P}\{\text{neither} \mid \text{spots}, B\} = 0.$$

So probability theory with the assignment (3) and (4) says that it is highly likely that the patient has chickenpox (smallpox is basically ruled out because it is extremely rare).

Here is an alternative way of solving this problem using maximum likelihood estimation. The maximum likelihood estimate in this case is smallpox because smallpox leads to a higher probability (0.9) of the observed data (spots) compared to chickenpox (0.8). Maximum Likelihood (widely used in statistics) is not based on probability theory and also seems to be based on the wrong conditional probabilities $\mathbb{P}\{\text{spots}|\text{smallpox}\}$ and $\mathbb{P}\{\text{spots}|\text{chickenpox}\}$ while we really should be calculating $\mathbb{P}\{\text{smallpox}|\text{spots}\}$ and $\mathbb{P}\{\text{chickenpox}|\text{spots}\}$.

2 Lecture Two

We shall continue our discussion of the scope of probability theory with more examples.

2.1 Example 3: Prisoner's dilemma

The following is a standard problem (see, for example, Mosteller [4, Problem 13]).

Example 2.1 (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. The prisoner A has a friend who is the warder of the prison and who knows which prisoners will be released. Prisoner A realizes that it would be unethical to ask the warder if he, A, is to be released, but decides to ask for the name of one prisoner other than himself who is to be released. The warder says "B will be released". What are the chances of A being released?*

We need to calculate

$$\mathbb{P}\{A \text{ will be released} \mid \text{Warder says B will be released}\}.$$

By the product rule of probability, the above probability is the same as

$$\frac{\mathbb{P}\{\text{A and B will be released}\}}{\mathbb{P}\{\text{Warder says B will be released}\}}$$

To calculate the numerator, it is natural to make the assignment

$$\mathbb{P}\{\text{A and B will be released}\} = \mathbb{P}\{\text{B and C will be released}\} = \mathbb{P}\{\text{A and C will be released}\} = 1/3.$$

For the denominator, we can split as

$$\begin{aligned}
&\mathbb{P}\{\text{Warder says B will be released}\} \\
&= \mathbb{P}\{\text{Warder says B will be released} \mid \text{A and B will be released}\}\,\mathbb{P}\{\text{A and B will be released}\} \\
&+ \mathbb{P}\{\text{Warder says B will be released} \mid \text{B and C will be released}\}\,\mathbb{P}\{\text{B and C will be released}\} \\
&+ \mathbb{P}\{\text{Warder says B will be released} \mid \text{A and C will be released}\}\,\mathbb{P}\{\text{A and C will be released}\} \\
&= 1 \times \frac{1}{3} + \mathbb{P}\{\text{Warder says B will be released} \mid \text{B and C will be released}\} \times \frac{1}{3} + 0 \times \frac{1}{3} \\
&= \frac{1}{3} + \frac{1}{3}\mathbb{P}\{\text{Warder says B will be released} \mid \text{B and C will be released}\}.
\end{aligned}$$

To proceed further, we need a probability assignment for

$$\mathbb{P}\{\text{Warder says B will be released} \mid \text{B and C will be released}\}$$

It is natural to assume that this probability equals 0.5. This means that, in the event that B and C are the two prisoners who will be released, the warder is equally likely to reveal the name of B or C to A. Under this assumption, we have

$$\mathbb{P}\{\text{Warder says B will be released}\} = \frac{1}{3} + \frac{1}{3} \times \frac{1}{2} = \frac{1}{2}$$

leading to

$$\mathbb{P}\{A \text{ will be released} \mid \text{Warder says B will be released}\} = \frac{1/3}{1/2} = \frac{2}{3}.$$

Note that this means that prisoner A's chances of being released remain the same in spite of the additional information revealed by the warder.

Here is an interesting wrinkle on this problem. Suppose that the conversation between the prisoner A and the Warder was overheard by prisoner C who then proceeds to calculate his own chances of being released in light of the additional information.

$$\begin{aligned}
&\mathbb{P}\{C \text{ will be released} \mid \text{Warder says B will be released}\} \\
&= \frac{\mathbb{P}\{C \text{ will be released, Warder says B will be released}\}}{\mathbb{P}\{\text{Warder says B will be released}\}}
\end{aligned}$$

The denominator is the same as before so it will be 1/2. For the numerator, note that C will be released either with A or with B. The case where C and A will be released is ruled out because the Warder said "B will be released". So the numerator equals:

$$\begin{aligned}
&\mathbb{P}\{\text{C and B will be released, Warder says B will be released}\} \\
&= \mathbb{P}\{\text{Warder says B will be released} \mid \text{B and C will be released}\}\,\mathbb{P}\{\text{B and C will be released}\} \\
&= \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}.
\end{aligned}$$

Therefore the probability of C's release given the additional information is $(1/6)/(1/2) = 1/3$. The additional information therefore significantly reduces the chances of C's release (from 2/3 to 1/3).

2.2 Example 4: Monty Hall Problems

Example 2.2 (Monty Hall Problem)**.** *Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?*

Let us suppose that I always pick Door 1 to start the game. We then need to calculate the conditional probability:

$$\mathbb{P}\{\text{car in door 2} \mid \text{host opened door 3}\}$$

which we can write as

$$\frac{\mathbb{P}\{\text{host door 3} \mid \text{car door 2}\}\,\mathbb{P}\{\text{car door 2}\}}{\mathbb{P}\{\text{host door 3} \mid \text{car door 2}\}\,\mathbb{P}\{\text{car door 2}\} + \mathbb{P}\{\text{host door 3} \mid \text{car door 1}\}\,\mathbb{P}\{\text{car door 1}\}}.$$

We now make the following natural probability assignment:

$$\mathbb{P}\{\text{host door 3}|\text{car door 2}\} = 1 \quad \text{and} \quad \mathbb{P}\{\text{host door 3}|\text{car door 1}\} = \frac{1}{2}$$

and also

$$\mathbb{P}\{\text{car door 1}\} = \mathbb{P}\{\text{car door 2}\} = \frac{1}{3}.$$

This leads to

$$\mathbb{P}\{\text{car door 2}|\text{host door 3}\} = \frac{1 * (1/3)}{1 * (1/3) + (1/2) * (1/3)} = \frac{2}{3}$$

and since this probability is more than 0.5, it makes sense for me to switch to door 2 from my original selction of door 1.

2.3 Example 5: MacKay sequence example

The following application of probability theory can be seen as a way of formalizing common sense. Probability theory has been described by some (for example, Laplace) as an extension of common sense. Here is a quote by Laplace on this: *It is seen in this essay that the theory of Probabilities is at bottom only common sense reduced to calculus; it makes us appreciate with exactitude that which exact minds feel by a sort of instinct without being able ofttimes to give a reason for it.*–Laplace.

The application given below is from the book MacKay [5, Chapter 28].

Problem 2.3. *Find the next number in the sequence:* $-1, 3, 7, 11$.

Note that this is a problem of reasoning under uncertainty as we are uncertain about the way this sequence of numbers has been generated. One way of using probability theory to solve this problem is the following. We can have two models for the number generation mechanism here:

1. **Model 1**: Arithmetic Progression i.e., $a_1 = \alpha$ and $a_{n+1} = a_n + \beta$.

2. **Model 2**: Random

Most people would look at the sequence and guess the next number as 15. In other words, they are using Model 1 (Arithmetic Progression). Probability theory can be used to justify this. We need to calculate

$$\mathbb{P}\{\text{Model } i \mid \text{data}\} \qquad \text{for } i = 1, 2.$$

What probability assignments would we need to calculate the above? We can use the Bayes Rule to write

$$\mathbb{P}\{\text{Model } i|\text{data}\} = \frac{\mathbb{P}\{\text{data}|\text{Model } i\}\,\mathbb{P}\{\text{Model } i\}}{\mathbb{P}\{\text{data}|\text{Model 1}\}\,\mathbb{P}\{\text{Model 1}\} + \mathbb{P}\{\text{data}|\text{Model 2}\}\,\mathbb{P}\{\text{Model 2}\}} \tag{5}$$

To be fair to each of the two models, we shall take

$$\mathbb{P}\{\text{Model } i\} = \frac{1}{2} \qquad \text{for each } i = 1, 2 \tag{6}$$

We now need to calculate $\mathbb{P}\{\text{data}|\text{Model } i\}$ for $i = 1, 2$. For $i = 1$, we have (below α and β are the parameters in Model 1):

$$\mathbb{P}\{\text{data}|\text{Model 1}\} = \mathbb{P}\{\alpha = -1, \beta = 4\}$$

To calculate the above, we need to make a probability assignment for the probability with which α and β take various values. MacKay [5, Chapter 28] assumes that α and β are integer-valued that they are independently uniformly distributed over the set $\{-50, -49, \ldots, 49, 50\}$ which has cardinality 101. Then

$$\mathbb{P}\{\text{data}|\text{Model 1}\} = \mathbb{P}\{\alpha = -1, \beta = 4\} = \mathbb{P}\{\alpha = -1\}\mathbb{P}\{\beta = 4\} = \left(\frac{1}{101}\right)^2 \approx 9.8 \times 10^{-5}. \tag{7}$$

For the second model, we need to specify what we mean by "random". We shall take this to mean that a_1, a_2, a_3, a_4 are independently distributed according to the uniform distribution on $\{-50, -49, \ldots, 49, 50\}$. Then

$$\begin{aligned}\mathbb{P}\{\text{data}|\text{Model 2}\} &= \mathbb{P}\{a_1 = -1, a_2 = 3, a_3 = 7, a_4 = 11\} \\ &= \mathbb{P}\{a_1 = -1\}\mathbb{P}\{a_2 = 3\}\mathbb{P}\{a_3 = 7\}\mathbb{P}\{a_4 = 11\} = \left(\frac{1}{101}\right)^4 \approx 9.6 \times 10^{-9}.\end{aligned}$$

Plugging in the above value (as well as (6) and (7)) in (5), we get

$$\mathbb{P}\{\text{Model 1}|\text{data}\} \approx \frac{101^{-2} \times 0.5}{101^{-2} \times 0.5 + 101^{-4} \times 0.5} = 0.999902 \quad \text{and } \mathbb{P}\{\text{Model 2}|\text{data}\} = .000098$$

This analysis clearly favors Model 1 compared to Model 2. The most interesting feature about this analysis is that

$$\mathbb{P}\{\text{Model 1}|\text{data}\} \gg \mathbb{P}\{\text{Model 2}|\text{data}\}$$

even though

$$\mathbb{P}\{\text{Model 1}\} = \mathbb{P}\{\text{Model 2}\}.$$

In other words, we did not dogmatically assert that the data was generated by an arithmetic progression but we gave a fair chance to the two models to explain the observed sequence.

In this example, some people argue in favor of Model 1 on the basis that Model 1 is "simpler" than Model 2. Our analysis above (based on probability theory) does not invoke

any vague notion of simplicity but does some formal calculations which in this case preferred Model 1 to Model 2. In another situation, Model 2 may well be the preferred model.

One can consider other alternative models in this problem. For example, MacKay [5, Chapter 28] considered the following cubic model:

Model 3 (Cubic): These numbers were generated by the formula: $a_1 = a$ and $a_{n+1} = ba_n^3 + ca_n^2 + d$ for an integer a and rational numbers b, c, d.

This cubic model explains the given data perfectly if and only if its four parameters a, b, c, d are chosen as $a = -1, b = -1/11, c = 9/11, d = 23/11$. As a result,

$$\mathbb{P}\{\text{data}|\text{Model 3}\} = \mathbb{P}\{a = -1, b = -1/11, c = 9/11, d = 23/11\}.$$

In order to explicitly calculate the above, we need to make probability assignments for a, b, c, d. MacKay [5] makes the following probability assignment: we assume that these four parameters are independent with a being uniform on $\{-50, -49, \ldots, 49, 50\}$ and b, c, d having the distribution of x/y where $x \sim \text{Unif}\{-50, -49, \ldots, 49, 50\}$ and $y \sim \text{Unif}\{1, \ldots, 50\}$ are independent. Under this assignment:

$$\begin{aligned}\mathbb{P}\{a = -1, b = -1/11, c = 9/11, d = 23/11\} &= \mathbb{P}\{a = -1\}\mathbb{P}\{b = -1/11\}\mathbb{P}\{c = 9/11\}\mathbb{P}\{d = 23/11\} \\ &= \left(\frac{1}{101}\right)\left(4 \cdot \frac{1}{101} \cdot \frac{1}{50}\right)\left(4 \cdot \frac{1}{101} \cdot \frac{1}{50}\right)\left(2 \cdot \frac{1}{101} \cdot \frac{1}{50}\right) \\ &\approx 2.5 \times 10^{-12}.\end{aligned}$$

In the above, we used $\mathbb{P}(b = -1/11) = 4 \cdot (1/101) \cdot (1/50)$ because $-1/11 = -2/22 = -3/33 = -4/44$ and each of these has the probability $(1/101) \cdot (1/50)$. A similar reasoning is used for $\mathbb{P}\{c = 9/11\}$ and $\mathbb{P}\{d = 23/11\}$.

If Model 1, 2, 3 are the only three models considered, the Bayes rule (5) becomes

$$\mathbb{P}\{\text{Model } i|\text{data}\} = \frac{\mathbb{P}\{\text{data}|\text{Model } i\}\,\mathbb{P}\{\text{Model } i\}}{\mathbb{P}\{\text{data}\}}$$

where the denominator should be calculated as:

$$\mathbb{P}\{\text{data}\} = \sum_{i=1}^{3} \mathbb{P}\{\text{data}|\text{Model } i\}\,\mathbb{P}\{\text{Model } i\}.$$

Under the fair assumption

$$\mathbb{P}\{\text{Model } i\} = \frac{1}{3} \qquad \text{for each } i = 1, 2, 3,$$

we obtain

$$\mathbb{P}\{\text{Model 1}|\text{data}\} = \frac{101^{-2} \times (1/3)}{101^{-2} \times (1/3) + 101^{-4} \times (1/3) + 2.5 \times 10^{-12} \times (1/3)} = 0.999902$$

and

$$\mathbb{P}\{\text{Model 2}|\text{data}\} \approx 9.8 \times 10^{-5}$$

and

$$\mathbb{P}\{\text{Model 3}|\text{data}\} \approx 9.8 \times 10^{-5} \approx 2.55 \times 10^{-8}.$$

Our preference for Model 1 is still as strong as before (when we only considered the two models Model 1 and Model 2).

The analysis given here depends on the specific choices of priors used for the three models. One can of course use alternative priors but the qualitative preference for Model 1 is unlikely to change for most **reasonable** prior choices.

2.4 Interpretation of Probability

There are multiple interpretations of probability and still some controversy regarding what probability really is. A proper understanding of the meaning of probability is very important for applications of probability to statistics and data analysis.

There are broadly two ways of understanding probability.

2.4.1 Frequentist/Objective Understanding of Probability

From the frequentist viewpoint, probability is applicable only in the context of "random experiments" (such as tossing coins and rolling dice). The probability $\mathbb{P}(A)$ of an event A is defined as the relative frequency that A occurs in N repeated trials of the experiment in the limit as $N \to \infty$:

$$\mathbb{P}(A) := \lim_{N\to\infty} \frac{n_A}{N}$$

where n_A is the number of trials out of N where A occurs. This definition of probability is the basis of frequentist statistics. Here are some examples:

1. The statement $\mathbb{P}(H) = 0.5$ means that the proportion of heads in a large number of tosses of the coin approaches 0.5.

2. The statement
$$\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$
means that if the experiment generating $\epsilon_1, \ldots, \epsilon_n$ is repeated a large number of times, the proportion of times the values of $(\epsilon_1, \ldots, \epsilon_n)$ lie in a set A approaches
$$\int_A \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x_i^2}{2\sigma^2}\right) dx_1 \ldots dx_n$$
and this should be true for all subsets A of $\mathbb{R}^n$.

3. The statement
$$\mathbb{P}\left\{\theta \in \left[\bar{X} - 1.96\frac{S}{\sqrt{n}}, \bar{X} + 1.96\frac{S}{\sqrt{n}}\right]\right\} = 0.95$$
means that if we repeat the experiment generating the random variables $\bar{X}$ and S a large number of times, then the proportion of times the interval
$$\left[\bar{X} - 1.96\frac{S}{\sqrt{n}}, \bar{X} + 1.96\frac{S}{\sqrt{n}}\right]$$
contains θ approaches 0.95.

The following are some obvious problems with the frequentist definition:

1. It is very restrictive and hardly ever applicable. In many simple situations where we would like to use probability, the frequency definition is simply does not apply:
   a) Is the suspect X guilty?
   b) What is the chance of rain in Berkeley today?
   c) What is the chance that Y is cancer positive given that they tested positive?

2. Even in situations where the frequency definition is seemingly applicable, closer thought might reveal some issues. For example, the frequentist probability that a coin comes up heads is 0.6 means that 60% of a large number of tosses of the coin should result in 0.6. But the mechanics of no two tosses are really identical and if two tosses are done exactly identically, then we would expect the same outcome by the laws of physics. So the term "identical and independent repetitions of an experiment" is ambiguous.

In the frequentist definition, probability is considered an intrinsic property of the object under investigation which is only accessible by an experiment generating samples of infinite size. The frequentist probability is also referred to as "objective probability". The implication is that we cannot assign it arbitrarily because any probability assignment that does not agree with the frequency in infinite trials is wrong. Unfortunately, the actual frequentist probabililty is seldom known because one cannot generally observe a large number of repetitions of an experiment and so almost all probability assignments are wrong from the frequentist point of view. This is one way of understanding the statistics aphorism: "All models are wrong" (usually attributed to George Box; see `https://en.wikipedia.org/wiki/All_models_are_wrong`).

3 Lecture Three

3.1 Interpretation of Probability

There are multiple interpretations of probability and still some controversy regarding what probability really is. A proper understanding of the meaning of probability is very important for applications of probability to statistics and data analysis.

There are broadly two ways of understanding probability.

3.1.1 Frequentist/Objective Understanding of Probability

From the frequentist viewpoint, probability is applicable only in the context of "random experiments" (such as tossing coins and rolling dice). The probability $\mathbb{P}(A)$ of an event A is defined as the relative frequency that A occurs in N repeated trials of the experiment in the limit as $N \to \infty$:

$$\mathbb{P}(A) := \lim_{N\to\infty} \frac{n_A}{N}$$

where n_A is the number of trials out of N where A occurs. This definition of probability is the basis of frequentist statistics.

According to this definition, the statement $\mathbb{P}(H) = 0.5$ means that the proportion of heads in a large number of tosses of the coin approaches 0.5.

The following are some obvious problems with the frequentist definition:

1. It is very restrictive and hardly ever applicable. In many simple situations where we would like to use probability, the frequency definition is simply does not apply:

   a) Is the suspect X guilty?

   b) What is the chance of rain in Berkeley today?

   c) What is the chance that Y is cancer positive given that they tested positive?

For an interesting anecdote about how this restrictive notion does not simply make sense in some important problems, see deGroot [3, pages 43-44].

2. Even in situations where the frequency definition is seemingly applicable, closer thought might reveal some issues. For example, the frequentist probability that a coin comes up heads is 0.6 means that 60% of a large number of tosses of the coin should result in 0.6. But the mechanics of no two tosses are really identical and if two tosses are done exactly identically, then we would expect the same outcome by the laws of physics. So the term "identical and independent repetitions of an experiment" is ambiguous.

In the frequentist definition, probability is considered an intrinsic property of the object under investigation which is only accessible by an experiment generating samples of infinite size. The frequentist probability is also referred to as "objective probability". The implication is that we cannot assign it arbitrarily because any probability assignment that does not agree with the frequency in infinite trials is wrong. Unfortunately, the actual frequentist probabililty is seldom known because one cannot generally observe a large number of repetitions of an experiment and so almost all probability assignments are wrong from the frequentist point of view. This is one way of understanding the statistics aphorism: "All models are wrong" (usually attributed to George Box; see `https://en.wikipedia.org/wiki/All_models_are_wrong`).

Here are some quotes by famous statisticians/probabilists illustrating how widespread frequentist thinking in probability is:

The numbers p_r should, in fact, be regarded as physical constants of the particular die that we are using, and the question as to their numerical values cannot be answered by the axioms of probability, any more than the size and the weight of the die are determined by the geometrical and mechanical axioms. However, experience shows that in a well-made die the frequency of any event r in any long series of throws usually approaches $1/6$*, and accordingly we shall often assume that all the p_r are equal to* $1/6$... – Cramér.

Here is Jaynes's response to the above quote (from page 317 of his book): *To a physicist, this statement seems to show utter contempt for the known laws of mechanics. The results of tossing a die many times do* ***not*** *tell us any definite number characteristic only of the die. They tell us also something about how the die was tossed. If you toss 'loaded' dice in different ways, you can easily alter the relative frequencies of the faces. With only slightly more difficulty, you can still do this if your dice are perfectly 'honest'.*

Here is a quote by Feller (see page 322 of the Jaynes book) illustrating the thinking that bridge hands possess physical probabilities and that the uniform probability assignment is a convention whose correctness can only be verified by observed frequencies in a random experiment : *The number of possible distributions of cards in bridge is almost* 10^{30}*. Usually we agree to consider them as equally probable. For a check of this convention more than* 10^{30} *experiments would be required – a billion of billion of years if every living person played one game every second, day and night.* – Feller.

In spite of these objections, one positive aspect of the frequentist meaning of probability is that the Rules of Probability follow easily from this definition.

3.1.2 Subjective or Bayesian Understanding of Probability

It should be clear that in order to use probability as a general method for reasoning under uncertainty, we need a much more general understanding of probability than that is allowed

by the frequentist notion. Expounding this general theory is the purpose of the Jaynes book [1].

The basic idea is to first give up on an objective definition of probability and just admit that *there is no such thing as a physical probability* (Jaynes [1, page 325]). Every probability is subjective and is relative to the person who is actually reasoning under uncertainty. More specifically, probability of an event is something that a specific individual (or a robot or a computer) either assigns based on their state of knowledge (available information) or calculates based on their probability assignments for related events. Generally, probability has nothing do with frequency (unless we are in certain special situations where frequency information is available; we shall see examples of this later). Here is a quote by Harold Jeffreys (who was one of the founders of this way of thinking about probability) related to this:

The essence of the present theory is that no probability, direct, prior, or posterior, is simply a frequency. – Jeffreys 1939.

To give a concrete example, from the Bayesian viewpoint, the statement $\mathbb{P}(H) = 0.5$ will considered to be the assignment made by some specific individual based on their background information. It means that, based on their background information, they have no reason at all to distinguish between H and T and thus they are totally confused about whether the specific toss will lead to an H or a T.

It is very interesting to note that the same statement $\mathbb{P}(H) = 0.5$ is an informative objective fact in the frequentist viewpoint while it is an uninformative assignment in the Bayesian viewpoint.

To understand the Bayesian interpretation, consider developing a spam filter that classifies incoming emails as spam or regular (there are many statistics/machine learning methods for doing this including, say, logistic regression or classification trees). Suppose you have a trained spam filter and you apply it to a specific incoming email. If the filter outputs a predicted probability of the email being spam as 0.5, you would conclude that the filter has no idea whether this email is spam or regular. This is exactly the Bayesian viewpoint.

3.1.3 Rules of Probability

Let us now look at the rules of the probability when probability is viewed from the Bayesian viewpoint which has nothing with do with frequencies:

1. $\mathbb{P}(A)$ always lies between 0 and 1. The probability of an impossible event is 0 and the probability of a certain event is 1.
2. Product rule: $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B|A) = \mathbb{P}(B)\mathbb{P}(A|B)$.
3. Sum rule: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ for disjoint events A and B.

These rules can be justified in a straightforward way if we use the frequency definition of probability. We are now using a more general form of probability which has nothing to do with frequencies and which are assignments made by a specific user. What is constraining the user to follow the above rules? The following quote by Fisher (1934) asks the same question:

Keynes establishes the laws of addition and multiplication of probabilities, by stating these laws in the form of definitions of the processes of addition and multiplication. The important

step of showing that, when these probabilities have numerical values, "addition" and "multiplication" are so defined, are equivalent to the arithmetical processes ordinarily known by these names, is omitted. The omission is an interesting one, since it shows the difficulty of establishing the laws of mathematical probability, without basing the notion of probability on the concept of frequency, for which these laws are really true, and from which they were originally derived.

It is important to be able to justify the rules of probability for this general form of probability. Otherwise, there will be no principled way of computing probabilities of things we really care about and the whole business will be quite arbitrary.

The following justification for the rules of probability is originally due to the physicist R. T. Cox and is the content of Chapter 1 and Chapter 2 of Jaynes [1]. I will give a sketch of the argument skipping some important technical details. For the full argument, please read Jaynes [1, Chapter 1 and 2].

Let us first remove all restrictions on probabilities and even allow them to take values outside the interval $[0, 1]$. To avoid confusion, let us use the term "plausibilities". We are assigning plausibilities of various events (or propositions) conditional on other events. Let us denote the plausibility of event A conditional on event B by $(A|B)$. Let us first make the assumption that plausibilities take values in the set of real numbers (no restriction now to be in the interval $[0, 1]$) and that a higher value of plausibility represents a greater belief.

3.1.4 Product Rule

Let us first investigate why the product rule should be true. The product rule in terms of probabilities states that

$$\mathbb{P}(AB|C) = \mathbb{P}(B|C)\mathbb{P}(A|BC)$$

Here AB denotes the event $A \cap B$. Should our plausibilities satisfy a similar inequality? Let us first assume that the plausibility $(AB|C)$ should really be determined by the two plausibilities $(B|C)$ and $(A|BC)$. This is basically because the process of deciding that AB is true can be broken down into first deciding whether B is true and then, having accepted B as true, deciding whether A is true. We shall therefore assume that there should be a function F such that

$$(AB|C) = F((B|C), (A|BC)).$$

We also assume that we should use the same function F for all possible events A, B, C (i.e., we are not using one function F for some A, B, C while calculating $(AB|C)$ from $(B|C)$ and $(A|BC)$ and using another function F for different A, B, C). This means in particular that

$$(AB|C) = F((A|C), (B|AC)).$$

It is also reasonable to assume that $F(x, y)$ is monotone increasing in each of its arguments and that it is continuous. If it is not continuous, then a small change in $(B|C)$ (or $(A|BC)$) might lead to a large change in $(AB|C)$ which is undesirable.

Now if we have four events A, B, C, D, we can write

$$(ABC|D) = F((BC|D), (A|BCD)) = F(F((C|D), (B|CD)), (A|BCD)).$$

We can also write

$$(ABC|D) = F((C|D), (AB|CD)) = F((C|D), F((B|CD), (A|BCD))).$$

We shall now make the following important consistency assumption: **If a plausibility can be calculated via two different methods, then both methods should give the same answer**. Clearly if this assumption were violated, then our answer to a plausibility calculation would depend on the specific method chosen to calculate and this would be highly undesirable. This assumption immediately implies that

$$F(F((C|D),(B|CD)),(A|BCD)) = F((C|D),F((B|CD),(A|BCD))).$$

for all A, B, C, D. If the individual plausibilities are arbitrary, we would get the following condition that the function F should satisfy

$$F(F(x,y),z) = F(x,F(y,z)) \qquad \text{for all real numbers } x,y,z.$$

It now turns out the only functions F which satisfy the above equation are of the form

$$F(x,y) = w^{-1}(w(x)w(y))$$

for a positive continuous increasing function w. I will skip this derivation (see Section 2.1, Chapter 2 of Jaynes [1]). We thus have

$$(AB|C) = F((B|C),(A|BC)) = w^{-1}\left(w(B|C)w(A|BC)\right). \tag{8}$$

This is equivalent to

$$w(AB|C) = w(B|C)w(A|BC).$$

Now if we take $B = A$, we get

$$w(A|C) = w(A|C)w(A|AC)$$

The event $A|AC$ can be seen as certainty so we get

$$w(A|C) = w(A|C)w(\text{certainty})$$

for all A and C. This can happen only if

$$w(\text{certainty}) = 1. \tag{9}$$

Also if we take $B = A^c$ in (8), we get

$$w(AA^c|C) = w(A^c|C)w(A|A^cC)$$

$AA^c|C$ and $A|A^cC$ can both be taken to represent impossibility so we get

$$w(\text{impossible}) = w(A^c|C)w(\text{impossible})$$

for all A and C which gives

$$w(\text{impossible}) = 0. \tag{10}$$

(9) and (10), along with the monotonicity of w, imply

$$0 \le w(A|B) \le 1 \qquad \text{for all } A \text{ and } B. \tag{11}$$

We have thus proved that $w(A|B)$ lies always between 0 and 1 (is 0 for impossibility and 1 for certainty) and it satisfies the product rule of probability:

$$w(AB|C) = w(A|C)w(B|AC) = w(B|C)w(A|BC). \tag{12}$$

In other words, if we apply this this function w to our plausibilities, then the resulting assignments satisfy the first two rules of probability.

3.1.5 Sum Rule

Below we shall sketch the argument for the sum rule of probability (the full details can be found in Section 2.2 of Jaynes [1]). For a proposition A, we denote its complement by A^c (i.e., A^c refers to the proposition that A is not true). Suppose that the plausibility $w(A^c|C)$ should be a function of $w(A|C)$:

$$w(A^c|C) = S(w(A|C)).$$

This is intuitively meaningful as the plausibility of A^c should be determined by the plausibility of A. We also assume that we use the same function S for every A, C. This function S maps $[0,1]$ to $[0,1]$ and it should be a self-reciprocal function because $S(S(w(A|C))) = S(w(A^c|C)) = w(A|C)$ i.e., $S(S(x)) = x$ or $S^{-1}(x) = S(x)$. It should also satisfy (by taking A to be certainty) $S(1) = 0$.

There is another condition that S needs to satisfy as a consequence of the fact that $w(A|B)$ satisfies the product rule. For three propositions A, B, C, we have

$$w(AB|C) = w(A|C)w(B|AC) = w(A|C)S(w(B^c|AC)) = w(A|C)S\left(\frac{w(AB^c|C)}{w(A|C)}\right).$$

Switching A and B, we get

$$w(AB|C) = w(B|C)S\left(\frac{w(A^cB|C)}{w(B|C)}\right).$$

We thus have

$$w(A|C)S\left(\frac{w(AB^c|C)}{w(A|C)}\right) = w(B|C)S\left(\frac{w(A^cB|C)}{w(B|C)}\right)$$

for all A, B, C.

Now suppose A and B are such that B^c is contained in A (or equivalently A^c is contained in B). This means that whenever B^c is true, A is also true. So we have $AB^c = B^c$ and $A^cB = A^c$. Thus

$$w(A|C)S\left(\frac{w(B^c|C)}{w(A|C)}\right) = w(B|C)S\left(\frac{w(A^c|C)}{w(B|C)}\right)$$

which is equivalent to

$$w(A|C)S\left(\frac{S(w(B|C))}{w(A|C)}\right) = w(B|C)S\left(\frac{S(w(A|C))}{w(B|C)}\right).$$

The above equation should be true for all A, B, C such that B^c is contained in A. Letting $x = w(A|C)$, $y = w(B|C)$ so that $S(y) = w(B^c|C) \leq w(A|C) = x$, we thus have

$$xS\left(\frac{S(y)}{x}\right) = yS\left(\frac{S(x)}{y}\right) \qquad \text{for all } 0 \leq x, y \leq 1 \text{ with } 0 \leq S(y) \leq x.$$

One can then show that the above condition implies that

$$S(x) = (1 - x^\alpha)^{1/\alpha} \qquad \text{for } x \in [0,1]$$

for some $\alpha > 0$. This argument is somewhat technical and you can read it in Jaynes [1, Section 2.2].

We have thus proved that

$$w(A^c|C) = (1 - w^\alpha(A|C))^{1/\alpha}$$

which is equivalent to

$$w^\alpha(A^c|C) = 1 - w^\alpha(A|C) \quad \text{or} \quad w^\alpha(A^c|C) + w^\alpha(A|C) = 1.$$

It can now be noted that the rules (9), (10), (11) and (12) that $w(A|B)$ satisfies are also satisfied by $w^\alpha(A|B)$. Thus $w^\alpha(A|B)$ satisfies all the three rules

1. $0 \leq w^\alpha(A|B) \leq 1$, $w^\alpha(\text{impossible}) = 0$, and $w^\alpha(\text{certain}) = 1$,
2. $w^\alpha(AB|C) = w^\alpha(A|C)w^\alpha(B|AC) = w^\alpha(B|C)w^\alpha(A|BC)$, and
3. $w^\alpha(A^c|C) + w^\alpha(A|C) = 1$.

We shall therefore denote w^α by $\mathbb{P}$ and call it probability. $\mathbb{P}$ then satisfies the rules:

1. $0 \leq \mathbb{P}(A|B) \leq 1$, $\mathbb{P}(\text{impossible}) = 0$, and $\mathbb{P}(\text{certain}) = 1$,
2. **Product Rule**: $\mathbb{P}(AB|C) = \mathbb{P}(A|C)\mathbb{P}(B|AC) = \mathbb{P}(B|C)\mathbb{P}(A|BC)$, and
3. **Sum Rule**: $\mathbb{P}(A^c|C) + \mathbb{P}(A|C) = 1$.

Note that these usual laws of probability hold because of the need for logical consistency and not because our probability has anything to do with frequencies.

The sum rule of probability is usually stated as

$$\mathbb{P}(A \cup B|C) = \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(AB|C) \tag{13}$$

where $A \cup B$ denotes the proposition "at least one of A and B is true" (Jaynes uses the notation $A + B$ for $A \cup B$). (14) can be derived from the stated rules as

$$\begin{aligned}
\mathbb{P}(A \cup B|C) &= 1 - \mathbb{P}((A \cup B)^c|C) \\
&= 1 - \mathbb{P}(A^c \cap B^c|C) \\
&= 1 - \mathbb{P}(B^c|A^cC)\mathbb{P}(A^c|C) \\
&= 1 - (1 - \mathbb{P}(B|A^cC))\,\mathbb{P}(A^c|C) \\
&= 1 - \mathbb{P}(A^c|C) + \mathbb{P}(B|A^cC)\mathbb{P}(A^c|C) \\
&= \mathbb{P}(A|C) + \mathbb{P}(A^cB|C) \\
&= \mathbb{P}(A|C) + \mathbb{P}(B|C)\mathbb{P}(A^c|BC) \\
&= \mathbb{P}(A|C) + \mathbb{P}(B|C)(1 - \mathbb{P}(A|BC)) \\
&= \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(B|C)\mathbb{P}(A|BC) = \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(AB|C).
\end{aligned}$$

4 Lecture Four

4.1 Recap: Derivation of the Rules of Probability for Subjective Probability

In the last class, we sketched the derivation of the rules of probability from logical consistency **without** relying on any imagined frequency considerations for defining probability. The argument (taken from Jaynes [1, Chapter 2]) proceeded in the following way.

We denoted the "plausibility" of a proposition A given some information in the form of proposition B by $(A \mid B)$. We assumed that these plausibilities take values in the set of real numbers (no restriction to be in the interval $[0, 1]$) and that a higher value of plausibility represents greater belief.

To deduce the usual product rule of probability, we assumed the existence of a continuous coordinate-wise increasing function F of two real variables such that

$$(AB \mid C) = F\left((B \mid C), (A \mid BC)\right)$$

for all A, B, C. This function can then be employed in two different orders to calculate $(ABC \mid D)$ for four propositions A, B, C, D:

$$(ABC \mid D) = F(F((C|D), (B|CD)), (A|BCD))$$
$$(ABC \mid D) = F((C|D), F((B|CD), (A|BCD))).$$

It is therefore natural to assume that the function F should be such that the right hand sides of the above two equations produce the same answer. From this, one gets the condition:

$$F(F(x, y), z) = F(x, F(y, z)) \qquad \text{for all real numbers } x, y, z.$$

As proved in Jaynes [1, Section 2.1], the only functions F which satisfy the above equation are of the form

$$F(x, y) = w^{-1}(w(x)w(y))$$

for a positive continuous increasing function w.

From here, we derived the following:

1. $w(AB \mid C) = w(A \mid C)w(B \mid AC) = w(B \mid C)w(A \mid BC)$.
2. $w(A \mid C)$ always lies between 0 and 1 with $w(\text{impossible}) = 0$ and $w(\text{certain}) = 1$.

In other words, the function w applied to the plausibilities leads to quantities which satisfy the first two rules of probability.

Next the goal is to derive the sum rule. Here we first assume that there exists a function $S : [0, 1] \to [0, 1]$ such that

$$w(A^c \mid C) = S(w(A \mid C))$$

for all A and C. Note that we are working with $w(A \mid C)$ instead of the raw plausibilities $(A \mid C)$. This allows us to use the product rule which has already been derived. To set up the characterizing equation for $S(\cdot)$, consider the setting of Figure 2.

In this setting, there are two different ways of calculating the plausibility of the proposition $R = AB$ in terms of $x := w(A)$ and $y := w(B)$ and the function S. Both these calculations use the product rule. The first method for calculating $w(R) = w(AB)$ is:

$$\begin{aligned} w(AB) &= w(A)w(B \mid A) \\ &= w(A)S\left(w(B^c \mid A)\right) \\ &= w(A)S\left(\frac{w(B^c)}{w(A)}\right) \\ &= w(A)S\left(\frac{S(w(B))}{w(A)}\right) = xS\left(\frac{S(y)}{x}\right). \end{aligned}$$

The second method for calculating $w(R) = w(AB)$ simply switches the roles of A and B in the first method:

$$\begin{aligned} w(AB) &= w(B)w(A \mid B) \\ &= w(B)S\left(w(A^c \mid B)\right) \\ &= w(B)S\left(\frac{w(A^c)}{w(B)}\right) \\ &= w(B)S\left(\frac{S(w(A))}{w(B)}\right) = yS\left(\frac{S(x)}{y}\right). \end{aligned}$$

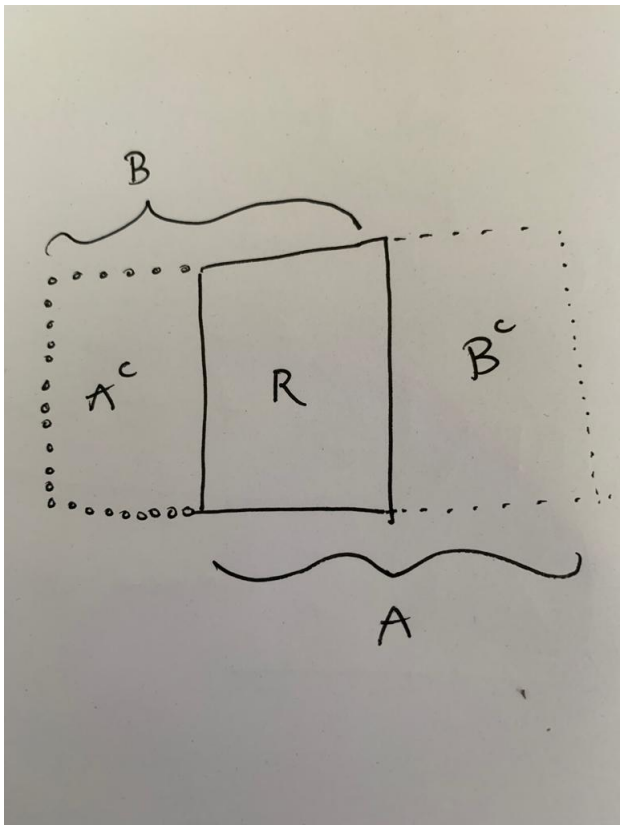

Figure 1: Setting for deriving the Sum Rule

It is therefore natural to assume that S satisfies:

$$xS\left(\frac{S(y)}{x}\right) = yS\left(\frac{S(x)}{y}\right).$$

Recall that here $x = w(A)$ and $y = w(B)$. The setting is such that x and y cannot be completely arbitrary. Indeed because $B^c \subseteq A$, we must have

$$w(B^c) \leq w(A) \quad \text{or equivalently} \quad S(y) \leq x.$$

Our condition on S is therefore

$$xS\left(\frac{S(y)}{x}\right) = yS\left(\frac{S(x)}{y}\right) \qquad \text{for all } 0 \leq x, y \leq 1 \text{ with } 0 \leq S(y) \leq x.$$

It is now proved in Jaynes [1, Section 2.2] that the above condition implies that

$$S(x) = (1 - x^\alpha)^{1/\alpha} \qquad \text{for } x \in [0, 1]$$

for some $\alpha > 0$.

We have thus proved that

$$w(A^c|C) = (1 - w^\alpha(A|C))^{1/\alpha}$$

which is equivalent to

$$w^\alpha(A^c|C) = 1 - w^\alpha(A|C) \quad \text{or} \quad w^\alpha(A^c|C) + w^\alpha(A|C) = 1.$$

It can now be noted that the first two rules that are satisfied by $w(A|B)$ are also satisfied by $w^\alpha(A|B)$. Thus $w^\alpha(A|B)$ satisfies all the three rules

1. $0 \leq w^\alpha(A|B) \leq 1$, $w^\alpha(\text{impossible}) = 0$, and $w^\alpha(\text{certain}) = 1$,
2. $w^\alpha(AB|C) = w^\alpha(A|C)w^\alpha(B|AC) = w^\alpha(B|C)w^\alpha(A|BC)$, and
3. $w^\alpha(A^c|C) + w^\alpha(A|C) = 1$.

Denoting w^α by $\mathbb{P}$, we have

1. $0 \leq \mathbb{P}(A|B) \leq 1$, $\mathbb{P}(\text{impossible}) = 0$, and $\mathbb{P}(\text{certain}) = 1$,
2. **Product Rule**: $\mathbb{P}(AB|C) = \mathbb{P}(A|C)\mathbb{P}(B|AC) = \mathbb{P}(B|C)\mathbb{P}(A|BC)$, and
3. **Sum Rule**: $\mathbb{P}(A^c|C) + \mathbb{P}(A|C) = 1$.

We have thus derived the rules of probability without invoking any relationship between probability and long-run frequency.

To summarize: If we argue in terms of plausibilities but we take care to ensure some natural constraints for logical consistency, then we cannot manipulate the plausibilities arbitrarily but have to reason according to the usual rules of probability after an appropriate transformation (this transformation is given by the function w^α).

The sum rule of probability is usually stated as

$$\mathbb{P}(A \cup B|C) = \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(AB|C) \tag{14}$$

where $A \cup B$ denotes the proposition "at least one of A and B is true" (Jaynes uses the notation $A + B$ for $A \cup B$). (14) can be derived from the stated rules as

$$\begin{aligned}
\mathbb{P}(A \cup B|C) &= 1 - \mathbb{P}((A \cup B)^c|C) \\
&= 1 - \mathbb{P}(A^c \cap B^c|C) \\
&= 1 - \mathbb{P}(B^c|A^cC)\mathbb{P}(A^c|C) \\
&= 1 - (1 - \mathbb{P}(B|A^cC))\,\mathbb{P}(A^c|C) \\
&= 1 - \mathbb{P}(A^c|C) + \mathbb{P}(B|A^cC)\mathbb{P}(A^c|C) \\
&= \mathbb{P}(A|C) + \mathbb{P}(A^cB|C) \\
&= \mathbb{P}(A|C) + \mathbb{P}(B|C)\mathbb{P}(A^c|BC) \\
&= \mathbb{P}(A|C) + \mathbb{P}(B|C)(1 - \mathbb{P}(A|BC)) \\
&= \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(B|C)\mathbb{P}(A|BC) = \mathbb{P}(A|C) + \mathbb{P}(B|C) - \mathbb{P}(AB|C).
\end{aligned}$$

4.2 Probability Assignment

As we discussed in Lecture One, probability theory works according to the two steps: (a) probability assignment, and (b) calculation. The calculation in the second step is based on the rules of probability and we have seen just the rationale behind the rules.

We shall now make some general comments on the probability assignment step. The most important word here is "Information". Probability assignment is always made by some specific individual based on available or assumed information. The term "assumed information" is important here because, quite often, certain aspects of available information might be hard to precisely quantify so one may choose to ignore such aspects in order work with a simpler probability assignment.

To make this first step of probability theory "rigorous", we need precise rules for transforming available/assumed information into probability assignments. The most fundamental

of these rules is known as the **Principle of Indifference** or the **Principle of Insufficient Reason** and it states the following:

If, on background information B, the propositions $A_1, \ldots, A_N$ are mutually exclusive and exhaustive, and B does not favor any one of them over any other, then

$$\mathbb{P}(A_i|B) = \frac{1}{N} \qquad \text{for } i = 1, \ldots, N. \tag{15}$$

In other words, the Principle of Indifference states that if the background information B is "symmetric" among $A_1, \ldots, A_N$, then one should use the probability assignment (15).

For a concrete example illustrating the Principle of Indifference, consider the following setting. suppose we want to assign a probability for a coin toss landing in H. Let us consider the following three kinds of information:

1. **Information** I_1: We don't know anything at all about the coin. We don't even know if it really has two sides H and T or both of its sides are of only one kind (either H or T). In addition, we don't know how exactly it will be tossed.

2. **Information** I_2: We know that it is a "regular" coin and that it has two sides H and T but we don't know how it will be tossed.

3. **Information** I_3: We know that this coin has been tossed a large number of times in the past and it landed heads 70% of the time.

In the first case, the Principle of Indifference applies and we shall assign

$$\mathbb{P}(H|I_1) = 0.5 \tag{16}$$

It is very important to recognize here that (16) is not a statement of frequency. Specifically (16) does not mean that if we toss the coin a large number of times, it will land hands 50% of the time. All we are saying is that we assign the probability of 0.5 for the coin landing heads in this one specific toss that we are reasoning about. We simply do not have any information to make speculations about the behaviour of a large number of tosses. In fact, our information does not even tell us that the coin indeed has the two sides H and T. So it might well be the case the coin tosses will result in $HHHHH\ldots$ or $TTTTT\ldots$. But we shall still assign (16) because our current information I_1 does not allow us to distinguish between H and T.

Now let us come to I_2 which is more informative than I_1. However, even here, there is nothing to distinguish between H and T. So the Principle of Indifference applies again and we assign

$$\mathbb{P}(H|I_2) = 0.5. \tag{17}$$

Again this has nothing to do with frequency. Even though it is a regular coin, it might be tossed in a way to produce more heads than tails. So if an actual experiment were performed, then depending on how the coin tosses were performed, the frequency of heads can be pretty much anything between 0 and 1. Because our probability has nothing to do with frequency, we shall simply assign (17) as our information I_2 is symmetric in H and T. Note that if an experiment is actually performed and the proportion of heads turns out to be different from 0.5, this does not contradict (17) at all. Because (17) is an assignment capturing our state of knowledge I_2 and is perfectly sensible under I_2.

Now let us come to I_3. Here the Principle of Indifference obviously does not apply. If we were using the frequency definition, we would immediately assign $\mathbb{P}(H|I_3) = 0.7$. But

because our probability has nothing to do with frequency, we cannot jump to this assignment immediately. The issue is that here we need to reason about not just one toss but this imminent toss along with all the previous tosses. Specifically, in this situation I_3, we are dealing with random variables $X_1, \ldots, X_N$ corresponding to the previous large number of tosses and the current toss X_{N+1} about which we are reasoning. We need to calculate:

$$\mathbb{P}\left\{X_{N+1} = 1 \middle| \frac{X_1 + \cdots + X_N}{N} = 0.7\right\} \tag{18}$$

To calculate this, we need to make a more basic probability assignment for the joint distribution of $X_1, \ldots, X_N, X_{N+1}$ which allows us to compute $\mathbb{P}(H|I_3)$. If we assume that

$$X_1, \ldots, X_{N+1} \overset{\text{i.i.d}}{\sim} \text{Ber}(0.5), \tag{19}$$

then X_{N+1} will be independent from $X_1, \ldots, X_N$ so that (18) will be 0.5 i.e., the given frequency information is irrelevant under the model. On the other hand, under the more complicated model assumption

$$X_1, \ldots, X_{N+1} \mid \Theta = \theta \overset{\text{i.i.d}}{\sim} \text{Ber}(\theta) \quad \text{and} \quad \Theta \sim \text{Unif}[0, 1],$$

we shall show later that the probability (18) will be very close to 0.7 when N is large.

Therefore, in the third situation, when we actually have frequency information, under the right kind of model, our analysis will lead to the frequency assignment. Thus in this theory of probability, frequency will appear naturally in probability assignments when frequency information is available and relevant to the problem.

Next we shall review standard probability distributions starting with the Hypergeometric Distribution. These standard distributions will be useful for us while making probability assignments.

4.3 Urn Problems: Hypergeometric Distribution

Consider an urn with N balls and assume that R of the N balls are red and the remaining $W := N - R$ are white. Assume that the balls are identical in every other respect.

Suppose we sample n balls from the urn without replacement. What is the probability of seeing exactly r red balls in the sample? The answer, as we shall see, is given by

$$\frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}}$$

where $w := n - r$ is the number of white balls in the sample. This requires $0 \leq r \leq R$ and $0 \leq w \leq W$. If these conditions are not satisfied, the required probability will be zero.

Here is one way of proving this probability statement. Let R_i denote the proposition that the i^{th} draw results in a red ball and let W_i denote the proposition that the i^{th} draw results in a white ball. Let us first consider the probability

$$\mathbb{P}\left\{R_1 \ldots R_r W_{r+1} \ldots W_n\right\}.$$

Using the product rule of probability we can calculate the above probability as

$$\mathbb{P}(R_1)\left[\prod_{i=2}^{r} \mathbb{P}\{R_i \mid R_1 \ldots R_{i-1}\}\right] \mathbb{P}(W_{r+1} \mid R_1 \ldots R_r) \prod_{j=r+2}^{n} \mathbb{P}\{W_j \mid R_1 \ldots R_r W_{r+1} \ldots W_{j-1}\}.$$

By the principle of indifference and the sum rule of probability, we get $\mathbb{P}(R_1) = R/N$,

$$\mathbb{P}\{R_i \mid R_1 \dots R_{i-1}\} = \frac{R-i+1}{N-i+1} \qquad \text{for } i = 2, \dots, r,$$

$$\mathbb{P}(W_{r+1} \mid R_1 \dots, R_r) = \frac{W}{N-r}$$

and

$$\mathbb{P}\{W_j \mid R_1 \dots R_r W_{r+1} \dots W_{j-1}\} = \frac{W-j+r+1}{N-j+1} \qquad \text{for } r+2 \le j \le n.$$

It then follows that

$$\mathbb{P}\left\{R_1 \dots R_r W_{r+1} \dots W_n\right\} = \frac{R(R-1)\dots(R-r+1)W(W-1)\dots(W-w+1)}{N(N-1)\dots(N-n+1)} = \frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}} \frac{1}{\binom{n}{r}}.$$

It can now be checked that, by the same argument, one also has

$$\mathbb{P}\{R_1 W_2 R_3 \dots R_{r+1} W_{r+2} \dots W_n\} = \frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}} \frac{1}{\binom{n}{r}}.$$

More generally, the probability of any specific sequence $RWRRW\dots$ with exactly r reds is given by

$$\frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}} \frac{1}{\binom{n}{r}}.$$

Because the number of such sequences of $R's$ and $W's$ with exactly r $R's$ is $\binom{n}{r}$, we obtain

$$\mathbb{P}\{r \text{ reds in sample of size } n\} = \frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}}.$$

As a function of r, this is known as the Hypergeometric Probability Mass Function. The mean of this distribution can be checked to be nR/N. Thus the average fraction of red balls in the sample of size n will match the fraction of red balls in the urn. We can also compute the most likely value of r i.e., the mode of the hypergeometric distribution. For this, let

$$h(r) = \frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}}.$$

and one can easily calculate that

$$\frac{h(r+1)}{h(r)} = \frac{R-r}{r+1} \frac{n-r}{W-(n-r)+1}$$

so that

$$\frac{h(r+1)}{h(r)} \ge 1 \iff r+1 \le \frac{(n+1)(R+1)}{N+2}.$$

This gives that

$$\lfloor \frac{(n+1)(R+1)}{N+2} \rfloor$$

can be taken to be the mode of the hypergeometric distribution. When n and N are large, the quantity above is approximately nR/N so that the most likely value of r is approximately such that the sample fraction of red balls matches the fraction of red balls in the urn.

See Jaynes [1, Chapter 3] for more calculations in this urn setting.

5 Lecture Five

5.1 The Hypergeometric Distribution

In the last class, we studied the Hypergeometric Distribution in the following urn setting. There is an urn with N balls of which R are red and the remaining $W := N - R$ are white. Assume that the balls are identical in every other respect. We then sample n balls from the urn without replacement. What is the probability of seeing exactly r red balls in the sample? We saw that the answer is given by:

$$\frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}}$$

where $w := n - r$ is the number of white balls in the sample.

We can let X to be the random variable denoting the number of red balls in the drawn sample of size n. Then

$$\mathbb{P}\{X = r\} = \frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}}. \tag{20}$$

This X is said to have the Hypergeometric Distribution with parameters N, R, n.

Let us go over the proof of (20) using notation that is different from last time. For each $i = 1, \ldots, n$, let X_i denote the binary random variable that equals 1 if the i^{th} draw results in a red ball and equals 0 if the i^{th} draw results in a white ball. Then it is easy to see that

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

By the argument given at the end of the last lecture, we have ($\sum_{i=1}^n x_i$ below plays the role of r in (20)):

$$\mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n\} = \frac{\binom{R}{\sum_{i=1}^n x_i}\binom{N}{n-\sum_{i=1}^n x_i}}{\binom{N}{n}} \frac{1}{\binom{n}{\sum_{i=1}^n x_i}}. \tag{21}$$

An important feature of (22) is that the right hand side depends on the individual $x_1, \ldots, x_n$ only through their sum $\sum_{i=1}^n x_i$. This implies that the distribution of $X_1, \ldots, X_n$ is the same as $X_{\pi_1}, \ldots, X_{\pi_n}$ for every permutation $\pi_1, \ldots, \pi_n$ of $1, \ldots, n$:

$$\mathbb{P}\{X_{\pi_1} = x_1, \ldots, X_{\pi_n} = x_n\} = \mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n\} \tag{22}$$

for every $x_1, \ldots, x_n \in \{0, 1\}$.

Random variables having the property (22) are known as **exchangeable**. Thus $X_1, \ldots, X_n$ are exchangeable random variables. Here are some consequences of exchangeability.

1. $X_1, \ldots, X_n$ have identical distributions i.e.,

$$\mathbb{P}\{X_1 = x\} = \mathbb{P}\{X_2 = x\} = \cdots = \mathbb{P}\{X_n = x\} \tag{23}$$

for every $x \in \{0, 1\}$. The reason behind (23) is that for every $i = 2, \ldots, n$,

$$\begin{aligned}\mathbb{P}\{X_i = x\} &= \sum_{x_j, j \neq i} \mathbb{P}\{X_1 = x_1, X_i = x, X_j = x_j \text{ for } j \neq 1, j \neq i\} \\ &= \sum_{x_j, j \neq i} \mathbb{P}\{X_1 = x, X_i = x_1, X_j = x_j \text{ for } j \neq 1, j \neq i\} = \mathbb{P}\{X_1 = x\}\end{aligned}$$

Even though $X_1, \ldots, X_n$ have identical distributions, they are not independent however because

$$\mathbb{P}\{X_2 = 1 \mid X_1 = 1\} = \frac{R-1}{N-1} \text{ and } \mathbb{P}\{X_2 = 1 \mid X_1 = 0\} = \frac{R-1}{N}.$$

In general,

$$\text{i.i.d} \implies \text{exchangeable but exchangeable} \not\Longrightarrow \text{i.i.d.}$$

2. The distribution of every pair (X_i, X_j) for $i \neq j$ is the same i.e.,

$$\mathbb{P}\{X_i = u, X_j = v\} = \mathbb{P}\{X_1 = u, X_2 = v\} \tag{24}$$

for all $u, v \in \{0, 1\}$. This is true because, for example,

$$\begin{aligned}
&\mathbb{P}\{X_3 = u, X_4 = v\} \\
&= \sum_{x_1, x_2, x_5, \ldots, x_n} \mathbb{P}\{X_1 = x_1, X_2 = x_2, X_3 = u, X_4 = v, X_5 = x_5, \ldots, X_n = x_n\} \\
&= \sum_{x_1, x_2, x_5, \ldots, x_n} \mathbb{P}\{X_1 = u, X_2 = v, X_3 = x_1, X_4 = x_2, X_5 = x_5, \ldots, X_n = x_n\} \\
&= \mathbb{P}\{X_1 = u, X_2 = v\}.
\end{aligned}$$

This proves that (X_3, X_4) has the same distribution as (X_1, X_2). The proof for other pairs is similar.

3. The distribution of every k-tuple $(X_{i_1}, \ldots, X_{i_k})$ for any distinct indices $i_1, \ldots, i_k$ from $\{1, \ldots, n\}$ is the same. The proof is similar to that of (23) and (24).

5.1.1 Mean and Variance of the Hypergeometric Distribution

The mean and variance of the random variable X having the Hypergeometric distribution (20) are given by:

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

These formulae can be easily derived using exchangeability of $X_1, \ldots, X_n$ and the fact that $X = X_1 + \cdots + X_n$ as follows:

$$\mathbb{E}X = \mathbb{E}(X_1 + \cdots + X_n) = n\mathbb{E}X_1 = n\mathbb{P}\{X_1 = 1\} = \frac{nR}{N}$$

and

$$\begin{aligned}
\mathbb{E}X^2 &= \mathbb{E}\left(X_1 + \cdots + X_n\right)^2 \\
&= \mathbb{E}\left(X_1^2 + \cdots + X_n^2 + \sum_{i \neq j} X_i X_j\right) \\
&= \mathbb{E}\left(X_1 + \cdots + X_n + \sum_{i \neq j} X_i X_j\right) \\
&= \frac{nR}{N} + n(n-1)\mathbb{E}(X_1 X_2) \\
&= \frac{nR}{N} + n(n-1)\mathbb{P}\{X_1 = 1, X_2 = 1\} \\
&= \frac{nR}{N} + n(n-1)\mathbb{P}\{X_1 = 1\}\mathbb{P}\{X_2 = 1 \mid X_1 = 1\} \\
&= \frac{nR}{N} + n(n-1)\frac{R}{N}\frac{R-1}{N-1}.
\end{aligned}$$

From here, one can deduce that

$$\text{var}(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \frac{nR}{N} + n(n-1)\frac{R}{N}\frac{R-1}{N-1} - \left(\frac{nR}{N}\right)^2 = \frac{nR(N-n)(N-R)}{N^2(N-1)}.$$

For more calculations involving the Hypergeometric Distribution, see Jaynes [1, Chapter 3].

5.2 Inverse Problem

We now address the following question. Suppose we actually draw n balls from the urn and see that r of them are red. What can we then infer about the contents of the original urn? Specifically, using the observed data n and r, what can we say about R and N?

This problem has applications in survey sampling. Suppose we want to know the number of people R in a city that support the republican party. We take a sample of n people out of which r support republican. What can we then say about R?

We shall study this in the following cases.

5.2.1 Case 1: N is known and R is unknown

We need to calculate

$$\mathbb{P}\left\{R = \tilde{R} \mid \text{data}, N\right\}.$$

where data just refers to the observed values of n and r. Here $\tilde{R}$ denotes a specific integer lying between 0 and N.

By the Bayes rule, we have

$$\begin{aligned}
\mathbb{P}\left\{R = \tilde{R} \mid \text{data}, N\right\} &\propto \mathbb{P}\left\{\text{data} \mid R = \tilde{R}, N\right\}\mathbb{P}\{R = \tilde{R} \mid N\} \\
&= \frac{\binom{\tilde{R}}{r}\binom{\tilde{W}}{w}}{\binom{N}{n}}\mathbb{P}\{R = \tilde{R} \mid N\} \propto \binom{\tilde{R}}{r}\binom{\tilde{W}}{w}\mathbb{P}\{R = \tilde{R} \mid N\}
\end{aligned}$$

To proceed further with the calculation, we need to make an assignment for $\mathbb{P}\{R = \tilde{R} \mid N\}$ which we take to be

$$\mathbb{P}\{R = \tilde{R} \mid N\} = \frac{1}{N+1} I\left\{\tilde{R} \in \{0, 1, \ldots, N\}\right\}.$$

This means that we are not expressing any preference for all the potential values $0, 1, \ldots, N$ that R can take. This gives

$$\mathbb{P}\left\{R = \tilde{R} \mid \text{data}, N\right\} = \binom{\tilde{R}}{r}\binom{\tilde{W}}{w}\frac{I\{\tilde{R} \in \{0, 1, \ldots, N\}\}}{C}$$

where C is the constant

$$C := \sum_{\tilde{R}=0}^{N}\binom{\tilde{R}}{r}\binom{\tilde{W}}{w} = \sum_{\tilde{R}=0}^{N}\binom{\tilde{R}}{r}\binom{N-\tilde{R}}{n-r}$$

A standard mathematical fact involving binomial coefficients (often referred to as a Chu-Vandermonde identity; see, for example, equation (9) in `https://en.wikipedia.org/wiki/Binomial_coefficient#Sums_of_the_binomial_coefficients`) now gives

$$C = \sum_{\tilde{R}=0}^{N}\binom{\tilde{R}}{r}\binom{N-\tilde{R}}{n-r} = \binom{N+1}{n+1}.$$

We thus get the following formula for the posterior distribution of R:

$$\mathbb{P}\left\{R = \tilde{R} \mid \text{data}, N\right\} = \binom{\tilde{R}}{r}\binom{\tilde{W}}{w}\frac{I\{\tilde{R} \in \{0, 1, \ldots, N\}\}}{\binom{N+1}{n+1}}.$$

Note how nicely this posterior distribution corresponds to common sense. For example, we automatically get 0 for the posterior when $\tilde{R} < r$ and when $\tilde{W} = N - \tilde{R} < w$. This is because $\binom{\tilde{R}}{r} = 0$ when $\tilde{R} < r$ and $\binom{\tilde{W}}{w} = 0$ when $\tilde{W} < w$. Also when there is no data (i.e., when $n = r = 0$), the posterior becomes equal to the prior.

We can summarize the above posterior distribution via its mean and variance which are given by (see Chapter 6 of the Jaynes book for details behind the answers below)

$$\mathbb{E}\left(R \mid \text{data}, N\right) = \frac{(N+2)(r+1)}{n+2} - 1$$

and

$$\text{var}\left(R \mid \text{data}, N\right) = \frac{p(1-p)}{n+3}(N+2)(N-n)$$

where

$$p := \frac{r+1}{n+2}. \tag{25}$$

Observe that when N (the number of balls in the urn) is large, we can write

$$\mathbb{E}\left(\frac{R}{N} \mid \text{data}, N\right) \approx \frac{r+1}{n+2} = p$$

and

$$\text{Var}\left(\frac{R}{N} \mid \text{data}, N\right) \approx \frac{p(1-p)}{n+3}.$$

A point estimate of $\frac{R}{N}$ can be taken to be p (when N is large) and the uncertainty of this point estimate can be taken to be the posterior standard deviation $\sqrt{\frac{p(1-p)}{n+3}}$. These of course can be calculated from the observed data r and n.

Let us now do a predictive calculation. Let R_{n+1} denote the proposition that the $(n+1)^{th}$ draw leads to a red ball. What is the probability of R_{n+1} given the observed data? This is obtained by

$$\mathbb{P}\left\{R_{n+1} \mid \text{data}, N\right\} = \sum_{\tilde{R}=0}^{N} \mathbb{P}\left\{R_{n+1} \mid R = \tilde{R}, \text{data}, N\right\} \mathbb{P}\left\{R = \tilde{R} \mid \text{data}, N\right\}.$$

Given $R = \tilde{R}$ and the data, it is clear that, just before the $(n+1)^{th}$ draw, the urn contains $\tilde{R} - r$ red balls and $N - r$ total balls. As a result

$$\mathbb{P}\left\{R_{n+1} \mid R = \tilde{R}, \text{data}, N\right\} = \frac{\tilde{R} - r}{N - n}.$$

Thus

$$\begin{aligned}
\mathbb{P}\left\{R_{n+1} \mid \text{data}, N\right\} &= \sum_{\tilde{R}=0}^{N} \frac{\tilde{R} - r}{N - n} \mathbb{P}\left\{R = \tilde{R} \mid \text{data}, N\right\} \\
&= \mathbb{E}\left(\frac{R - r}{N - n} \mid \text{data}, N\right) \\
&= \frac{\mathbb{E}\left(R \mid \text{data}, N\right) - r}{N - n} \\
&= \frac{\frac{(N+2)(r+1)}{n+2} - 1 - r}{N - n} = \frac{r+1}{n+2} = p
\end{aligned}$$

because of (25). This equation is known as the Laplace Rule of Succession (more details on this will be provided later). Observe that when n (the sample size) is large, we have

$$p = \frac{r+1}{n+2} \approx \frac{r}{n}$$

so that $\mathbb{P}\left\{R_{n+1} \mid \text{data}, N\right\}$ is basically equal to the observed fraction of red balls in the sample.

We can write the above formula in slightly different notation. Let $X_1, \ldots, X_{n+1}$ be as before i.e., X_i equals 1 if the i^{th} draw leads to red and 0 if the i^{th} draw leads to white. Then, when n is large,

$$\mathbb{P}\{X_{n+1} = 1 \mid X_1 = x_1, \ldots, X_n = x_n\} \approx \frac{x_1 + \cdots + x_n}{n}.$$

This reveals a connection between probability and observed frequency that naturally appears by a probability calculation.

5.2.2 Case 2: N is unknown and R is known

This situation arises in the Capture-Recapture problem in ecology. Suppose there is a given pond with some fish and we want to estimate the number of fish in the pond. We take a first fish sample of size R. We then color red (or just tag by some label) all the fish in our sample and then let them back into the pond. Now the pond is like a urn with R red fish

and the remaining $W = N - R$ non-red fish. We now take a second sample of size n and observe that r of this second sample of fish are red. Based on knowledge of r, n, R, what can we infer about N?

The relevant calculation now is

$$\begin{aligned}\mathbb{P}\left\{N = \tilde{N} \mid \text{data}, R\right\} &\propto \mathbb{P}\left\{\text{data} \mid N = \tilde{N}, R\right\} \mathbb{P}\{N = \tilde{N} \mid R\} \\ &= \frac{\binom{R}{r}\binom{\tilde{N}-R}{w}}{\binom{\tilde{N}}{n}} \mathbb{P}\{N = \tilde{N} \mid R\} \propto \frac{\binom{\tilde{N}-R}{w}}{\binom{\tilde{N}}{n}} \mathbb{P}\{N = \tilde{N} \mid R\}\end{aligned}$$

To proceed further with the calculation, we need to make an assignment for $\mathbb{P}\{N = \tilde{N} \mid R\}$ which we take to be

$$\mathbb{P}\{N = \tilde{N} \mid R\} = \frac{1}{N_{\max} - R} I\left\{R \leq \tilde{N} < N_{\max}\}\right\}$$

for some large number $N_{\max}$. This gives

$$\mathbb{P}\left\{N = \tilde{N} \mid \text{data}, R\right\} \propto \frac{\binom{\tilde{N}-R}{w}}{\binom{\tilde{N}}{n}} I\left\{\tilde{N} < N_{\max}\right\}.$$

The presence of the binomial coefficient means that the posterior probability above is zero unless $\tilde{N} \geq R + w$. This posterior will give us everything that we need to know about $\tilde{N}$ after observing the data. If the posterior depends on $N_{\max}$, we need to be careful with the results as it would mean that the data is not very informative for N. This would be the case for example if $r = 0$.

5.2.3 Case 3: Both N and R are unknown

In this case, we need to place a prior for

$$\mathbb{P}(N = \tilde{N}, R = \tilde{R}).$$

If we assume that $R \mid N$ is uniform on $\{0, \ldots, N\}$, we would get

$$\mathbb{P}(N = \tilde{N}, R = \tilde{R}) = \mathbb{P}(N = \tilde{N}) \frac{I\{0 \leq \tilde{R} \leq \tilde{N}\}}{\tilde{N} + 1}.$$

The posterior then becomes

$$\mathbb{P}\left(N = \tilde{N}, R = \tilde{R} \mid \text{data}\right) \propto \frac{\binom{\tilde{R}}{r}\binom{\tilde{W}}{w}}{\binom{\tilde{N}}{n}} \frac{I\{0 \leq \tilde{R} \leq \tilde{N}\}}{\tilde{N} + 1} \mathbb{P}\{N = \tilde{N}\}$$

In this case, we can do useful inference on R/N but we cannot learn anything nontrivial from the data about N. To see this, calculate the marginal posterior probability of N and show that

$$\mathbb{P}\left(N = \tilde{N} \mid \text{data}\right) \propto \mathbb{P}(N = \tilde{N}) I\{\tilde{N} \geq n\}.$$

This means that the posterior for N is just the prior truncated to the set $\{n, n+1, \ldots\}$ so we basically don't learn anything about N from the sample other than the fact that it is at least n. Nontrivial inference about N is only possible if we R and N are known to be linked in some manner and we use a prior reflecting that link.

To learn more about inference of the parameters of the hypergeometric distribution, see Jaynes [1, Chapter 6].

6 Lecture Six

6.1 Random Variables

We shall go over the Binomial and Negative Binomial distributions today. It will be convenient to use the language of random variables. The term "random variable" can be used to describe any varying quantity (taking real values) about which we are uncertain about. Many real-life quantities such as (a) The average temperature in Berkeley tomorrow, (b) The height of the tallest 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. The term "random" in the phrase "random variable" refers to the uncertainty of a specific individual about the specific value that will be taken by the variable. Note, in particular, that variable may not be intrinsically random but it is random from the point of view of a specific individual because of their uncertainty. For example, it is perfectly fine for me to treat Joe Biden's current height as a random variable even though it is actually non-random.

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 infinite 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 $a_1, a_2, \ldots$ that it takes together with a specification of the probabilities:

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

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 takes a continuous set of values, its distribution is often described by a function called the *probability density function* (pdf). We shall formally define this later.

6.1.1 Independence of Random Variables

We say that random variables $X_1, \ldots, X_n$ are independent if, for every subset $S \subseteq \{1, \ldots k\}$, conditioning on any proposition involving $X_i, i \notin S$ does not change the probability of any proposition involving $X_i, i \in S$. From here one can easily derive properties of independence such as

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

for all possible choices of $A_1, \ldots, A_k$.

Independence can be a subtle concept in modeling. Suppose I am uncertain about $X_1 :=$ Joe Biden's height and $X_2 :=$ Donald Trump's height. Would it be a reasonable for me to assume that X_1 and X_2 are independent?

6.2 Common Discrete Distributions

6.2.1 Bernoulli $Ber(p)$ 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?

6.2.2 Binomial $Bin(n,p)$ 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 $0, 1, \ldots, n$ with pmf given by

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

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

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

The binomial distribution arises in basically two contexts:

1. **Approximation of the hypergeometric distribution**: In the last class, we looked at the hypergeometric distribution which corresponds to the probabilities

$$\begin{aligned}\frac{\binom{R}{r}\binom{W}{w}}{\binom{N}{n}} &= \binom{n}{r} \frac{R(R-1)\ldots(R-r+1)W(W-1)\ldots(W-w+1)}{N(N-1)\ldots(N-n+1)} \\ &= \binom{n}{r} \frac{R(R-1)\ldots(R-r+1)}{N(N-1)\ldots(N-r+1)} \frac{W(W-1)\ldots(W-w+1)}{(N-r)(N-r-1)\ldots(N-n+1)}.\end{aligned}$$

   This arises as the probability of seeing exactly r red balls when n ($= r + w$) balls are drawn without replacement from an urn with R red balls and $N - R$ white balls. When N is much larger than n, we can write

$$\frac{R-i}{N-i} = \frac{R/N - i/N}{1 - i/N} \approx \frac{R}{N}$$

   and

$$\frac{W-j}{(N-r-j)} = \frac{W/N - j/N}{1-(r+j)/N} \approx \frac{W}{N}.$$

   As a result, the hypergeometric probability simplifies to

$$\binom{n}{r} \left(\frac{R}{N}\right)^r \left(\frac{W}{N}\right)^w$$

   which corresponds to the binomial distribution with parameters n and $p := R/N$.

2. **Number of Successes in Repeated Trials**: Suppose

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

   where each $X_i \sim \text{Ber}(p)$ and $X_1, \ldots, X_n$ are independent. Then it can be checked that $X \sim \text{Bin}(n,p)$.

Example 6.1 (Fairness testing)**.** *Suppose a coin is tossed 12 times leading to the outcome: TTTTHTHTTTTH (this has 3 heads and 9 tails). What is your assessment of the fairness of the coin?*

For the usual frequentist answer to this question, we assume that the observed sequence of outcomes are the realization of random variables $X_1, \ldots, X_n$ (with $n = 12$) that are independently distributed according to the Bin(n, p) distribution for some unknown p. We need to test the (null) hypothesis that $p = 0.5$ against, say, the alternative $p < 0.5$. This can be done by calculating the p-value which is the probability (under the assumption $p = 0.5$) of getting 3 or lower heads. The distribution of the number of heads under the null distribution is Bin$(n, 0.5)$ so the p-value is

$$\left(\binom{12}{3} + \binom{12}{2} + \binom{12}{1} + \binom{12}{0}\right) \frac{1}{2^{12}} = \frac{299}{4096} = 0.073 = 7.3\%$$

which does not lead to a rejection of the null hypothesis at the usual 5% level.

6.2.3 Negative Binomial NB(n, p) distribution

Let X denote the number of tosses (of a coin with probability of heads p) required to get the k^{th} head. What is the probability distribution of X?

The distribution of X is 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 $NB(k, p)$).

Example 6.2 (Fairness Testing (continued))**.** *Let us get back to the fairness testing problem in Example 6.1 where a coin was tossed 12 times leading to the outcome: TTTTHTHTTTTH (this has 3 heads and 9 tails). In our previous p-value calculation, we implicitly assumed that the experiment consisted of tossing the coin 12 times where 12 was a priori chosen by the coin tosser. Consider now the alternative scenario where the coin tosser wanted to toss the coin until the point where 3 heads are observed. Now for the same outcome, the p-value will change. Indeed now the random variable of interest will become N = number of tosses and the p-value will equal the probability of needing to toss the coin 12 or more times to get the 3 heads (assuming fairness). This is calculated using the negative binomial distribution as:*

$$1 - \sum_{n=3}^{11} \binom{n-1}{2} 2^{-n} = \frac{134}{4096} = 0.0327 = 3.27\%$$

and this leads to rejection of the null hypothesis at the 5% level.

Note that the "likelihood function" is the same function $p^3(1-p)^9$ whether the sample size was predetermined or whether the coin was tossed till 3 heads are observed. But the procedure obtained for testing $p = 0.5$ has changed from the binomial to the negative binomial case. This means that p-valued based frequentist inference violates the Likelihood Principle (the likelihood principle states that "all the evidence in a sample relevant to model parameters is

contained in the likelihood function"). Here is a story from the wikipedia article on the "Likelihood Principle" (see `https://en.wikipedia.org/wiki/Likelihood_principle`) which puts an interesting context to these numbers:

Suppose a number of scientists are assessing the probability of a certain outcome (which we shall call 'success') in experimental trials. Conventional wisdom suggests that if there is no bias towards success or failure then the success probability would be one half. Adam, a scientist, conducted 12 trials and obtains 3 successes and 9 failures. One of those successes was the 12th and last observation. Then Adam left the lab.

Bill, Adam's boss in the same lab, continued Adam's work and published Adam's results, along with a significance test. He tested the null hypothesis that θ, the success probability, is equal to a half, versus $\theta < 0.5$. The probability that out of 12 trials, 3 or fewer (i.e. more extreme) were successes, if H_0 is true, is 7.3%. Thus the null hypothesis is not rejected at the 5% significance level.

*Adam actually stopped immediately after 3 successes, because his boss Bill had instructed him to do so. After the publication of the statistical analysis by Bill, Adam realizes that he has missed a **later instruction** from Bill to instead conduct 12 trials, and that Bill's paper is based on this second instruction. Adam is very glad that he got his 3 successes after exactly 12 trials, and explains to his friend Charlotte that by coincidence he executed the second instruction. But Charlotte then explains to Adam that the p-value should now be changed to 3.27% and the result becomes significant at the 5% level. Adam is astonished to hear this.*

For more comments on the violation of the likelihood principle by p-values, read MacKay [5, Section 37.2].

To contrast with the above p-value based analysis, let us look at a Bayesian/probability theory approach to this testing problem. The goal is to calculate:

$$\mathbb{P}\{\text{fairness} \mid \text{data}\}$$

where data refers to $TTTTHTHTTTTH$. By the Bayes rule, we can write

$$\mathbb{P}\{\text{fairness} \mid \text{data}\} = \frac{\mathbb{P}\{\text{data} \mid \text{fairness}\}\mathbb{P}\{\text{fairness}\}}{\mathbb{P}\{\text{data} \mid \text{fairness}\}\mathbb{P}\{\text{fairness}\} + \mathbb{P}\{\text{data} \mid \text{not fair}\}\mathbb{P}\{\text{not fair}\}}$$

We clearly have

$$\mathbb{P}\{\text{data} \mid \text{fairness}\} = 2^{-n}.$$

What assignments do we use for

$$\mathbb{P}\{\text{fairness}\}, \quad \mathbb{P}\{\text{not fair}\} \quad \text{and} \quad \mathbb{P}\{\text{data} \mid \text{not fair}\}?$$

For concreteness, let us assume

$$\mathbb{P}\{\text{fairness}\} = 0.5 \quad \text{and} \quad \mathbb{P}\{\text{not fair}\} = 0.5. \tag{26}$$

This is actually a very strong assumption in favor of fairness because a coin can be not fair in many many variety of ways. So to assume that the probability of fairness is the same as the combined probability of the many variety of ways in which the coin can be non-fair seems quite strong.

Let us now come to $\mathbb{P}\{\text{data} \mid \text{not fair}\}$. If the coin is not fair, we can assume that it has a heads probability of p and that the coin tosses are still independent. We can then write

$$\mathbb{P}\{\text{data} \mid \text{not fair}\} = \int_0^1 \mathbb{P}\{\text{data} \mid \text{not fair}, p\} f_{p|\text{not fair}}(p)dp = \int_0^1 p^3(1-p)^9 f_{p|\text{not fair}}(p)dp.$$

To proceed further, we need to assign $f_{p|\text{not fair}}(p)$. One concrete assumption might be that

$$f_{p|\text{not fair}}(p) = 1 \qquad \text{for every } p \in [0,1]. \tag{27}$$

This corresponds to the assumption that, under the alternative (not fair), p has the uniform distribution on $[0,1]$. Then (using an online integrator)

$$\mathbb{P}\{\text{data} \mid \text{not fair}\} = \int_0^1 p^3(1-p)^9 dp = \frac{1}{2860}.$$

We then get

$$\mathbb{P}\{\text{fairness} \mid \text{data}\} = \frac{2^{-12} * 0.5}{2^{-12} * 0.5 + \frac{1}{2860} * 0.5} = 0.4111558.$$

Note that this Bayesian probability calculation does not depend at all on whether the number of tosses ($n = 12$) was decided a priori or whether it was decided to toss until getting 3 heads. It is the same for both those cases.

Also note that the Bayesian approach (based on (26) and (27)) is only slightly supporting the alternative hypothesis (roughly 60% to the null 40%) while the frequentist p-values are fairly small indicating more evidence for the alternative. This discrepancy also persists when the sample size is large. Consider the following example.

Example 6.3. *In a certain city, 49581 boys and 48870 girls have been born over a certain time period (note* $49581/(49581 + 48870) = 0.5036109$*). Assuming that the number of male births is binomially distributed with parameters* $n = 49581 + 48870 = 98451$ *and* p*, test the hypothesis* $H_0 : p = 0.5$*.*

The usual frequentist p-value is:

$$\mathbb{P}\left\{Bin(n, 0.5) \geq 49581\right\} \approx 0.01163$$

which is fairly small.

On the other hand, the Bayesian method above with the priors (26) *and* (27) *gives*

$$\mathbb{P}\left\{p = 0.5 \mid data\right\} = \frac{2^{-n} * 0.5}{2^{-n} * 0.5 + B(x+1, n-x+1) * 0.5} = 0.950523.$$

Here $x = 49581$*,* $n = 98451$ *and* $B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp$ *is the Beta function.*

Thus the Bayesian method gives a high probability to the null while the frequentist method will reject the null hypothesis. The reason why the Bayesian method is so supportive of the null hypothesis is that the prior choice (26) *gives strong support to* $p = 0.5$ *over nearby values of* p *(such as* $p \in (0.49, 0.51)$*).*

This discrepancy between the Bayesian and Frequentist solutions in this problem is referred to as the Jeffreys-Lindley paradox (see `https://en.wikipedia.org/wiki/Lindley%27s_paradox`*).*

7 Lecture Seven

7.1 Geometric Distribution

The Geometric distribution is a special case of the Negative Binomial distribution for $k = 1$. It corresponds to the number of independent tosses (of a coin with probability of heads

p) required to get the first head. Formally, 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 $1, 2, \ldots$ with the probabilities:

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

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

$$\mathbb{P}\{X > m+n | X > n\} = \mathbb{P}\{X > m\}. \tag{28}$$

This is easy to check as $\mathbb{P}\{X > m\} = (1-p)^m$. It is also interesting that the Geometric distribution is the only distribution on $\{1, 2, \ldots\}$ which satisfies the memorylessness property (28). To see this, suppose that X is a random variable satisfying (28) which takes values in $\{1, 2, \ldots\}$. Let $G(m) := \mathbb{P}\{X > m\}$ for $m = 1, 2, \ldots$. Then (28) 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 $p = \mathbb{P}\{X = 1\}$, then

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

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

7.2 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 is small.

To intuitively see why this is true, just see that

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

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

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

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

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, it is common to use the Poisson distribution to model 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 etc. Can you justify why the Poisson distribution might be appropriate for these random variables?

7.3 Continuous Random Variables

Continuous random variables are random variables that potentially take a continuous set of values. The distribution of a continuous random variable X 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 X 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 pdf 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 the pdf $f(x)$ 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 for probabilities. This is because usually (as long as f is continuous at x):

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

If X is a continuous random variable with density (pdf) f, the expectation of $g(X)$ is defined as

$$\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)dx. \tag{29}$$

We shall next look at some standard Continuous Distributions.

7.4 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 $U \sim U[0, 1]$.

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:

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

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

7.5 The Gaussian or Normal Distribution

A random variable X has the Gaussian or 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 $\phi(\cdot)$:

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

The following is the reason for the presence of the factor $\sqrt{2\pi}$ above:

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{2\pi}. \tag{30}$$

To see why (30) is true, note that

$$\begin{aligned} \left(\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2}\right) dx\right)^2 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2+y^2}{2}\right) dxdy \\ &= \int_0^{\infty} \exp\left(-\frac{r^2}{2}\right)(2\pi r)dr = 2\pi \int_0^{\infty} e^{-z}dz = 2\pi. \end{aligned}$$

From the above (and by a change of variable), we can derive

$$\int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \sqrt{2\pi\sigma^2}.$$

If $X \sim N(\mu, \sigma^2)$, then $\mathbb{E}(X) = \mu$ and $Var(X) = \sigma^2$.

7.5.1 The Gauss Derivation of the Normal Distribution

We shall next look at the Gauss derivation of the normal distribution. Gauss applied the normal distribution in the context of data analysis. The basic question that Gauss addressed is the following: Suppose we take measurements $x_1, \ldots, x_n$ on some physical quantity θ. When is $\bar{x} := (x_1 + \cdots + x_n)/n$ the right estimate for θ? Before the work of Gauss, some prominent mathematicians had doubts about the use of $\bar{x}$ to estimate θ. Jaynes (see Jaynes [1, Section 7.4]) writes that Euler thought that combining many observations would make their errors multiply instead of canceling. As is clear from quote below (taken from Jaynes [1, Page 204]), Daniel Bernoulli thought that taking the average of observations amounts to assuming that the individuals errors in the observations are uniformly distributed and that assuming that the errors are uniformly distributed contradicts common sense:

Now is it not self-evident that the hits must be assumed to be thicker and more numerous on any given band the nearer this is to the mark? If all the places on the vertical plane, whatever their distance from the mark, were equally liable to be hit, the most skillful shot would have no advantage over a blind man. That, however, is the tacit assertion of those who use the common rule (the arithmetic mean) in estimating the value of various discrepant observations, when they treat them all indiscriminately.

The quote above (by Daniel Bernoulli) is in the context of an archer shooting at a vertical line drawn on a target and contemplating on the number of shoots landing on vertical bands on either side of the vertical line.

Gauss showed that taking the average of the distributions is the right way of estimating θ when the errors have the Gaussian distribution. Gauss first assumed that the errors have a distribution f. More specifically, assume that

$$x_i = \theta + \epsilon_i$$

for $i = 1, \ldots, n$ with

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} f$$

for a density f. The maximum likelihood estimator of θ is then given by the maximizer $\hat{\theta}$ of

$$\sum_{i=1}^{n} \log f(x_i - \theta).$$

Letting $g(u) = \log f(u)$, we can say that $\hat{\theta}$ maximizes

$$\sum_{i=1}^{n} g(x_i - \theta)$$

which means (assuming g is smooth)

$$\sum_{i=1}^{n} g'(x_i - \hat{\theta}) = 0.$$

Gauss asked for what density f is it true that the maximum likelihood estimator $\hat{\theta}$ equals the mean $\bar{x}$ for every dataset $x_1, \ldots, x_n$. More precisely, for what f (or equivalently g) do we have

$$\sum_{i=1}^{n} g'(x_i - \bar{x}) = 0 \qquad \text{for every } n \geq 1 \text{ and } x_1, \ldots, x_n.$$

Gauss showed that this equation leads to g' being the linear function:

$$g'(u) = au \tag{31}$$

for some $a \in \mathbb{R}$. This means $g(u) = au^2/2 + b$ so that

$$f(u) = \exp\left(\frac{au^2}{2} + b\right).$$

For f to be a density over $(-\infty, \infty)$, we need $a < 0$ in which case f will be normal:

$$f(u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{u^2}{2\sigma^2}\right)$$

for some $\sigma > 0$. The density of the observations $x_1, \ldots, x_n$ is then

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

Gauss therefore showed that the only distribution on the errors which leads to maximum likelihood estimates being averages is the Normal. This accounts for the popularity (since Gauss) of normal error assumptions in data analysis.

Here is the argument for (31). If we take $n = 2$, we get

$$g'((x_1 - x_2)/2) = -g'((x_2 - x_1)/2) \qquad \text{for every } x_1 \text{ and } x_2$$

which means $g'(0) = 0$ and $g'(-x) = -g'(x)$. Now taking

$$x_1 = nu \quad \text{and} \quad x_2 = \cdots = x_n = 0,$$

for some u (so that $\bar{x} = u$), we get

$$g'((n-1)u) + (n-1)g'(-u) = 0$$

which gives (combining with $g'(-x) = -g'(x)$)

$$g'((n-1)u) = (n-1)g'(u).$$

Taking $m = n - 1$, we have proved that

$$g'(mu) = mg'(u) \qquad \text{for every } u \in \mathbb{R} \text{ and } m \geq 0.$$

This can also be written as (replacing u by u/m) $g'(u/m) = g'(u)/m$ and thus we have

$$g'(\frac{n}{m}u) = \frac{n}{m}g'(u)$$

for every $n, m \geq 1$ which is same as $g'(ru) = rg'(u)$ for every positive rational r and real u. If we now assume that g' is continuous, we obtain $g'(uv) = vg'(u)$ for every $v > 0$ and u which gives $g'(u) = ug'(1)$. This proves (31) with $a = g'(1)$. Something can still be said without continuity (see the wiki article on the Cauchy Functional Equation `https://en.wikipedia.org/wiki/Cauchy%27s_functional_equation`).

For more comments on Gauss's derivation of the normal distribution, see Jaynes [1, Section 7.4].

8 Lecture Eight

8.1 The Normal Distribution as an Approximation to the Binomial Distribution

The normal distribution shows up as an approximation to the Binomial Distribution and, in fact, this was the way the normal density was first discovered by De Moivre. We shall go over this today. Fix $n \geq 1$ and $p \in (0,1)$. Suppose $X \sim \text{Bin}(n,p)$. Then

$$\mathbb{P}\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \qquad \text{for } k = 0, 1, \ldots, n \tag{32}$$

In what sense do the above probabilities resemble the normal density? To understand this, the first step is to approximate the factorials using the Stirling Approximation. A brief overview of Stirling's approximation is discussed next.

8.1.1 Stirling Approximation

The Stirling Approximation is an approximation for $n!$ that is quite accurate even for small values of n. It states that

$$n! \sim \left(\frac{n}{e}\right)^n \sqrt{2\pi n}.$$

The Stirling Approximation is quite accurate even for small n (even for $n = 1$) as can be checked by evaluating the right hand side and comparing it to $n!$. The accuracy of approximation can be gauged from the bound:

$$\exp\left(\frac{1}{12n+1}\right) < \frac{n!}{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}} < \exp\left(\frac{1}{12n}\right).$$

Here is a heuristic justification for the Stirling Approximation using the Laplace Method for approximating integrals. We start with the basic formula

$$n! = \int_0^\infty x^n e^{-x} dx$$

which can be proved by induction over n. Rewrite the integral as

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

The change of variable $x = ny$ gives

$$n! = n^n \cdot n \int_0^\infty \exp\left(n(\log y - y)\right) dy.$$

The function $y \mapsto \log y - y$ attains its maximum value of -1 at $y = 1$. Because of the presence of n in the exponent, the integral will be dominated by the points y which are close to 1 (at least for large n). We shall therefore use the second order Taylor expansion:

$$g(y) := \log y - y \approx g(1) + g'(1)(y-1) + \frac{1}{2} g''(1)(y-1)^2 \approx -1 - \frac{1}{2}(y-1)^2$$

in the exponent to get

$$\begin{aligned} n! &\approx n^n \cdot n \int_0^\infty \exp\left(-n - \frac{n}{2}(y-1)^2\right) dy \\ &= n\left(\frac{n}{e}\right)^n \int_0^\infty \exp\left(-\frac{n}{2}(y-1)^2\right) dy \\ &\approx n\left(\frac{n}{e}\right)^n \int_{-\infty}^\infty \exp\left(-\frac{n}{2}(y-1)^2\right) dy = n\left(\frac{n}{e}\right)^n \sqrt{\frac{2\pi}{n}} = \left(\frac{n}{e}\right)^n \sqrt{2\pi n} \end{aligned}$$

which is exactly the Stirling Approximation.

8.1.2 Entropy Approximation of $Bin(n,p)$

Let us now get back to the problem of approximating the binomial probabilities (32). A reference for these calculations is Sinai [2, Chapter 3] (available for free through the Berkeley Library website). Using the Stirling approximation

$$m! \sim \left(\frac{m}{e}\right)^m \sqrt{2\pi m}$$

for each of the factorials in (32), we get

$$\begin{aligned} \mathbb{P}\{X=k\} &\approx \frac{\left(\frac{n}{e}\right)^n \sqrt{2\pi n}}{\left(\frac{k}{e}\right)^k \sqrt{2\pi k}\left(\frac{n-k}{e}\right)^{n-k}\sqrt{2\pi(n-k)}} p^k (1-p)^{n-k} \\ &= \frac{1}{\sqrt{2\pi n \frac{k}{n}\frac{n-k}{n}}} \left(\frac{k}{n}\right)^{-k} \left(\frac{n-k}{n}\right)^{-(n-k)} p^k (1-p)^{n-k} \\ &= \frac{1}{\sqrt{2\pi n f(1-f)}} \exp\left[-n\left(f\log\frac{f}{p} + (1-f)\log\frac{1-f}{1-p}\right)\right] \qquad \text{where } f := \frac{k}{n} \end{aligned}$$

Note that $f = k/n$ denotes the fraction of heads whose probability we are calculating. Using the notation

$$D(f\|p) := f\log\frac{f}{p} + (1-f)\log\frac{1-f}{1-p},$$

we have

$$\mathbb{P}\{Bin(n,p)=k\} \approx \frac{1}{\sqrt{2\pi n f(1-f)}} \exp\left[-nD(f\|p)\right]. \tag{33}$$

The quantity $D(f\|p)$ is known variously as either the **Relative Entropy** of $(f, 1-f)$ with respect to $(p, 1-p)$ or as the **Kullback-Leibler divergence** between $(f, 1-f)$ and $(p, 1-p)$. We shall refer to (33) as the **Entropy Approximation to** $Bin(n,p)$. The relative entropy $D(f\|p)$ has the following two important properties:

1. **Nonnegativity**: $D(f\|p)$ is always nonnegative. This is basically a consequence of the elementary inequality $\log x \le x - 1$ because

$$\begin{aligned} D(f\|p) &= f\log\frac{f}{p} + (1-f)\log\frac{1-f}{1-p} \\ &= -f\log\frac{p}{f} - (1-f)\log\frac{1-p}{1-f} \ge -f\left(\frac{p}{f}-1\right) - (1-f)\left(\frac{1-p}{1-f}-1\right) = 0. \end{aligned}$$

2. **Zero if and only if** $f = p$: This basically follows from the above argument and the fact that $\log x = x - 1$ if and only if $x = 1$.

The quantity $D(f\|p)$ is often seen as a measure of discrepancy or distance or divergence between the two discrete probability distributions $(f, 1-f)$ and $(p, 1-p)$.

The entropy approximation (33) can be rewritten in the following way. The proportion f represents the empirical proportion of heads while p represents the theoretical (or true) proportion of heads. Thus

$$\mathbb{P}\{\text{Empirical Proportion of Heads and Tails is } (f, 1-f)\} \sim \frac{\exp\left[-nD(f\|p)\right]}{\sqrt{2\pi nf(1-f)}}$$

which shows clearly how the probability decays the further $(f, 1-f)$ moves from $(p, 1-p)$ as measured by the Kullback-Leibler divergence. The subject "Large Deviations Theory" in Probability extends such probability facts to more complicated scenarios.

8.1.3 Normal Approximation of $Bin(n, p)$

To obtain the normal approximation for the Binomial, we approximate

$$G(f) := D(f\|p) = f \log \frac{f}{p} + (1-f) \log \frac{1-f}{1-p}$$

by its Taylor expansion around $f = p$:

$$G(f) = G(p) + G'(p)(f-p) + \frac{1}{2}G''(p)(f-p)^2 + \frac{1}{6}G'''(g)(f-p)^3$$

for some g between f and p. One can directly verify (by calculating derivatives of G) that

$$G(p) = 0 \quad G'(p) = 0 \quad G''(p) = \frac{1}{p(1-p)} \quad G'''(g) = \frac{2g-1}{g^2(1-g)^2}.$$

As a result

$$G(f) = \frac{(f-p)^2}{2p(1-p)} + \frac{2g-1}{6g^2(1-g)^2}(f-p)^3.$$

Plugging this in the formula for $\mathbb{P}\{Bin(n, p) = k\}$, we obtain

$$\mathbb{P}\{Bin(n, p) = k\} \sim \frac{1}{\sqrt{2\pi nf(1-f)}} \exp\left(-\frac{(n-p)^2}{2p(1-p)}\right) \exp\left(\frac{n(2g-1)(f-p)^3}{6g^2(1-g)^2}\right).$$

If the third term above is close to one, then we can drop it which will lead to the normal approximation for $\mathbb{P}\{Bin(n, p) = k\}$. In order to do so, we need

$$\text{Remainder} := \left|\frac{n(2g-1)(f-p)^3}{6g^2(1-g)^2}\right|$$

to be small. Because $|2g-1| \leq 1$ (as g lies between 0 and 1) and $g \geq \min(f, p)$ and $1-g \geq \min(1-p, 1-f)$, we can write

$$\text{Remainder} \leq \frac{n|f-p|^3}{6\left(\min(f,p)\right)^2\left(\min(1-f,1-p)\right)^2}.$$

If p is away from 0 and 1 and $n|f-p|^3$ is small, then the above quantity will be small when n is large. In this situation, we can ignore the remainder term to obtain the approximation:

$$\mathbb{P}\{Bin(n, p) = k\} \sim \frac{1}{\sqrt{2\pi nf(1-f)}} \exp\left(-\frac{(n-p)^2}{2p(1-p)}\right).$$

Also in the case when p is away from 0 and 1 and when $n|f-p|^3$ is small, we have f/p is close to 1 for large n. We can thus replace f by p in the multiplicative term to obtain

$$\begin{aligned}\mathbb{P}\{Bin(n,p)=k\} &\sim \frac{1}{\sqrt{2\pi np(1-p)}}\exp\left(-\frac{(n-p)^2}{2p(1-p)}\right)\\ &= \frac{1}{\sqrt{2\pi np(1-p)}}\exp\left(-\frac{(k-np)^2}{2np(1-p)}\right).\end{aligned}$$

The above is the normal density with mean np and variance $np(1-p)$ evaluated at k. We thus have

$$\mathbb{P}\{Bin(n,p)=k\} \approx \phi(k; np, np(1-p)) \tag{34}$$

where $\phi(x;\mu,\sigma^2)$ denotes the normal density with mean μ and variance σ^2 evaluated at x.

(35) is the normal approximation for $\mathbb{P}\{Bin(n,p)=k\}$. Note that this requires p to be away from 0 and 1 and $n|(k/n)-p|^3$ to small. If these conditions are violated, the normal approximation will not be accurate. The Entropy Approximation, on the other hand, is accurate for a much larger range of k (it is accurate as long as the Stirling Approximation is accurate for $n-k$ and k and the Stirling approximation is quite accurate even for small integers).

For a concrete example, consider the following two situations:

1. Suppose $n=3000, k=2500, p=0.5$. Here $f=k/n=5/6$ which is quite far from p. For example, $n|f-p|^3$ is quite large. One can then verify on the computer that:

   $$\mathbb{P}\{Bin(n,p)=k\} \approx 1.7\times 10^{-318} \quad \text{and} \quad \phi(k;np,np(1-p)) \approx 4.3\times 10^{-292}$$

   Thus the normal approximation is off by many orders of magnitude. On the other hand, the entropy approximation gives

   $$\frac{\exp\left[-nD(f\|p)\right]}{\sqrt{2\pi nf(1-f)}} \approx 1.265\times 10^{-318}$$

   which is quite close to $\mathbb{P}\{Bin(n,p)=k\}$.

2. Suppose $n=3000, k=1525, p=0.5$. Here $f=0.5083$ which is quite close to p. Also $n*|f-p|^3=0.00173$ is quite small. Then

   $$\mathbb{P}\{Bin(n,p)=k\} \approx 0.0096037 \quad \text{and} \quad \phi(k;np,np(1-p)) \approx 0.0096033$$

   so the normal approximation is very accurate. The entropy approximation here is:

   $$\frac{\exp\left[-nD(f\|p)\right]}{\sqrt{2\pi nf(1-f)}} \approx 0.0096032$$

In both these situations, the entropy approximation is accurate while the normal approximation works well only in the second situation.

8.1.4 Implication for the chi-squared test

The Normal Approximation to the Binomial is the key ingredient in the popular Chi-Squared test for goodness of fit. Because the normal approximation is not always valid, the chi-squared test comes with certain warnings recommending against its use in some exceptional cases (such as situations in which some of the cells have low counts). Use of the chi-squared test in such situations leads to paradoxical conclusions. This is very nicely illustrated in the following simple example (taken from Jaynes [1, Section 9.12]).

Example 8.1. *Suppose that a coin toss can give three different results: H (heads), T (tails) and edge (when the coin just stands on its edge). Suppose that a person A assigns probabilities* $0.499, 0.499, 0.002$ *to these three outcomes and another person B assigns probabilities* $1/3, 1/3, 1/3$ *to these outcomes. Suppose now that we perform an experiment with this coin by tossing it* $n = 29$ *times and this led to* 14 *heads,* 14 *tails and* 1 *edge.*

We are now interested in measuring the fit between each of the two models (of person A and B) and the observed data. If we use the chi-squared criterion:

$$\sum_{\text{all outcomes}} \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

for measuring goodness of fit, we would obtain

$$\chi^2_A = \frac{(14 - 29 \times 0.499)^2}{29 \times 0.499} + \frac{(14 - 29 \times 0.499)^2}{29 \times 0.499} + \frac{(1 - 29 \times 0.002)^2}{29 \times 0.002} = 15.33$$

as the goodness of fit for A and

$$\chi^2_B = \frac{(14 - 29/3)^2}{29/3} + \frac{(14 - 29/3)^2}{29/3} + \frac{(1 - 29/3)^2}{29/3} = 11.66$$

as the goodness of fit for B. This clearly runs against intuition as clearly A's model seems closer to the observed data compared to B. The reason for this strangeness is that the underlying normal approximation is not working.

The right approach is simply to calculate probabilities of the observed data for each of the two models. Specifically,

$$\mathbb{P}\{data \mid A\} = 0.499^{14} 0.499^{14} (0.002)^1 = 7 \times 10^{-12}$$

and

$$\mathbb{P}\{data \mid B\} = \left(\frac{1}{3}\right)^{14} \left(\frac{1}{3}\right)^{14} \left(\frac{1}{3}\right)^{1} = 1.46 \times 10^{-14}.$$

So A's model assigns much higher probability (about 483 times higher) to the observed data compared to B's model. This certainly matches our intuition. Note that if we don't know the specific order of the 29 outcomes, we can multiply the above probabilities by the multinomial coefficient

$$\frac{29!}{14!14!1!}$$

but this factor will not change anything as both the above probabilities will be multiplied by this same factor.

The moral of this example is to always calculate binomial/multinomial probabilities directly (or use the Entropy Approximation if an approximation is necessary). The normal approximation probabilities should be calculated only when one is sure that the normal approximation is accurate.

9 Lecture Nine

9.1 Normal Approximation for the Binomial: CLT

In the last class, we studied the normal approximation to the Binomial. We proved that

$$\mathbb{P}\{Bin(n,p) = k\} \approx \phi(k; np, np(1-p)) = \frac{1}{\sqrt{2\pi np(1-p)}} \exp\left(-\frac{(k-np)^2}{2np(1-p)}\right) \tag{35}$$

This approximation is not always good. It is only accurate when $f := k/n$ is closed to p. More precisely, we argued in the last class that the approximation is good provided

$$\frac{n|f-p|^3}{6\left(\min(f,p)\right)^2\left(\min(1-f,1-p)\right)^2} \tag{36}$$

is small. Note the presence of n in the numerator above. This means that, for the approximation to be accurate, f needs to be much closer to p when n is large.

The normal approximation of the binomial is usually stated in the form of the **De Moivre-Laplace Central Limit Theorem** as discussed below.

9.1.1 De Moivre-Laplace Central Limit Theorem

Fact 9.1. *Let $X \sim Bin(n,p)$ with fixed $0 < p < 1$. For every fixed pair of real numbers a and b with $a < b$, we have*

$$\lim_{n\to\infty} \mathbb{P}\left\{a \le \frac{X-np}{\sqrt{np(1-p)}} \le b\right\} = \int_a^b \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz. \tag{37}$$

Here is a sketch of the proof of (37). First write

$$\begin{aligned}\mathbb{P}\left\{a \le \frac{X-np}{\sqrt{np(1-p)}} \le b\right\} &= \mathbb{P}\left\{np + a\sqrt{np(1-p}} \le X \le np + b\sqrt{np(1-p}}\right\} \\ &= \sum_{k\in\left[np+a\sqrt{np(1-p)},np+b\sqrt{np(1-p)}\right]} \mathbb{P}\{X=k\}.\end{aligned}$$

The key now is to observe that, in the range $k \in \left[np + a\sqrt{np(1-p)}, np + b\sqrt{np(1-p)}\right]$, the normal approximation is accurate. To see this, first note that

$$f = \frac{k}{n} \in \left[p + a\sqrt{\frac{p(1-p)}{n}}, p + b\sqrt{\frac{p(1-p)}{n}}\right]$$

As a result, $|f-p|$ is at most of order $n^{-1/2}$ which implies that $n|f-p|^3$ will be of order $n^{-1/2}$ and thus small (the denominator in (36) will behave like a constant as p is fixed away from 0 and 1 and f is close to p). As a result, we can use the approximation (35) to write

$$\mathbb{P}\left\{a \le \frac{X-np}{\sqrt{np(1-p)}} \le b\right\} \approx \sum_{k\in\left[np+a\sqrt{np(1-p)},np+b\sqrt{np(1-p)}\right]} \phi\left(k; np, np(1-p)\right).$$

Using the formula

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

we get

$$\phi\left(k; np, np(1-p)\right) = \frac{1}{\sqrt{np(1-p)}}\phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$

so that

$$\mathbb{P}\left\{a \le \frac{X-np}{\sqrt{np(1-p)}} \le b\right\} \approx \sum_{k\in\left[np+a\sqrt{np(1-p)},np+b\sqrt{np(1-p)}\right]} \frac{1}{\sqrt{np(1-p)}}\phi\left(\frac{k-np}{\sqrt{np(1-p)}}\right)$$

Let us now write

$$z_k = \frac{k - np}{\sqrt{np(1-p)}} \qquad \text{so that } z_k - z_{k-1} = \frac{1}{\sqrt{np(1-p)}}.$$

Thus

$$\mathbb{P}\left\{a \leq \frac{X - np}{\sqrt{np(1-p)}} \leq b\right\} \approx \sum_{z_k \in [a,b]} (z_k - z_{k-1})\phi(z_k).$$

The left hand side above is a Riemann sum for $\int_a^b \phi(z)dz$ and it approaches that integral as $n \to \infty$ (note that $z_k - z_{k-1} \to 0$ as $n \to \infty$). This proves (37).

The statement (37) is also true if $a = -\infty$ and/or $b = +\infty$. This can be derived as a consequence of (37) for finite a and b. This argument is technical and omitted; if interested, see Corollary 3.2 of the book *Probability Theory* by Yakov G. Sinai.

9.2 The Exponential Distribution

The exponential distribution is given by 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. This can be justified by a discretization argument as follows.

Suppose we divide the time starting now into a large number of small intervals each of length δ. In each time interval, assume that there can be at most one phone call and that the probability of a phone call is a small real number p. Also assume independence of getting phone calls in distinct time intervals. In this setup, suppose X is the random variable denoting the time of the first phone call. For a positive real number x,

$$\mathbb{P}\{x \leq X < x + \delta\}$$

is the probability that there is no phone call in the first x/δ time intervals and that there is a phone call in the $(1 + (x/\delta))^{th}$ interval. Thus

$$\mathbb{P}\{x \leq X < x + \delta\} \approx (1-p)^{\frac{x}{\delta}} p.$$

The quantity p is quite small so we can use the approximation $1 - p \approx e^{-p}$. This gives

$$\mathbb{P}\{x \leq X < x + \delta\} \approx p \exp\left(-\frac{p}{\delta}x\right) = \delta \frac{p}{\delta} \exp\left(-\frac{p}{\delta}x\right).$$

Now the quantity $\frac{p}{\delta}$ is the average number of phone calls in unit time (note that there are $1/\delta$ intervals in unit time). This can therefore be termed as the "rate" of arrival of phone calls:

$$\lambda = \frac{p}{\delta}.$$

We can then write

$$\mathbb{P}\{x \leq X < x + \delta\} \approx \delta\lambda \exp(-\lambda x)$$

which is same as

$$f_X(x) = \lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{P}\{x \leq X < x + \delta\} = \lambda \exp(-\lambda x)$$

for $x > 0$.

Observe also that in the above setup, the distribution of the number of phone calls in any time interval of length T is $Bin(T/\delta, p)$ (as the number of small time intervals each of length δ in the original time interval of length T equals T/δ). The quantity

$$\frac{T}{\delta}p^2 = T\delta\left(\frac{p}{\delta}\right)^2 = T\delta\lambda^2$$

which is small if δ is small and T and λ are held fixed. The Poisson approximation holds and we can approximate the distribution of the number of phone calls in any time interval of length T as $Poi(T\frac{p}{\delta}) = Poi(\lambda T)$.

Also, by the assumption of independence, the number of phone calls in *disjoint* time intervals will be independent.

These two assumptions (Poisson distribution of arrivals in any time interval and independence of number of arrivals in disjoint time intervals) are characteristic of the Poisson process. We have therefore assumed that the phone call arrivals form a Poisson process. The waiting time for the first phone call then is Exponentially Distributed.

The exponential density has the memorylessness property (just like the Geometric distribution in the discrete case). To see this, first note that

$$\mathbb{P}\{X > x\} = \int_x^\infty \lambda e^{-\lambda u} du = e^{-\lambda x}.$$

which gives

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

The property

$$\mathbb{P}\{X > a + b \mid X > b\} = \mathbb{P}\{X > a\} \qquad \text{for all } a > 0, b > 0 \tag{38}$$

is called memorylessness.

The exponential density is the only density on $(0, \infty)$ that has the memorylessness property (proof left as exercise). In this sense, the Exponential distribution can be treated as the continuous analogue of the Geometric distribution. Note that a Geometric random variable would satisfy (38) when a, b are integers but not when a, b are arbitrary real numbers.

9.3 The Gamma Distribution

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

$$f(x) \propto x^{\alpha-1}e^{-\lambda x} I\{x > 0\}. \tag{39}$$

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 (39) 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 $X_1, \ldots, X_k$ 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 k, it can be shown that

$$\Gamma(k) = (k-1)! \qquad \text{for } k \geq 1. \tag{40}$$

The above inequality is a consequence of the property

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha) \qquad \text{for } \alpha > 0 \tag{41}$$

and the trivial fact that $\Gamma(1) = 1$. You can easily verify (41) 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 $\int e^{-x^2/2} dx = \sqrt{2\pi}$).

When $\alpha = k$ is a positive integer, the $\Gamma(k, \lambda)$ distribution arises as the distribution of the k^{th} arrival in a Poisson process of rate λ. To see, consider the same binomial setup (that we used for the modeling the arrivals of phone calls in the previous section). Then if X denotes the waiting time for the k^{th} phone call, then (for $x > 0$)

$$\mathbb{P}\{x \leq X < x + \delta\}$$

is the probability that there are exactly $k-1$ phone calls in the first x/δ small time intervals (of length δ) and an additional phone call in the $(x/\delta + 1)^{th}$ interval. Thus

$$\mathbb{P}\{x \leq X < x + \delta\} \approx \binom{x/\delta}{k-1} p^k \left(1-p\right)^{\frac{x}{\delta} - k + 1}.$$

If x and k are fixed, then x/δ is much larger than $k-1$ so we can approximate the right hand side above as (also as p is small)

$$\mathbb{P}\{x \leq X < x + \delta\} \approx \frac{(x/\delta)^{k-1}}{(k-1)!} p^k (1-p)^{x/\delta}.$$

Now using $1 - p \approx e^{-p}$ and writing $\lambda = p/\delta$,

$$\mathbb{P}\{x \leq X < x + \delta\} \approx \delta \frac{\lambda^k}{(k-1)!} x^{k-1} e^{-\lambda x}$$

which means that X has the $Gamma(k, \lambda)$ density.

9.4 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_X and $T(\cdot)$ is a smooth function, then it is fairly straightforward to write down the density of $Y = T(X)$ in terms of f_X. In the case when T is invertible and T^{-1} is differentiable, there is the following formula:

$$f_{T(X)}(y) = f_X(T^{-1}y) \left| \frac{dT^{-1}(y)}{dy} \right|.$$

We shall look at the ideas behind this formula (as well how to solve this problem when T is not invertible) in the next class. We will look at the following problem in the next class.

Example 9.2. *Suppose $X \sim U(-\pi/2, \pi/2)$. What is the density of $Y = \tan(X)$? We shall prove, in the next class, that Y has the **Cauchy** density:*

$$f_Y(y) = \frac{1}{\pi(1+y^2)} \qquad \text{for } -\infty < y < \infty.$$

10 Lecture Ten

10.1 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_X and $T(\cdot)$ is a smooth function, then it is fairly straightforward to write down the density of $Y = T(X)$ in terms of f_X. There are some general formulae for doing this but it is better to learn how to do it from first principles. The general idea will be clear from the following two examples.

Example 10.1. *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

$$\mathbb{P}\{y < Y < y + \delta\} \approx g(y)\delta \qquad \textit{when } \delta \textit{ is small.} \tag{42}$$

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 (42), 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.*

The answer derived in the above example is a special case of the following formula:

$$f_{T(X)}(y) = f_X(T^{-1}y) \left| \frac{dT^{-1}(y)}{dy} \right|. \tag{43}$$

which makes sense as long as T is invertible and T^{-1} is differentiable. If the function T is not invertible, then the formula above cannot be directly used but the method (based on first principles) used to derive the above formula is applicable in all cases. Here is an example illustrating this.

Example 10.2. *Suppose X has the standard normal density. What is the density of $Y = X^2$?*

The function $T(x) = x^2$ is not invertible so the formula (43) cannot be used directly. Instead we argue from first principles as follows. Let $y > 0$. The density of Y at y is given by

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

For small $\delta > 0$, we can write

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

Thus

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

This is the density of the chi-squared random variable with 1 degree of freedom. This is also the Gamma random variable with shape parameter $\alpha = 0.5$ and rate parameter $\lambda = 0.5$.

Note that we can also try to calculate the density of Y at 0 by the above method:

$$\mathbb{P}\{0 < Y < \delta\} = \mathbb{P}\{-\sqrt{\delta} < X < \sqrt{\delta}\} \approx 2\phi(0)\sqrt{\delta}$$

so that

$$f_Y(0) = \lim_{\delta \downarrow 0} \frac{1}{\delta} \mathbb{P}\{0 < Y < \delta\} = 2\phi(0) \lim_{\delta \downarrow 0} \frac{\sqrt{\delta}}{\delta} = 2\phi(0) \lim_{\delta \downarrow 0} \delta^{-1/2} = \infty.$$

This ∞ does not affect any calculation of probabilities $\mathbb{P}\{Y \in A\}$ as these densities are calculated by the integral $\int_A f_Y(y)dy$ and the value of f_Y at the one point 0 does not affect the value of this integral.

10.2 The Cumulative Distribution Function and the Quantile Transform

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

If the random variable X has a density f_X, then its cdf is given by

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

and, in this case, it is generally true that $F'(x) = f_X(x)$.

The *inverse* of the cdf is used to define quantiles. Given a random variable X and a number $u \in (0, 1)$, the u-quantile of the distribution of X is given by a real number $q_X(u)$ satisfying

$$\mathbb{P}\{X \leq q_X(u)\} = u \tag{44}$$

provided such a number $q_X(u)$ exists uniquely. If F_X is the cdf of X, the equation (44) simply becomes

$$F_X(q_X(u)) = u$$

so we can write

$$q_X(u) = F_X^{-1}(u).$$

Here are some simple examples.

Example 10.3 (Uniform)**.** *Suppose X has the uniform distribution on $(0, 1)$. Then $F_X(x) = x$ for $x \in (0, 1)$ and thus $F_X^{-1}(u)$ exists uniquely for every $u \in (0, 1)$ and equals u. We thus have $q_X(u) = u$ for every $u \in (0, 1)$.*

Example 10.4 (Normal)**.** *Suppose X has the standard normal distribution. Then $F_X(x) = \Phi(x)$ where*

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt = \int_{-\infty}^{x} (2\pi)^{-1/2} \exp\left(\frac{-t^2}{2}\right) dt.$$

There is no closed form expression for Φ but its values can be obtained in R (for example) using the function pnorm. Φ is a strictly increasing function from $(-\infty, \infty)$ to $(0,1)$ so its inverse exists uniquely and we thus have

$$q_X(u) = \Phi^{-1}(u) \qquad \text{for every } u \in (0,1).$$

There is no closed form expression for $q_X = \Phi^{-1}$ but its values can be obtained from R by the function qnorm.

Example 10.5 (Cauchy)**.** *Suppose X has the standard Cauchy density:*

$$f_X(x) := \frac{1}{\pi} \frac{1}{1+x^2} \qquad \text{for } -\infty < x < \infty.$$

Its cdf is given by

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\pi} \frac{dt}{1+t^2} = \frac{1}{\pi} \arctan(t) \Big|_{-\infty}^{x} = \frac{1}{\pi} \arctan(x) + \frac{1}{2}.$$

It is easy to see that this is a strictly increasing function from (∞, ∞) to $(0,1)$ and its inverse is given by

$$F_X^{-1}(u) = \tan\left(\pi\left(u - 0.5\right)\right).$$

Thus the quantile function for the Cauchy distribution is given by

$$q_X(u) = \tan\left(\pi\left(u - 0.5\right)\right) \qquad \text{for every } u \in (0,1).$$

How to define the u-quantile when there is no solution or multiple solutions to the equation $F_X(q) = u$? No solutions for $F_X(q) = u$ can happen for discrete distributions (for example, for $X \sim Ber(0.5)$ and $u = 0.25$, there is no q satisfying $\mathbb{P}\{X \le q\} = u$). Multiple solutions can also happen. For example, if X is uniformly distributed on the set $[0,1] \cup [2,3]$ and $u = 0.5$, then every $q \in [1,2]$ satisfies $F_X(q) = 0.5$. In such cases, it is customary to define the u-quantile via

$$q_X(u) := \inf\{x \in \mathbb{R} : F_X(x) \ge u\}. \tag{45}$$

This can be seen as a generalization of $F_X^{-1}(u)$. Indeed, if there is a unique q such that $F_X(q) = u$, it is easy to see then that $q_X(u) = q$.

The function $q_X : (0,1) \to (-\infty, \infty)$ defined by (45) is called the quantile function or the quantile transform of the random variable X. It can be checked that the definition (45) ensures that

$$\mathbb{P}\{X \le q_X(u)\} \ge u \quad \text{and} \quad \mathbb{P}\{X < q_X(u)\} \le u. \tag{46}$$

Example 10.6 (Bernoulli)**.** *Suppose $X \sim Ber(p)$ i.e., $\mathbb{P}\{X = 0\} = 1-p$ and $\mathbb{P}\{X = 1\} = p$. When then is $q_X(u)$ for $u \in (0,1)$? It can be checked that*

$$q_X(u) = \begin{cases} 0 & \text{for } 0 < u \le 1-p \\ 1 & \text{for } 1-p < u < 1 \end{cases}$$

The quantile transform is important for the following reason.

Proposition 10.7. *The following two statements are true.*

1. *Suppose U is a random variable distributed according to the uniform distribution on $(0,1)$. Then $q_X(U)$ has the same distribution as X. In other words, the function q_X transforms the uniform distribution to the distribution of X.*

2. *Suppose X is a random variable with a **continuous** cdf F_X. Then $F_X(X)$ has the uniform distribution on $(0,1)$. In other words, the function F_X transforms the distribution of X into the $Unif(0,1)$ distribution (provided the distribution of X is continuous).*

It should be stressed that the first conclusion of the above Proposition holds for every X (discrete or continuous) while the second conclusion is only true if F_X is continuous. To see why the second conclusion is false when F_X is not continuous, suppose that $X \sim Ber(p)$ so that X takes only the two values 0 and 1. Then $F_X(X)$ also takes only two values: $F_X(0) = 1-p$ and $F_X(1) = 1$; thus $F_X(X)$ cannot have the uniform distribution on $(0,1)$.

Example 10.8 (Cauchy)**.** *We have seen in Example 10.5 that for a standard Cauchy random variable, $q_X(u) = \tan(\pi(u-0.5))$. Proposition 10.7 then gives that if $U \sim Unif(0,1)$, then*

$$\tan(\pi(U-0.5)) \sim Cauchy.$$

Note that $\pi(U-0.5) \sim Unif(-\pi/2, \pi/2)$. Thus the tan function applied to a uniformly distributed random variable on $(-\pi/2, \pi/2)$ results in a random variable having the Cauchy distribution (as we have seen in Example 10.1).

Example 10.9 (Bernoulli)**.** *The quantile function for $Ber(p)$ was calculated in* (10.6) *as*

$$q(u) = I\{1-p < u < 1\}.$$

As a result, the first conclusion of Proposition 10.7 states that, for $U \sim Unif(0,1)$,

$$q(U) = I\{1-p < U < 1\} \sim Ber(p)$$

as can be checked directly. Note that this function q is not the only function with the property that $q(U) \sim Ber(p)$. For example, the function $\tilde{q}(u) = I\{0 < u < p\}$ also satisfies $\tilde{q}(U) \sim Ber(p)$.

11 Lecture Eleven

11.1 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.

A real-valued function of two-variables $f(\cdot,\cdot)$ is called a joint density if

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

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

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

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

Here are two simple examples of joint densities.

Example 11.1. *Consider the function*

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

First note that this is a valid joint density as this function is always nonnegative and

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{1}{2\pi}\exp\left(-\frac{x^2+y^2}{2}\right)dxdy \\ &= \left[\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx\right]\left[\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right)dy\right] = 1\end{aligned}$$

Suppose a pair of random variables (X, Y) have this as their joint density. Then probabilities involving X, Y are calculated by integrating the density $f(x, y)$ in the appropriate region. For example,

$$\begin{aligned}\mathbb{P}\{-1 \le X+Y \le 2\} &= \int\int I\{(x,y) : -1 \le x+y \le 2\} f(x,y)dxdy \\ &= \frac{1}{2\pi}\int\int I\{(x,y) : -1 \le x+y \le 2\} \exp\left(-\frac{1}{2}(x^2+y^2)\right)dxdy\end{aligned}$$

The set $\{(x, y) : -1 \le x + y \le 2\}$ represents the region betwen the two lines $x + y = 2$ and $x + y = -1$.

Example 11.2. *Consider the function*

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

This function takes the value 1 on the set $\{(x, y) : 0 \le x \le 1, 0 \le y \le 1\}$ and can also be written succinctly as

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

This is clearly a density function as this is nonnegative and integrates to one (the area of the unit square $\{(x, y) : 0 \le x \le 1, 0 \le y \le 1\}$ equals 1). Suppose the random variables X and Y have this joint density f, then

$$\begin{aligned}\mathbb{P}\{(X,Y) \in B\} &= \int\int I\{(x,y) \in B\} f(x,y)dxdy \\ &= \int\int I\{(x,y) \in B\} I\{0 \le x, y \le 1\}dxdy \\ &= \text{area of } B \cap \{(x,y) : 0 \le x, y \le 1\}\end{aligned}$$

For example,

$$\mathbb{P}\{X^2+Y^2 \le 1\} = \text{area of } \left(\{(x,y) : 0 \le x, y \le 1 \text{ and } x^2+y^2 \le 1\}\right) = \frac{\pi}{4}.$$

In order to calculate joint densities, the following formula is very useful. If Δ is a *small* region in $\mathbb{R}^2$ around a point (x, y), we have

$$\mathbb{P}\{(X,Y) \in \Delta\} \approx (\text{area of } \Delta)\, f_{X,Y}(x,y). \tag{47}$$

More formally,

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

where the limit is taken as Δ shrinks to (x,y). Here are two special cases of this formula:

1. By taking Δ to be the rectangle $\{(a,b) : x \le a \le x+\delta, y \le b \le y+\epsilon\}$ for small δ and ϵ, we get
$$f_{X,Y}(x,y) = \lim_{\delta \downarrow 0, \epsilon \downarrow 0} \frac{\mathbb{P}\{x \le X \le x+\delta, y \le Y \le y+\epsilon\}}{\delta\epsilon}. \tag{48}$$

2. By taking Δ to be the circle centered at (x,y) of radius r, we get
$$f_{X,Y}(x,y) = \lim_{r \downarrow 0} \frac{\mathbb{P}\{(X-x)^2 + (Y-y)^2 \le r^2\}}{\pi r^2}.$$

The usefulness of the formula (47) is illustrated in the following two examples.

Example 11.3. *Suppose (X,Y) have the joint density:*

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

Define two new random variables R and Θ as follows: $R := \sqrt{X^2+Y^2}$ and Θ is the angle made by the vector (X,Y) with the positive X-axis (measured from the positive X-axis in the counterclockwise direction) (note that Θ takes values between 0 and 2π). What is the joint density of (R,Θ)?

Let us calculate the joint density $f_{R,\Theta}(r,\theta)$ of (R,Θ) at a fixed point (r,θ). Let us assume that $r > 0$ and $0 < \theta < 2\pi$. One way of calculating this is via (48):

$$f_{R,\Theta}(r,\theta) = \lim_{\delta,\epsilon \downarrow 0} \frac{\mathbb{P}\{r < R < r+\delta, \theta < \Theta < \theta+\epsilon\}}{\delta\epsilon} \tag{49}$$

We can calculate $\mathbb{P}\{r < R < r+\delta, \theta < \Theta < \theta+\epsilon\}$ in the following way:

$$\mathbb{P}\{r < R < r+\delta, \theta < \Theta < \theta+\epsilon\} = \mathbb{P}\{(X,Y) \in S\}$$

where S is the set of all points (x,y) such that $r < \sqrt{x^2+y^2} < r+\delta$ and the angle made by (x,y) with the positive x-axis lies between θ and $\theta+\epsilon$. As can be seen from Figure 2, when δ, ϵ are small, the set S is a small region around the point $(r\cos\theta, r\sin\theta)$. Moreover, its area is approximately equal to $r\epsilon\delta$. We thus get

$$\begin{aligned}\mathbb{P}\{(X,Y) \in S\} &\approx f_{X,Y}(r\cos\theta, r\sin\theta) \times \textit{area of } S \\ &= \frac{1}{2\pi} \exp\left(-\frac{(r\cos\theta)^2 + (r\sin\theta)^2}{2}\right) r\delta\epsilon = r\exp\left(-\frac{r^2}{2}\right)\frac{1}{2\pi}\delta\epsilon.\end{aligned}$$

Combining the above with (58), we deduce that

$$f_{R,\Theta}(r,\theta) = r\exp\left(-\frac{r^2}{2}\right)\frac{1}{2\pi} \text{ whenever } r > 0, 0 < \theta < 2\pi$$

Example 11.4. *Suppose X, Y have joint density $f_{X,Y}$. What is the joint density of U and V where $U = X$ and $V = X+Y$?*

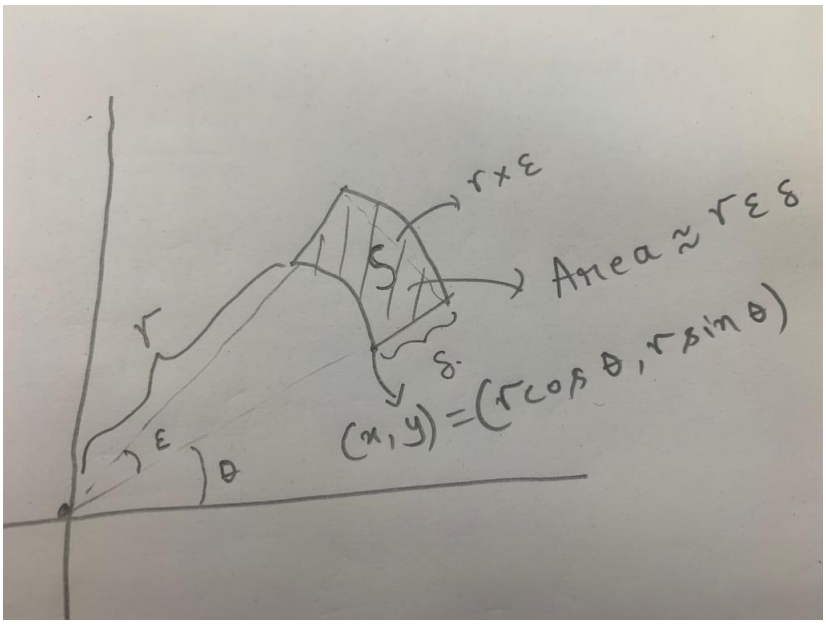

Figure 2: The set S

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*

$$\mathbb{P}\{u \leq U \leq u + \delta, v \leq V \leq v + \epsilon\} \approx \delta\epsilon f_{U,V}(u, v). \tag{50}$$

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*

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

where $S(R)$ *is the image of the rectangle* R *under the mapping* S*. How does* $S(R)$ *look like? It 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*

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

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 (50)*, 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)*.*

The logic behind the above two examples can be extended to obtain formulae for the joint density of an arbitrary transformation of a pair of random variables with known joint density. We shall first consider linear transformations (as in Example 11.4) and, in the next class, consider nonlinear transformations.

11.2 Joint Densities under General Linear Invertible transformations

Let us first recall some basic properties of linear transformations.

11.2.1 Linear Transformations

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

$$L(x, y) := M \begin{pmatrix} x \\ y \end{pmatrix} + c \tag{51}$$

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 (51) is invertible if and only if the matrix M is invertible. We shall only be dealing with invertible linear transformations. The following are two standard properties of linear transformations that you need to familiar with.

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.

11.2.2 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 previous example, we had $T(x, y) = (x, x + y)$ 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

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

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

$$f_{U,V}(u,v) = f_{X,Y}(S(u,v))\left|\det M_S\right|. \tag{52}$$

Remember again that M_S is the 2×2 matrix corresponding to the linear transformation S.

In the next class, we shall see extensions of the formula (54) for nonlinear transformations.

12 Lecture Twelve

12.1 Last Class: Joint Density

In the last lecture, we started discussing joint densities. Any function $f(x,y)$ of two real variables x and y is a joint density provided it is nonnegative and integrates (over both x and y) to 1. If two random variables X and Y have joint density $f_{X,Y}$, then

$$\mathbb{P}\left\{(X,Y) \in B\right\} = \int\int I\{(x,y) \in B\} f_{X,Y}(x,y)dxdy$$

for every set $B \subseteq \mathbb{R}^2$. One further has

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

which implies that

$$\mathbb{P}\{(X,Y) \in \Delta\} \approx f_{X,Y}(x,y) \times \text{area of } \Delta \tag{53}$$

provided Δ is a small region around the point (x,y). The precise shape of the small region Δ is immaterial for (53).

12.2 Marginal Densities corresponding to a Joint Density

Suppose X and Y have joint density $f_{X,Y}$. Then probabilities involving only the random variable X can be calculated as:

$$\begin{aligned}
\mathbb{P}\{X \in A\} &= \mathbb{P}\{X \in A, -\infty < Y < \infty\} \\
&= \int\int f_{x,Y}(x,y) I\{x \in A, -\infty < y < \infty\} dxdy \\
&= \int I\{x \in A\} \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right) dx \\
&= \int_A \left(\int_{-\infty}^{\infty} f_{X,Y}(x,y)dy\right) dx.
\end{aligned}$$

Comparing this with the formula for the density $f_X(x)$ of a single random variable X:

$$\mathbb{P}\{X \in A\} = \int_A f_X(x)dx,$$

we immediately deduce that $f_X(x)$ can be written in terms of $f_{X,Y}(x,y)$ as:

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

Analogous, the density $f_Y(y)$ of Y is given by:

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

In words, the density of a single random variable can be derived by integrating the joint density (of this random variable and another random variable) with respect to the other variable.

When discussing a joint density $f_{X,Y}$, individual densities f_X of X and f_Y of Y are referred to as *marginal* densities.

12.3 Independence in terms of Joint Densities

Independence of two random variables X and Y can be characterized in terms of their joint density $f_{X,Y}$ using any of the following statements. The following statements are all equivalent to each other:

1. The random variables X and Y are independent.
2. The joint density $f_{X,Y}(x,y)$ factorizes into the product of a function depending on x alone and a function depending on y alone.
3. $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ for all x, y.

Example 12.1. *The joint density*

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

factorizes into the product of a function depending on x alone and a function depending on y alone because

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

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 $[0,1]$.

Example 12.2. *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?)

12.4 How linear transformations change joint densities

In the last lecture, we looked at the following fact. 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. Then the joint density of U and V is given by

$$f_{U,V}(u, v) = f_{X,Y}(S(u, v)) \left|\det M_S\right|. \tag{54}$$

where M_S is the 2×2 matrix corresponding to the linear transformation S.

As an example suppose $U = X$ and $V = X + Y$ so that $T(x, y) = (x, x + y)$ and $S(u, v) = (u, v - u)$. The matrix corresponding to S is

$$M_S = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}.$$

The determinant of M_S is clearly 1. The formula (54) then gives

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

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

12.5 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

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

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 in the last class) 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

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

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

$$f_{U,V}(u,v) = f_{X,Y}(S(u,v)) \left|\det J_S(u,v)\right|. \tag{55}$$

The following is an example of this formula (we derived the result in this example using first principles in the last class)

Example 12.3. *Suppose X and Y are two random variables having joint density $f_{X,Y}$. Define two new random variables R and Θ in the following way. $R := \sqrt{X^2 + Y^2}$ and Θ is the angle made by the vector (X, Y) with the positive X-axis in the counterclockwise direction. 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 (R, Θ) at (r, θ) is zero unless $r > 0$ and $0 < \theta < 2\pi$. The formula (55) then gives

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(r\cos\theta, r\sin\theta) \left|\det \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix}\right| = f_{X,Y}(r\cos\theta, r\sin\theta) \times (r)$$

We have thus derived the formula:

$$f_{R,\Theta}(r,\theta) = r f_{X,Y}(r\cos\theta, r\sin\theta) \qquad \text{for every } r > 0 \text{ and } 0 < \theta < 2\pi. \tag{56}$$

We can also write this formula as:

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{x^2+y^2}} f_{R,\Theta}(\sqrt{x^2+y^2}, \theta(x,y)) \qquad \text{for every } -\infty < x, y < \infty. \tag{57}$$

where $\sqrt{x^2 + y^2}$ represents r and $\theta(x, y)$ is the made by the vector (x, y) with the positive X-axis in the counterclockwise direction.

The formulae (56) and (57) have an important connection to the Herschel-Maxwell derivation of the normal distribution which we discuss next.

12.6 The Herschel-Maxwell Derivation of the Normal Distribution

In the context of Example 12.3, the astronomer John Herschel derived the normal distribution in the following way (this derivation was extended to the three dimensional case by the physicist James Clerk Maxwell). See Jaynes [1, Section 7.2] for more details. Their result is the following.

Fact 12.4. *Suppose X and Y are two random variables. Suppose that R and Θ are defined as in Example 12.3. Assume the following three conditions:*

1. *X and Y are independent and identically distributed*
2. *R and Θ are independent*
3. *Θ is uniformly distributed on $(0, 2\pi)$.*

Then X and Y have the normal distribution $N(0, \sigma^2)$ with mean zero and some variance σ^2.

Before proving (12.4), let us note that if X and Y are independently distributed as $N(0, \sigma^2)$, then all the conditions above are true. To see this, observe first that, by independence,

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

The formula (56) then gives

$$f_{R,\Theta}(r, \theta) = r f_{X,Y}(r\cos\theta, r\sin\theta) = r\frac{1}{2\pi\sigma^2} \exp\left(-\frac{(r\cos\theta)^2 + (r\sin\theta)^2}{2\sigma^2}\right) = r\frac{1}{2\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right).$$

The above is the joint density of R and Θ at (r, θ) provided $r > 0$ and $0 < \theta < 2\pi$. To make the ranges of the variables r and θ clear, we write

$$f_{R,\Theta}(r, \theta) = r\frac{1}{2\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) I\{r > 0\} I\{0 < \theta < 2\pi\}.$$

The right hand side above clearly factorizes into the product of a function depending on r alone and a function depending on θ alone. This implies that R and Θ are independent. The marginal distribution of Θ is given by integrating over r:

$$f_\Theta(\theta) = I\{0 < \theta < 2\pi\} \int_0^\infty r\frac{1}{2\pi\sigma^2} \exp\left(-\frac{r^2}{2\sigma^2}\right) dr$$

The substitution $s = \frac{r^2}{2\sigma^2}$ (so $ds = rdr/\sigma^2$) leadds to

$$f_\Theta(\theta) = I\{0 < \theta < 2\pi\} \int_0^\infty \frac{e^{-s}}{2\pi} ds = \frac{I\{0 < \theta < 2\pi\}}{2\pi}.$$

This means that Θ is uniformly distributed over $(0, 2\pi)$. We have thus proved that all the three conditions of Fact 12.4 are satisfied when X, Y are independently distributed as $N(0, \sigma^2)$.

We shall now prove Fact 12.4 by showing that X, Y being $N(0, \sigma^2)$ is the only way all the three conditions are satisfied.

Proof of Fact 12.4. We shall work with (57) which connects the joint density of (X, Y) to the joint density of (R, Θ). Because X and Y are assumed to be independent and identically distributed, the left hand side of (57) becomes

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

where f is the common density of X and Y. On the other hand, because R and Θ are independent and $\Theta \sim \text{Unif}(0, 2\pi)$, the right hand side of (57) becomes

$$\begin{aligned}\frac{1}{\sqrt{x^2+y^2}} f_{R,\Theta}(\sqrt{x^2+y^2}, \theta(x,y)) &= \frac{1}{\sqrt{x^2+y^2}} f_R(\sqrt{x^2+y^2}) f_\Theta(\theta(x,y)) \\ &= \frac{1}{\sqrt{x^2+y^2}} f_R(\sqrt{x^2+y^2}) \frac{1}{2\pi}.\end{aligned}$$

We thus obtain

$$f(x)f(y) = \frac{1}{\sqrt{x^2+y^2}} f_R(\sqrt{x^2+y^2}) \frac{1}{2\pi} \tag{58}$$

for every $-\infty < x, y < \infty$. Plugging in $y = 0$ above, we obtain

$$\frac{1}{|x|} f_R(|x|) \frac{1}{2\pi} = f(x)f(0) \qquad \text{for every } -\infty < x < \infty.$$

Plugging in $\sqrt{x^2+y^2}$ in place of x above, we obtain

$$\frac{1}{\sqrt{x^2+y^2}} f_R(\sqrt{x^2+y^2}) \frac{1}{2\pi} = f(\sqrt{x^2+y^2}) f(0).$$

Combining the above identity with (58), we deduce

$$f(x)f(y) = f(\sqrt{x^2+y^2})f(0) \qquad \text{for every } -\infty < x, y < \infty. \tag{59}$$

This identity implies that f is a symmetric function (i.e., $f(x) = f(-x) = f(|x|)$) because if we take $y = 0$, we get $f(x)f(0) = f(|x|)f(0)$ or $f(x) = f(|x|)$.

Let $h : [0, \infty) \to [0, \infty)$ be defined by

$$h(u) = f(\sqrt{u}) \qquad \text{for every } u \geq 0.$$

Then (59) implies

$$h(u)h(v) = h(u+v)h(0) \qquad \text{for every } u, v \geq 0,$$

or equivalently

$$\frac{h(u+v)}{h(0)} = \frac{h(u)}{h(0)} \frac{h(v)}{h(0)} \qquad \text{for every } u, v \geq 0.$$

This implies that for every nonnegative integer m and $u \geq 0$,

$$\frac{h(mu)}{h(0)} = \frac{h(u + \cdots + u)}{h(0)} = \frac{h(u)}{h(0)} \cdots \frac{h(u)}{h(0)} = \left(\frac{h(u)}{h(0)}\right)^m \tag{60}$$

Two consequences of the above are:

$$\frac{h(mu/n)}{h(0)} = \left(\frac{h(u/n)}{h(0)}\right)^m$$

which is obtained by replacing u by u/n in (60), and

$$\frac{h(u/n)}{h(0)} = \left(\frac{h(u)}{h(0)}\right)^{1/n}$$

which is obtained by replacing u by u/n in (60) and taking $m = n$. Combining the above two equations, we obtain

$$\frac{h(mu/n)}{h(0)} = \left(\frac{h(u/n)}{h(0)}\right)^m = \left(\frac{h(u)}{h(0)}\right)^{m/n}.$$

As m and n are nonnegative integers, we have deduced (take $u = 1$)

$$\frac{h(x)}{h(0)} = \left(\frac{h(1)}{h(0)}\right)^x$$

whenever $x \geq 0$ is a rational number (i.e., of the form m/n for some integers m and n). If we now assume that h is continuous, we can deduce the above for every $x \geq 0$. We have thus proved that

$$h(x) = h(0) \exp\left(x \log \frac{h(1)}{h(0)}\right) = c \exp(xb)$$

for some constants c ($c = h(0)$) and b ($b = \log \frac{h(1)}{h(0)}$). As $f(\sqrt{u}) = h(u)$ (and f is symmetric), we get

$$f(\pm\sqrt{u}) = c \exp(ub).$$

We have thus proved that

$$f(x) = c \exp\left(x^2 b\right) \qquad \text{for every } -\infty < x < \infty.$$

As f needs to be a valid density, we must have $b < 0$ so we can write $b = -\frac{1}{2\sigma^2}$ for some $\sigma > 0$. This will necessarily imply that $c = \frac{1}{\sqrt{2\pi}\sigma}$ leading to f being the $N(0, \sigma^2)$ density. This completes the proof of Fact 12.4. □

13 Lecture Thirteen

13.1 Joint Density under Transformations

Let (X, Y) have joint density $f_{X,Y}$. We transform (X, Y) to two new random variables (U, V) via $(U, V) = T(X, Y)$. Suppose that T is invertible (having an inverse $S = T^{-1}$) and differentiable. In the last class, we saw the following formula relating the joint density of (U, V) to $f_{X,Y}$:

$$f_{U,V}(u, v) = f_{X,Y}(S(u, v)) \left|\det J_S(u, v)\right|. \tag{61}$$

Let us start today by working out the following simple application of the formula (61).

Example 13.1. *Suppose X and Y are independent random variables with*

$$X \sim \mathit{Gamma}(\alpha_1, \lambda) \quad \text{and} \quad Y \sim \mathit{Gamma}(\alpha_2, \lambda).$$

Note the the rate parameter is the same in both the Gamma distributions. Now define

$$U := X + Y \quad \text{and} \quad V := \frac{X}{X + Y}.$$

What is the joint density of U and V? Here the transformation T is given by $T(x, y) = (x + y, x/(x + y))$ and its inverse transformation can be checked to be $S(u, v) = (uv, u(1-v))$. The formula (61) then gives that for every $u > 0$ and $0 < v < 1$ (we are taking $u > 0$ because the random variable U is always positive and V is between 0 and 1):

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

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

$$f_{U,V}(u, v) = \frac{\lambda^{\alpha_1+\alpha_2}}{\Gamma(\alpha_1 + \alpha_2)} u^{\alpha_1+\alpha_2-1} e^{-\lambda u} I\{u > 0\} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} v^{\alpha_1-1}(1 - v)^{\alpha_2-1} I\{0 < v < 1\}. \tag{62}$$

This implies that $U \sim Gamma(\alpha_1 + \alpha_2, \lambda)$. (62) *also implies that the density of* V *is*

$$f_V(v) = \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} v^{\alpha_1 - 1}(1-v)^{\alpha_2 - 1} I\{0 < v < 1\}.$$

The above density is known as the Beta density with parameters α_1 *and* α_2*:* $Beta(\alpha_1, \alpha_2)$. *Using the notation*

$$B(\alpha_1, \alpha_2) = \frac{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\Gamma(\alpha_1 + \alpha_2)} = \int_0^1 v^{\alpha_1 - 1}(1-v)^{\alpha_2 - 1} dv, \tag{63}$$

the Beta density can also be written as

$$f_V(v) = \frac{1}{B(\alpha_1, \alpha_2)} v^{\alpha_1 - 1}(1-v)^{\alpha_2 - 1} I\{0 < v < 1\}.$$

The name "Beta density" is derived from the Beta function which is the name given to the function (63) *in mathematics.*

One of the conclusions of the above example is that if $X_1 \sim Gamma(\alpha_1, \lambda)$ and $X_2 \sim Gamma(\alpha_2, \lambda)$ are independent, then

$$X_1 + X_2 \sim Gamma(\alpha_1 + \alpha_2, \lambda). \tag{64}$$

More generally, if $X_1, \ldots, X_n$ are independent random variables with $X_i \sim Gamma(\alpha_i, \lambda)$, then

$$X_1 + \cdots + X_n \sim Gamma(\alpha_1 + \cdots + \alpha_n, \lambda). \tag{65}$$

(65) can be proved from (64) by, for example, mathematical induction over n. One consequence of (65) is that if $Z_1, \ldots, Z_n$ are independent random variables having the standard normal distribution, then

$$Z_1^2 + \cdots + Z_n^2 \sim Gamma\left(\frac{n}{2}, \frac{1}{2}\right).$$

This is because, as we saw earlier in Lecture 10 (see Example 1.2 of Lecture 10), the square of a normal random variable has the $Gamma(1/2, 1/2)$ distribution. The distribution of the sum of squares of n independent normal random variables is also known as the chi-squared distribution with n degrees of freedom: χ_n^2. We thus have

$$\chi_n^2 = Gamma\left(\frac{n}{2}, \frac{1}{2}\right).$$

13.2 Conditional Densities for Continuous Random Variables

Consider two random variables X and Y with joint density $f_{X,Y}$. How do we calculate the conditional probability:

$$\mathbb{P}\{X \in A \mid Y = y_0\} \tag{66}$$

for some subset $A \subseteq \mathbb{R}$ and $y_0 \in \mathbb{R}$. The naive way to calculate the above probability is to write it as

$$\mathbb{P}\{X \in A \mid Y = y_0\} = \frac{\mathbb{P}\{X \in A \mid Y = y_0\}}{\mathbb{P}\{Y = y_0\}}.$$

The denominator on the right hand side above equals 0 because Y is a continuous random variable. As a result, the numerator is also equal zero. Thus the right hand side equals $\frac{0}{0}$ and hence undefined.

The proper way to define (66) is to think of the conditioning event $Y = y_0$ as $y_0 - \epsilon/2 \leq Y \leq y_0 + \epsilon/2$ for some small ϵ. We then have

$$\begin{aligned}
\mathbb{P}\{X \in A \mid Y = y_0\} &\approx \mathbb{P}\{X \in A \mid y_0 - \epsilon/2 \leq Y \leq y_0 + \epsilon/2\} \\
&= \frac{\mathbb{P}\{X \in A, y_0 - \epsilon/2 \leq Y \leq y_0 + \epsilon/2\}}{\mathbb{P}\{y_0 - \epsilon/2 \leq Y \leq y_0 + \epsilon/2\}} \\
&= \frac{\int_A \int_{y_0-\epsilon/2}^{y_0+\epsilon/2} f_{X,Y}(x,y)dxdy}{\int_{y_0-\epsilon/2}^{y_0+\epsilon/2} f_Y(y)dy} \\
&\approx \frac{\int_A (f_{X,Y}(x,y_0) \times \epsilon)\, dx}{f_Y(y_0) \times \epsilon} = \int_A \left(\frac{f_{X,Y}(x,y_0)}{f_Y(y_0)}\right) dx.
\end{aligned}$$

Motivated by the above calculation, we define the conditional density of X given $Y = y$ as

$$f_{X|Y=y}(x) := \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{67}$$

This is well-defined as long as $f_Y(y) > 0$. The result we just derived can also be written as

$$\mathbb{P}\{X \in A \mid Y = y_0\} = \int_A f_{X|Y=y}(x)dx.$$

Here are important facts about conditional densities.

13.2.1 Conditional Density is Proportional to Joint Density

As a function of x (and keeping y fixed), $f_{X|Y=y}(x)$ is a valid density i.e.,

$$f_{X|Y=y}(x) \geq 0 \text{ for every } x \quad \text{and} \quad \int_{-\infty}^{\infty} f_{X|Y=y}(x)dx = 1.$$

The integral above equals one because

$$\int_{-\infty}^{\infty} f_{X|Y=y}(x)dx = \int_{-\infty}^{\infty} \frac{f_{X,Y}(x,y)}{f_Y(y)}dx = \frac{\int_{-\infty}^{\infty} f_{X,Y}(x,y)dx}{f_Y(y)} = \frac{f_Y(y)}{f_Y(y)} = 1.$$

Because $f_{X|Y=y}(x)$ integrates to one as a function of x and because the denominator $f_Y(y)$ in the definition (67) does not depend on x, it is common to write

$$f_{X|Y=y}(x) \propto f_{X,Y}(x,y). \tag{68}$$

The symbol $\propto$ here stands for "proportional to" and the above statement means that $f_{X|Y=y}(x)$, as a function of x, is proportional to $f_{X,Y}(x,y)$. The proportionality constant then has to be $f_Y(y)$ because that is equal to the value of the integral of $f_{X,Y}(x,y)$ as x ranges over $(-\infty, \infty)$.

The proportionality statement (68) often makes calculations involving conditional densities much simpler.

13.2.2 Conditional Densities and Independence

X and Y are independent if and only if $f_{X|Y=y} = f_X$ for every value of y such that $f_Y(y) > 0$. This latter statement is precisely equivalent to $f_{X,Y}(x,y) = f_X(x)f_Y(y)$. By switching roles of X and Y, it also follows that X and Y are independent if and only if $f_{Y|X=x} = f_Y$ for every x such that $f_X(x) > 0$.

13.2.3 Law of Total Probability for Continuous Random Variables

Note first that from the definition of $f_{X|Y=y}(x)$, it directly follows that

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

This tells us how to compute the joint density of X and Y using knowledge of the marginal of Y and the conditional density of X given Y.

From here (and the fact that integrating the joint density with respect to one of the variables gives the marginal density of the other random variable), it is easy to derive the formula

$$f_X(x) = \int f_{X|Y=y}(x) f_Y(y) dy. \tag{69}$$

This formula, known as the **Law of Total Probability**, allows us to deduce the marginal density of X using knowledge of the conditional density of X given Y and the marginal density of Y.

Here are two applications of the Law of Total Probability.

Example 13.2. *Suppose X and Y are independent standard normal random variables. What is the density of $U = X/Y$?*

Using the Law of Total Probability, we get

$$f_U(u) = \int_{-\infty}^{\infty} f_{U|Y=v}(u) f_Y(v) dv = \int_{-\infty}^{\infty} f_{\frac{X}{Y}|Y=v}(u) f_Y(v) dv.$$

Now (below $\stackrel{d}{=}$ stands for equality in distribution: $A \stackrel{d}{=} B$ means that the random variables A and B have the same distribution)

$$\frac{X}{Y} \mid Y = v \stackrel{d}{=} \frac{X}{v} \mid Y = v \stackrel{d}{=} \frac{X}{v}$$

where the last equality follows because X and Y are independent. We thus get

$$f_U(u) = \int_{-\infty}^{\infty} f_{\frac{X}{Y}|Y=v}(u) f_Y(v) dv = \int_{-\infty}^{\infty} f_{\frac{X}{v}}(u) f_Y(v) dv.$$

By the change of variable formula in the univariate case, we get

$$f_{\frac{X}{v}}(u) = f_X(uv) \left| \frac{d}{du}(uv) \right| = f_X(uv)|v|.$$

Thus

$$\begin{aligned} f_U(u) &= \int_{-\infty}^{\infty} f_X(uv)|v| f_Y(v) dv \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2v^2}{2}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v^2}{2}\right) |v| dv \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{(1+u^2)v^2}{2}\right) |v| dv \\ &= 2\int_{0}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{(1+u^2)v^2}{2}\right) v dv = \frac{1}{\pi(1+u^2)}. \end{aligned}$$

The last equality is derived by the change of variable $w = v^2/2$ to evaluate the integral. We have therefore proved that U has the Cauchy density.

The Cauchy density is a special case of the t-density when the degrees of freedom is equal to one (i.e., the t-density with one degree of freedom is the same as the Cauchy density). The t-density for n degrees of freedom can also be derived as a consequence of the law of total probability (this is done in the next example).

Example 13.3. *Suppose $Z, X_1, \ldots, X_n$ are independent random variables all having the standard normal distribution. The distribution of the random variable*

$$T := \frac{Z}{\sqrt{\frac{X_1^2+\ldots X_n^2}{n}}}$$

is said to be the t-distribution with n degrees of freedom. Its density can be calculated using the Law of Total Probability as shown below. First let

$$V := X_1^2 + \cdots + X_n^2 \quad \textit{and note that} \quad V \sim \chi_n^2 = \textit{Gamma}\left(\frac{n}{2}, \frac{1}{2}\right).$$

As a result

$$\begin{aligned}
f_T(t) &= f_{\frac{Z}{\sqrt{V/n}}}(t) \\
&= \int_0^\infty f_{\frac{Z}{\sqrt{V/n}}|V=v}(t) f_V(v) dv \\
&= \int_0^\infty f_{\frac{Z}{\sqrt{v/n}}|V=v}(t) f_V(v) dv \\
&= \int_0^\infty f_{\frac{Z}{\sqrt{v/n}}}(t) f_V(v) dv \\
&= \int_0^\infty f_Z\left(t\sqrt{\frac{v}{n}}\right)\sqrt{\frac{v}{n}} f_V(v) dv \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{vt^2}{2n}\right)\sqrt{\frac{v}{n}} \frac{(1/2)^{n/2}}{\Gamma(n/2)} v^{(n/2)-1} e^{-v/2} dv \\
&= \frac{1}{\sqrt{2\pi n}} \frac{(1/2)^{n/2}}{\Gamma(n/2)} \int_0^\infty \exp\left(-\frac{v}{2}\left(1+\frac{t^2}{n}\right)\right) v^{(n/2)-(1/2)} dv.
\end{aligned}$$

The integrand in the integral above is equal to the main part of the $Gamma(\alpha, \lambda)$ density with

$$\alpha = \frac{n+1}{2} \quad \textit{and} \quad \lambda = \frac{1}{2}\left(1+\frac{t^2}{n}\right).$$

Thus the value of the integral is simply the normalization constant of the Gamma density:

$$\int_0^\infty \exp\left(-\frac{v}{2}\left(1+\frac{t^2}{n}\right)\right) v^{(n/2)-(1/2)} dv = \frac{\Gamma(\alpha)}{\lambda^\alpha} = \Gamma\left(\frac{n+1}{2}\right) 2^{(n+1)/2} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

We have thus proved:

$$\begin{aligned}
f_T(t) &= \frac{1}{\sqrt{2\pi n}} \frac{(1/2)^{n/2}}{\Gamma(n/2)} \Gamma\left(\frac{n+1}{2}\right) 2^{(n+1)/2} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \\
&= \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{\pi n}\Gamma\left(\frac{n}{2}\right)} \left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}.
\end{aligned}$$

This density, which is proportional to $\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}$*, is the t-density with* n *degrees of freedom. When* $n=1$*, this density if proportional to* $(1+t^2)^{-1}$ *so the t-density with 1 degree of freedom is exactly equal to the Cauchy density. When* n *becomes large, the tails of the t-density become less heavy and it eventually becomes the standard normal density. Indeed, when* n *is large, we can write (for each fixed* u*)*

$$\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}} \approx \exp\left(-t^2\frac{n+1}{2n}\right) \approx e^{-t^2/2}.$$

13.2.4 Bayes Rule for Continuous Random Variables

A direct consequence of the definition of the conditional density is:

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

This is the Bayes rule and it is useful for calculating the conditional density of Y given $X=x$ from knowledge of the conditional density of X given $Y=y$ (and the marginal density of Y). We shall see many applications of this rule in the next few lectures.

14 Lecture Fourteen

14.1 Recap: Last Class

In the last class, we looked at conditional densities for continuous random variables. Given two continuous random variables X and Y having a joint density $f_{X,Y}(x,y)$, the conditional density of X given $Y=y$ is defined as

$$f_{X|Y=y}(x) := \frac{f_{X,Y}(x,y)}{f_Y(y)}. \tag{70}$$

This definition makes sense when $f_Y(y) > 0$. Also when $f_Y(y) > 0$, the above is a valid density in x i.e., $f_{X|Y=y}(x) \geq 0$ for all x and

$$\int_{-\infty}^{\infty} f_{X|Y=y}(x)dx = 1.$$

Using the conditional density, the conditional probabilities involving X given $Y=y$ are calculated as:

$$\mathbb{P}\{X \in A \mid Y=y\} := \int_A f_{X|Y=y}(x)dx.$$

14.2 Law of Total Probability (LTP) and Bayes Rule for Continuous Variables

Conditional densities are used all over the place in probability and statistics. They are particulary useful while making probability assignments in Bayesian analyses. Here it is convenient to denote the two random variables by Θ and X (as opposed to X and Y). Θ typically denotes an unobserved parameter while X denotes observed data. A Bayesian analysis starts by making an assignment for the probability distribution of Θ and X. Directly modeling the joint density is usually difficult. One therefore models the marginal distribution

of Θ (this is called the prior distribution) and the conditional distribution of X given $\Theta = \theta$ (this is called the likelihood):

$$f_\Theta(\theta) \quad \text{and} \quad f_{X|\Theta=\theta}(x).$$

From these, the joint density can be written as

$$f_{\Theta,X}(\theta, x) = f_\theta(\theta) f_{X|\Theta}(x).$$

This completely specifies the joint probability distribution of Θ and X. Unspecified quantities such as the marginal distribution of X and the conditional distribution of Θ given $X = x$ are then calculated by using the rules of probability.

For calculating the marginal distribution of X, we use the law of total probability:

$$f_X(x) = \int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta.$$

For calculating the conditional distribution of Θ given $X = x$, we use the Bayes rule:

$$\begin{aligned} f_{\Theta|X=x}(\theta) &= \frac{f_{X|\Theta=\theta}(x) f_\Theta(\theta)}{f_X(x)} \\ &= \frac{f_{X|\Theta=\theta}(x) f_\Theta(\theta)}{\int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta} \\ &\propto f_{X|\Theta=\theta}(x) f_\Theta(\theta). \end{aligned}$$

In the statistical context, we shall refer to $f_{\Theta|X=x}(\cdot)$ as the posterior density of Θ and $f_X(\cdot)$ as the Evidence.

Here are a simple application of these formulae. More interesting examples will be studied later.

Example 14.1. *Suppose* $\Theta \sim N(\mu, \tau^2)$ *and* $X|\Theta = \theta \sim N(\theta, \sigma^2)$. *Then*

$$X \sim N(\mu, \tau^2 + \sigma^2) \quad \textit{and} \quad \Theta|X = x \sim N\left(\frac{x/\sigma^2 + \mu/\tau^2}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\sigma^2 + 1/\tau^2}\right). \tag{71}$$

I will sketch the proof of the above results below. The intuition behind the posterior distribution is as follows. For a normal density with mean m *and variance* v^2*, the inverse of the variance* $1/v^2$ *is called the precision. Skinnier normal distributions have high precision and vice versa.*

The formula for the posterior distribution given above implies that the precision of the conditional distribution of Θ *given* $X = x$ *equals the sum of the precisions of the distribution of* Θ *and the distribution of* X *respectively which means that the posterior is skinnier compared to the prior and the likelihood normal distributions. Also the mean of the posterior distribution equals a weighted linear combination of the prior mean and the data with the weights being proportional to the precisions.*

To derive the first part of (71)*, we use the LTP:*

$$f_X(x) = \int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta$$

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

$$\frac{(\theta-\mu)^2}{\tau^2}+\frac{(x-\theta)^2}{\sigma^2}=\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\theta^2-2\theta\left(\frac{\mu}{\tau^2}+\frac{x}{\sigma^2}\right)+\frac{\mu^2}{\tau^2}+\frac{x^2}{\sigma^2}$$
$$=\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\theta-\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2+\frac{(x-\mu)^2}{\tau^2+\sigma^2}$$

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

$$X\sim N(0,\tau^2+\sigma^2).$$

For the posterior distribution in (71)*, we use the Bayes rule (and the above derived expressions for* $f_{X|\Theta=\theta}(x)f_\Theta(\theta)$ *and* $f_X(x)$*):*

$$f_{\Theta|X=x}(\theta)=\frac{f_{X|\Theta=\theta}(x)f_\Theta(\theta)}{f_X(x)}=\frac{\sqrt{\tau^2+\sigma^2}}{\sqrt{2\pi\tau^2\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{1}{\tau^2}+\frac{1}{\sigma^2}\right)\left(\theta-\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2}\right)^2\right)$$

which immediately implies:

$$\Theta|X=x\sim N\left(\frac{x/\sigma^2+\mu/\tau^2}{1/\sigma^2+1/\tau^2},\frac{1}{1/\sigma^2+1/\tau^2}\right).$$

14.3 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 marginal 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 marginal 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.

14.3.1 X and Θ are both discrete

The LTP is

$$\mathbb{P}\{X = x\} = \sum_{\theta} \mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}$$

and the Bayes rule is

$$\mathbb{P}\{\Theta = \theta|X = x\} = \frac{\mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\mathbb{P}\{X = x\}} = \frac{\mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\sum_{\theta} \mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}.$$

14.3.2 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

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x) f_{\Theta}(\theta)}{f_X(x)} = \frac{f_{X|\Theta=\theta}(x) f_{\Theta}(\theta)}{\int f_{X|\Theta=\theta}(x) f_{\Theta}(\theta) dx}.$$

14.3.3 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

$$f_{\Theta|X=x}(\theta) = \frac{\mathbb{P}\{X = x|\Theta = \theta\} f_{\Theta}(\theta)}{\mathbb{P}\{X = x\}} = \frac{\mathbb{P}\{X = x|\Theta = \theta\} f_{\Theta}(\theta)}{\int \mathbb{P}\{X = x|\Theta = \theta\} f_{\Theta}(\theta) d\theta}.$$

14.3.4 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 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 $X = x$.

The following is an example of the LTP and Bayes Rule when Θ is continuous and X is discrete.

Example 14.2. *Suppose that Θ has the $Beta(\alpha, \beta)$ distribution on $(0, 1)$ and let $X|\Theta = \theta$ has the binomial distribution with parameters n and 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 $0, 1, \ldots, n$) 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 $x = 0, 1, \ldots, n$)

$$\begin{aligned}\mathbb{P}\{X = x\} &= \int \mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta) d\theta \\ &= \int_0^1 \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)} d\theta \\ &= \binom{n}{x} \frac{B(x+\alpha, n-x+\beta)}{B(\alpha, \beta)}.\end{aligned}$$

Let us now calculate the posterior distribution of Θ given $X = x$. Using the Bayes rule, we obtain

$$\begin{aligned}f_{\Theta|X=x}(\theta) &\propto \mathbb{P}\{X = x|\Theta = \theta\} f_\Theta(\theta) \propto \theta^x (1-\theta)^{n-x} \theta^{\alpha-1}(1-\theta)^{\beta-1} I\{0 < \theta < 1\} \\ &= \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} I\{0 < \theta < 1\}.\end{aligned}$$

Thus

$$\Theta|X = x \sim Beta(x + \alpha, n - x + \beta).$$

15 Lecture Fifteen

15.1 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 marginal 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 marginal distribution of Θ.

The precise formulae for the LTP and Bayes Rule can be broken down into four cases according to whether X and/or Θ is discrete or continuous.

15.1.1 X and Θ are both discrete

The LTP is

$$\mathbb{P}\{X = x\} = \sum_\theta \mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}$$

and the Bayes rule is

$$\mathbb{P}\{\Theta = \theta|X = x\} = \frac{\mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\mathbb{P}\{X = x\}} = \frac{\mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}{\sum_\theta \mathbb{P}\{X = x|\Theta = \theta\}\mathbb{P}\{\Theta = \theta\}}.$$

15.1.2 X and Θ are both continuous

Here LTP is

$$f_X(x) = \int f_{X|\Theta=\theta}(x) f_\Theta(\theta) d\theta \tag{72}$$

and Bayes rule is

$$f_{\Theta|X=x}(\theta) = \frac{f_{X|\Theta=\theta}(x)f_\Theta(\theta)}{f_X(x)} = \frac{f_{X|\Theta=\theta}(x)f_\Theta(\theta)}{\int f_{X|\Theta=\theta}(x)f_\Theta(\theta)dx}.$$

15.1.3 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

$$f_{\Theta|X=x}(\theta) = \frac{\mathbb{P}\{X = x|\Theta = \theta\}f_\Theta(\theta)}{\mathbb{P}\{X = x\}} = \frac{\mathbb{P}\{X = x|\Theta = \theta\}f_\Theta(\theta)}{\int \mathbb{P}\{X = x|\Theta = \theta\}f_\Theta(\theta)d\theta}.$$

15.1.4 X is continuous while Θ is discrete

LTP is

$$f_X(x) = \sum_\theta f_{X|\Theta=\theta}(x)\mathbb{P}\{\Theta = \theta\} \tag{73}$$

and Bayes rule is

$$\mathbb{P}\{\Theta = \theta|X = x\} = \frac{f_{X|\Theta=\theta}(x)\mathbb{P}\{\Theta = \theta\}}{f_X(x)} = \frac{f_{X|\Theta=\theta}(x)\mathbb{P}\{\Theta = \theta\}}{\sum_\theta f_{X|\Theta=\theta}(x)\mathbb{P}\{\Theta = \theta\}} \tag{74}$$

These formulae are useful when the conditional distribution of X given $\Theta = \theta$ as well as the marginal distribution of Θ 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 look at some more applications of these formulae today.

15.2 A Simple Model Selection Application

Suppose Θ has the $Ber(0.5)$ distribution i.e.,

$$\mathbb{P}\{\Theta = 0\} = \mathbb{P}\{\Theta = 1\} = 0.5.$$

Next assume that $X_1, \ldots, X_n$ have the following distributions conditional on $\Theta = \theta$:

$$X_1, \ldots, X_n \mid \Theta = 0 \overset{\text{i.i.d}}{\sim} f_0 \qquad \text{where } f_0(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$$

and

$$X_1, \ldots, X_n \mid \Theta = 1 \overset{\text{i.i.d}}{\sim} f_1 \qquad \text{where } f_1(x) := \frac{1}{\sqrt{2\pi}} \exp\left(-|x|\sqrt{\frac{2}{\pi}}\right).$$

f_0 is the standard normal density and f_1 is a Laplace density. Both densities have the same maximal value of $1/\sqrt{2\pi}$. Based on the information given, calculate the conditional distribution of Θ given $X_1 = x_1, X_2 = x_2, \ldots, X_6 = x_6$ (i.e., $n = 6$) where

$$x_1 = -0.55, x_2 = -1.11, x_3 = 1.23, x_4 = 0.29, x_5 = 1.56, x_6 = -1.64. \tag{75}$$

Here is the context for this question. We observe data $x_1, \ldots, x_n$ with $n = 6$. We want to use one of the models f_0 or f_1 for this data. The random variable Θ is used to describe the choice of the model. We want to treat both the models on an equal footing so we assumed that Θ has the uniform prior distribution on $\{0, 1\}$.

To calculate the conditional distribution of Θ given the data, we use the formula (74) because Θ is discrete and the data $X_1, \ldots, X_n$ are continuous. This gives

$$\begin{aligned}
&\mathbb{P}\{\Theta = 0 \mid X_1 = x_1, \ldots, X_n = x_n\} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n)\mathbb{P}\{\Theta = 0\}}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n)\mathbb{P}\{\Theta = 0\} + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n)\mathbb{P}\{\Theta = 1\}} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) \times \frac{1}{2}}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) \times \frac{1}{2} + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n) \times \frac{1}{2}} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n)}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n)} \\
&= \frac{f_0(x_1)f_0(x_2)\ldots f_0(x_n)}{f_0(x_1)f_0(x_2)\ldots f_0(x_n) + f_1(x_1)f_1(x_2)\ldots f_1(x_n)} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right) + \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sqrt{\frac{2}{\pi}}\sum_{i=1}^n |x_i|\right)}.
\end{aligned}$$

Similarly

$$\begin{aligned}
&\mathbb{P}\{\Theta = 0 \mid X_1 = x_1, \ldots, X_n = x_n\} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sqrt{\frac{2}{\pi}}\sum_{i=1}^n |x_i|\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=1}^n x_i^2\right) + \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\sqrt{\frac{2}{\pi}}\sum_{i=1}^n |x_i|\right)}.
\end{aligned}$$

Plugging in the above formula the data values given in (75) for $x_1, \ldots, x_6$, we obtain

$$\mathbb{P}\{\Theta = 0 \mid X_1 = x_1, \ldots, X_6 = x_6\} = 0.72 \quad \text{and} \quad \mathbb{P}\{\Theta = 1 \mid X_1 = x_1, \ldots, X_6 = x_6\} = 0.28$$

Thus, conditioning on the data, we have a 72% probability for the normal model compared to 28% probability for the Laplace model. Now suppose that we add in an additional observation $x_7 = 5$. It can be checked that

$$\mathbb{P}\{\Theta = 0 \mid X_1 = x_1, \ldots, X_7 = x_7\} = 0.001 \quad \text{and} \quad \mathbb{P}\{\Theta = 1 \mid X_1 = x_1, \ldots, X_7 = x_7\} = 0.999$$

Now there is overwhelming preference for the Laplace model. This is because $x_7 = 5$ is an outlying observation to which the Laplace model gives much higher probability compared to the Normal model owing to heavy tails of the Laplace density.

15.3 Model Selection with unknown parameters

Suppose Θ has the $Ber(0.5)$ distribution i.e.,

$$\mathbb{P}\{\Theta = 0\} = \mathbb{P}\{\Theta = 1\} = 0.5.$$

Next assume that $X_1, \ldots, X_n$ have the following distributions conditional on $\Theta = \theta$:

$$X_1, \ldots, X_n \mid \Theta = 0 \overset{\text{i.i.d}}{\sim} N(0, \sigma_0^2) \qquad \text{for some } \sigma_0 > 0$$

and

$$X_1, \ldots, X_n \mid \Theta = 1 \overset{\text{i.i.d}}{\sim} Lap(0, \sigma_1) \qquad \text{for some } \sigma_1 > 0.$$

Here $Lap(0, \sigma_1)$ denotes the Laplace density centered at 0 and having scale σ_1; its density is given by

$$x \mapsto \frac{1}{2\sigma_1} \exp\left(-\frac{|x|}{\sigma_1}\right).$$

Based on this information, calculate the conditional distribution of Θ given $X_1 = x_1, \ldots, X_n = x_n$ where $n = 10$ and $x_1, \ldots, x_n$ are given by

$$-0.69, -4.26, 0.14, -0.86, 0.42, 24.21, 0.51, -1.23, 2.30, 4.15. \tag{76}$$

Once again, this is a model selection problem where we need to choose between the normal model and the Laplace model based on the observed data given above. We can proceed exactly as in the last section and write down the conditional probabilities of Θ given $X_1 = x_1, \ldots, X_n = x_n$. However the answers would depend on σ_0 and σ_1. We would not be able to make a decision between the two models because of this annoying dependence on σ_0, σ_1. To get rid of the dependence on the specific values of σ_0, σ_1, a natural strategy is to treat σ_0 and σ_1 as unknown parameters and further make distributional assumptions on them to reflect our ignorance of their precise values. One way of doing this is to assume that:

$$\log \sigma_0 \mid \Theta = 0 \sim \text{Unif}(-C, C) \quad \text{and} \quad X_1, \ldots, X_n \mid \sigma_0, \Theta = 0 \overset{\text{i.i.d}}{\sim} N(0, \sigma_0^2) \tag{77}$$

as well as

$$\log \sigma_1 \mid \Theta = 1 \sim \text{Unif}(-C, C) \quad \text{and} \quad X_1, \ldots, X_n \mid \sigma_1, \Theta = 1 \overset{\text{i.i.d}}{\sim} Lap(0, \sigma_1) \tag{78}$$

for a large constant value C. In other words, we are using the uniform distribution on $(-C, C)$ to reflect our ignorance of $\log \sigma_0$ and $\log \sigma_1$. We can now calculate the conditional distribution of Θ given $X_1 = x_1, \ldots, X_n = x_n$ in the following way. As in the previous section, we first obtain

$$\begin{aligned}
&\mathbb{P}\{\Theta = 0 \mid X_1 = x_1, \ldots, X_n = x_n\} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n)\mathbb{P}\{\Theta = 0\}}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n)\mathbb{P}\{\Theta = 0\} + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n)\mathbb{P}\{\Theta = 1\}} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) \times \frac{1}{2}}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) \times \frac{1}{2} + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n) \times \frac{1}{2}} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n)}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n)}
\end{aligned} \tag{79}$$

and similarly

$$\mathbb{P}\{\Theta = 1 \mid X_1 = x_1, \ldots, X_n = x_n\} = \frac{f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n)}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n)}. \tag{80}$$

We therefore need to calculate

$$f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) \quad \text{and} \quad f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n)$$

These densities are not directly given to us (unlike in the problem of the previous section) but they are given conditionally on the parameters σ_0 and σ_1. We shall therefore calculate

them using the Law of Total Probability (72) (note that $X_1, \ldots, X_n$ as well as σ_0, σ_1 are all continuous parameters). We thus have

$$\begin{aligned}
f_{X_1,\ldots,X_n|\Theta=0}(x_1,\ldots,x_n) &= \int f_{X_1,\ldots,X_n|\Theta=0,\sigma_0}(x_1,\ldots,x_n) f_{\sigma_0|\Theta=0}(\sigma_0) d\sigma_0 \\
&= \int \left[\prod_{i=1}^n f_{X_i|\Theta=0,\sigma_0}(x_i)\right] f_{\sigma_0|\Theta=0}(\sigma_0) d\sigma_0 \\
&= \int \left[\prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{x_i^2}{2\sigma_0^2}\right)\right] f_{\sigma_0|\Theta=0}(\sigma_0) d\sigma_0 \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \int \sigma_0^{-n} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2}\right) f_{\sigma_0|\Theta=0}(\sigma_0) d\sigma_0.
\end{aligned}$$

Because we assumed that $\log \sigma_0$ has the uniform distribution on $(-C, C)$ (conditional on $\Theta = 0$), we get

$$f_{\sigma_0|\Theta=0}(\sigma_0) = f_{\log\sigma_0|\Theta=0}(\log\sigma_0)\left|\frac{d}{d\sigma_0}(\log\sigma_0)\right| = f_{\log\sigma_0|\Theta=0}(\log\sigma_0)\frac{1}{\sigma_0} = \frac{I\{-C < \log\sigma_0 < C\}}{2C\sigma_0}.$$

As a result

$$\begin{aligned}
f_{X_1,\ldots,X_n|\Theta=0}(x_1,\ldots,x_n) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n \int \sigma_0^{-n} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2}\right) \frac{I\{-C < \log\sigma_0 < C\}}{2C\sigma_0} d\sigma_0 \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C} \int_{e^{-C}}^{e^C} \sigma_0^{-n-1} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2}\right) d\sigma_0.
\end{aligned}$$

Because C is large, the limits in the above integral will effectively be between 0 and ∞ (as $e^{-C} \approx 0$ and $e^C \approx \infty$). Thus

$$f_{X_1,\ldots,X_n|\Theta=0}(x_1,\ldots,x_n) \approx \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C} \int_0^{\infty} \sigma_0^{-n-1} \exp\left(-\frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2}\right) d\sigma_0.$$

To evaluate the integral above, we use the change of variable

$$u = \frac{\sum_{i=1}^n x_i^2}{2\sigma_0^2} \quad \text{so that} \quad \sigma_0 = \sqrt{\frac{\sum_{i=1}^n x_i^2}{2}} u^{-1/2} \quad \text{and} \quad d\sigma_0 = \sqrt{\frac{\sum_{i=1}^n x_i^2}{2}} \left(-\frac{1}{2}u^{-3/2}\right) du$$

which gives

$$\begin{aligned}
&f_{X_1,\ldots,X_n|\Theta=0}(x_1,\ldots,x_n) \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C} \int_0^{\infty} \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-(n+1)/2} u^{(n+1)/2} e^{-u} \sqrt{\frac{\sum_{i=1}^n x_i^2}{2}} \left(\frac{1}{2}u^{-3/2}\right) du \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C} \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2} \int_0^{\infty} u^{(n/2)-1} e^{-u} du \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C} \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2}\Gamma\left(\frac{n}{2}\right).
\end{aligned}$$

Similarly

$$\begin{aligned}
f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) &= \int f_{X_1,\ldots,X_n|\Theta=1,\sigma_1}(x_1,\ldots,x_n) f_{\sigma_1|\Theta=1}(\sigma_1) d\sigma_1 \\
&= \int \left[\prod_{i=1}^{n} f_{X_i|\Theta=1,\sigma_1}(x_i)\right] f_{\sigma_0|\Theta=1}(\sigma_1) d\sigma_1 \\
&= \int \left[\prod_{i=1}^{n} \frac{1}{2\sigma_1} \exp\left(-\frac{|x_i|}{\sigma_1}\right)\right] \frac{I\{-C < \log\sigma_1 < C\}}{2C\sigma_1} d\sigma_1 \\
&= \left(\frac{1}{2}\right)^n \frac{1}{2C} \int_{e^{-C}}^{e^C} \sigma_1^{-n-1} \exp\left(-\frac{\sum_{i=1}^n |x_i|}{\sigma_1}\right) d\sigma_1 \\
&\approx \left(\frac{1}{2}\right)^n \frac{1}{2C} \int_{0}^{\infty} \sigma_1^{-n-1} \exp\left(-\frac{\sum_{i=1}^n |x_i|}{\sigma_1}\right) d\sigma_1
\end{aligned}$$

The change of variable

$$u = \frac{\sum_{i=1}^n |x_i|}{\sigma}$$

leads to

$$\begin{aligned}
&f_{X_1,\ldots,X_n|\Theta=1}(x_1,\ldots,x_n) \\
&= \left(\frac{1}{2}\right)^n \frac{1}{2C} \left(\sum_{i=1}^n |x_i|\right)^{-n} \int_0^\infty u^{n-1} e^{-u} du = \left(\frac{1}{2}\right)^n \frac{1}{2C} \left(\sum_{i=1}^n |x_i|\right)^{-n} \Gamma(n).
\end{aligned}$$

Plugging these expressions in (79) and (80), we obtain

$$\begin{aligned}
&\mathbb{P}\{\Theta = 0 \mid X_1 = x_1, \ldots, X_n = x_n\} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C} \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2}\Gamma\left(\frac{n}{2}\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C} \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2}\Gamma\left(\frac{n}{2}\right) + \left(\frac{1}{2}\right)^n \frac{1}{2C} \left(\sum_{i=1}^n |x_i|\right)^{-n} \Gamma(n)} \\
&= \frac{\left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2}\Gamma\left(\frac{n}{2}\right)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2}\Gamma\left(\frac{n}{2}\right) + \left(\frac{1}{2}\right)^n \left(\sum_{i=1}^n |x_i|\right)^{-n} \Gamma(n)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\{\Theta = 1 \mid X_1 = x_1, \ldots, X_n = x_n\} \\
&= \frac{\left(\frac{1}{2}\right)^n \left(\sum_{i=1}^n |x_i|\right)^{-n} \Gamma(n)}{\left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2}\Gamma\left(\frac{n}{2}\right) + \left(\frac{1}{2}\right)^n \left(\sum_{i=1}^n |x_i|\right)^{-n} \Gamma(n)}.
\end{aligned}$$

Now the observed data values in (76) can be plugged in to compute the posterior probabilities. This gives

$$\mathbb{P}\{\Theta = 0 \mid X_1 = x_1, \ldots, X_n = x_n\} = 0.008 \text{ and } \mathbb{P}\{\Theta = 1 \mid X_1 = x_1, \ldots, X_n = x_n\} = 0.992$$

Thus the observed data in (76) (which seems to contain some outliers) overwhelmingly favors the Laplace model compared to the Normal model.

15.3.1 Considering one more model

Suppose now that Θ has the distribution given by:

$$\mathbb{P}\{\Theta = 0\} = \mathbb{P}\{\Theta = 1\} = \mathbb{P}\{\Theta = 2\} = \frac{1}{3}$$

and that $X_1, \ldots, X_n$ have the following distributions conditional on $\Theta = \theta$:

$$X_1, \ldots, X_n \mid \Theta = 0 \overset{\text{i.i.d}}{\sim} N(0, \sigma_0^2) \qquad \text{for some } \sigma_0 > 0$$

and

$$X_1, \ldots, X_n \mid \Theta = 1 \overset{\text{i.i.d}}{\sim} Lap(0, \sigma_1) \qquad \text{for some } \sigma_1 > 0.$$

and

$$X_1, \ldots, X_n \mid \Theta = 2 \overset{\text{i.i.d}}{\sim} C(0, \sigma_2) \qquad \text{for some } \sigma_2 > 0.$$

Here $C(0, \sigma_2)$ is the Cauchy density with location parameter 0 and scale parameter σ_2. This density is given by $\frac{1}{\pi}\frac{\sigma_2}{x^2+\sigma_2^2}$. What then is the conditional distribution of Θ given $X_1 = x_1, \ldots, X_n = x_n$ for the same data (76)?

This is basically the same problem as that considered in the previous section except that we are considering the Cauchy model in addition to the normal and the Laplace models.

Using the Bayes rule, it is easy to see that

$$\begin{aligned}
&\mathbb{P}\left\{\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n\right\} \\
&= \frac{f_{X_1,\ldots,X_n|\Theta=\theta}(x_1, \ldots, x_n)}{f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) + f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n) + f_{X_1,\ldots,X_n|\Theta=2}(x_1, \ldots, x_n)}
\end{aligned}$$

for each $\theta = 0, 1, 2$. The calculation for $f_{X_1,\ldots,X_n|\Theta=\theta}(x_1, \ldots, x_n)$ for $\theta = 0$ and $\theta = 1$ is done based on the assumptions (77) and (78), and this leads to exactly the same values as in the previous section. More specifically

$$f_{X_1,\ldots,X_n|\Theta=0}(x_1, \ldots, x_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \frac{1}{2C}\left(\frac{\sum_{i=1}^n x_i^2}{2}\right)^{-n/2} \frac{1}{2}\Gamma\left(\frac{n}{2}\right).$$

and

$$f_{X_1,\ldots,X_n|\Theta=1}(x_1, \ldots, x_n) = \left(\frac{1}{2}\right)^n \frac{1}{2C}\left(\sum_{i=1}^n |x_i|\right)^{-n} \Gamma(n).$$

For $f_{X_1,\ldots,X_n|\Theta=2}(x_1, \ldots, x_n)$, we shall make the assumption (analogous to (77) and (78)):

$$\log \sigma_2 \mid \Theta = 2 \sim \text{Unif}(-C, C) \quad \text{and} \quad X_1, \ldots, X_n \mid \sigma_2, \Theta = 2 \overset{\text{i.i.d}}{\sim} C(0, \sigma_2).$$

This leads to

$$\begin{aligned}
f_{X_1,\ldots,X_n|\Theta=2}(x_1, \ldots, x_n) &= \int_{e^{-C}}^{e^C} \left[\prod_{i=1}^n \left(\frac{1}{\pi}\frac{\sigma_2}{x_i^2 + \sigma_2^2}\right)\right] \frac{1}{2C\sigma_2} d\sigma_2 \\
&= \int_0^{\infty} \left[\prod_{i=1}^n \left(\frac{1}{\pi}\frac{\sigma_2}{x_i^2 + \sigma_2^2}\right)\right] \frac{1}{2C\sigma_2} d\sigma_2.
\end{aligned}$$

It is probably difficult to calculate this integral in closed form. But it is quite straightforward to compute this numerically (in the code file, I computed this after the change of variable $t = \log \sigma_2$).

For the dataset in (76), the above analysis leads to

$$\mathbb{P}\left\{\Theta = \theta \mid X_1 = x_1, \ldots, X_n = x_n\right\} = \begin{cases} 0.0003 & \text{for } \theta = 0 \\ 0.0417 & \text{for } \theta = 1 \\ 0.958 & \text{for } \theta = 2 \end{cases}$$

Thus the Cauchy model has the highest probability. The Laplace model which received the highest probability when only the two models (Normal and Laplace) were being considered now gets low probability when we are also considering the Cauchy model.

This approach for Model Selection is often known as Bayesian Model Selection. An important role in this approach is played by the quantities $f_{X_1,\ldots,X_n|\Theta=\theta}(x_1, \ldots, x_n)$ for different values of θ. In the Machine Learning literature, this quantity is known as the **Evidence** for the model represented by $\Theta = \theta$ in light of the observed data $x_1, \ldots, x_n$. When each individual model contains additional parameters (such as in the present case where the i^{th} model is expressed in terms of the parameter σ_i), the Evidence is calculated (via the Law of Total Probability) as the integral of the probability for each value of the parameter with respect to a prior on the parameter. Thus the Evidence for a model is also known as the **Integrated Likelihood** of the model.

For more on Bayesian model selection using Evidences, see Jaynes [1, Chapter 20] or MacKay [5, Chapter 28].

16 Lecture Sixteen

16.1 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.

16.1.1 Law of Iterated/Total Expectation

The law of total expectation states that

$$\mathbb{E}(Y) = \begin{cases} \int \mathbb{E}(Y|X=x)\, f_X(x)dx & : \text{ if } X \text{ is continuous} \\ \sum_x \mathbb{E}(Y|X=x)\, \mathbb{P}\{X=x\} & : \text{ if } X \text{ is discrete} \end{cases}$$

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 16.1. *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,

$$\mathbb{E}(Y) = \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(X/2) = \ell/4.$$

Example 16.2. *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 16.3 (Sum of a random number of i.i.d random variables)**.** *Suppose $X_1, X_2, \ldots$ 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 $\{1, 2, \ldots, \}$ and that is independent of $X_1, X_2, \ldots$. 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****.*

16.1.2 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:

$$\mathbb{E}\left[(Y - g(x))^2 \,|X = x\right]$$

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

$$R(g) := \mathbb{E}\,|g(X) - Y|$$

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

$$R(g) = \mathbb{E}\,|g(X) - Y| = \mathbb{E}\left\{\mathbb{E}\left[|g(X) - Y|\,|X\right]\right\}$$

The value $g^*(x)$ which minimizes the inner expectation:

$$\mathbb{E}\left[|Y - g(x)|\,|X = x\right]$$

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

$$g^*(x) = \text{minimizer of } \mathbb{E}(L(Y,c)|X=x) \text{ over } c \in \mathbb{R}.$$

16.2 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,

$$Var(Y|X=x) := \mathbb{E}\left[\left(Y - \mathbb{E}(Y|X=x)\right)^2 |X=x\right] = \mathbb{E}\left(Y^2|X=x\right) - \left(\mathbb{E}(Y|X=x)\right)^2.$$

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

$$Var(Y|X) = \mathbb{E}(Y^2|X) - \left(\mathbb{E}(Y|X)\right)^2. \tag{81}$$

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 16.4. *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 16.5 (Sum of a random number of i.i.d random variables)**.** *Suppose $X_1, X_2, \ldots$ 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 $\{1, 2, \ldots, \}$ and that is independent of $X_1, X_2, \ldots$. 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.

$$Var(S) = \mathbb{E}(Var(S|N)) + Var(\mathbb{E}(S|N)) = \mathbb{E}(N\sigma^2) + Var(N\mu) = \sigma^2(\mathbb{E}N) + \mu^2 Var(N).$$

.

17 Lecture Seventeen

We shall study the multivariate normal and multivariate t-distributions today. Before going over them, let us first recall these densities in the univariate case.

17.1 Univariate normal and t densities

Let us start by looking at the standard normal distribution. We say that Z is standard normal if its density is given by

$$f_Z(z) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right).$$

The mean of Z is zero and its variance equals 1.

From standard normal, we define general normal distributions via a scale and location change. Specifically, for two real numbers μ and a, define

$$X = \mu + aZ.$$

The density of X is given by (using the change of variable formula):

$$f_X(x) = \frac{1}{\sqrt{2\pi}|a|} \exp\left(-\frac{(x-\mu)^2}{2a^2}\right).$$

This density depends on the two quantities μ and a^2 and is denoted by $N(\mu, a^2)$. Note that the quantity a can be positive or negative but the density only depends on $|a|$ or a^2. This density is called the Normal density with parameters μ and a^2. It is easy to check (using $X = \mu + aZ$) that the mean of X equals μ and variance of X equals a^2.

Next we come to the t-distribution. This is obtained by a further scale change involving an independent chi-squared distributed random variable. Specifically consider a random variable V that has the χ^2_k distribution (χ^2_k is the chi-squared distribution with k degrees of freedom; recall that $\chi^2_k = Gamma(k/2, 1/2)$) and assume that V and Z are independent. Define

$$T = \mu + \frac{1}{\sqrt{V/k}} aZ.$$

Thus T is very similar to X except for the additional scale change involving the random variable V/k. This random variable has mean and variance given by:

$$\mathbb{E}\left(\frac{V}{k}\right) = \frac{\mathbb{E}V}{k} = \frac{\mathbb{E}(\chi^2_k)}{k} = \frac{k}{k} = 1$$

and

$$\text{var}\left(\frac{V}{k}\right) = \frac{\text{var}(V)}{k^2} = \frac{\text{var}(\chi^2_k)}{k^2} = \frac{2k}{k^2} = \frac{2}{k}.$$

Thus when k is large, V/k is a random variable with mean 1 and very small variance which implies that V/k will be highly concentrated around 1. Thus the additional scale change involving V/k will play very little role if k is large. But if k is not very large, then it will make the distribution of T considerably different from that of X. The density of T can be explicitly calculated using the following argument.

$$f_T(y) = \int_0^\infty f_{T|V=x}(y) f_V(x) dx.$$

Observe now that

$$T \mid V = x \quad = \mu + \frac{aZ}{\sqrt{\frac{x}{k}}} \sim N\left(\mu, a^2 \frac{k}{x}\right)$$

so that

$$f_{T|V=x}(y) = \frac{\sqrt{x}}{\sqrt{2\pi} a \sqrt{k}} \exp\left(-\frac{x}{2a^2 k}(y-\mu)^2\right).$$

As a result

$$\begin{aligned} f_T(y) &= \int_0^\infty f_{T|V=x}(y) f_V(x) dx \\ &\propto \int_0^\infty \frac{\sqrt{x}}{\sqrt{2\pi} a \sqrt{k}} \exp\left(-\frac{x}{2a^2 k}(y-\mu)^2\right) x^{\frac{k}{2}-1} e^{-x/2} dx \\ &\propto \int_0^\infty x^{\frac{k}{2}-\frac{1}{2}} \exp\left(-\frac{x}{2}\left(1 + \frac{(y-\mu)^2}{ka^2}\right)\right) dx. \end{aligned}$$

The change of variable

$$t = x\left(1 + \frac{(y-\mu)^2}{ka^2}\right)$$

now leads to

$$f_T(y) \propto \frac{1}{\left(1 + \frac{(y-\mu)^2}{ka^2}\right)^{\frac{k+1}{2}}} \int_0^\infty t^{\frac{k}{2}-1} e^{-t/2} dt \propto \frac{1}{\left(1 + \frac{(y-\mu)^2}{ka^2}\right)^{\frac{k+1}{2}}}.$$

The density of this random variable T will be denoted by $t_k(\mu, a^2)$. In other words, the density of $t_k(\mu, a^2)$ is proportional to

$$y \mapsto \frac{1}{\left(1 + \frac{(y-\mu)^2}{ka^2}\right)^{\frac{k+1}{2}}}.$$

This density has heavier tails compared to the normal density $N(\mu, \sigma^2)$. The mean of $t_k(\mu, a^2)$ exists if and only if $k > 1$ and equals μ. Its variance exists if and only if $k > 2$ and equals $\frac{k}{k-2}a^2$.

17.2 Random Vectors and Covariance Matrices

In order to discuss the multivariate normal distribution, we shall the language of random vectors and covariance matrices which are defined next.

A finite number of random variables can be viewed together as a random vector. More precisely, a random vector is a vector whose entries are random variables. Let $Y = (Y_1, \ldots, Y_n)^T$ be an $n \times 1$ random vector. Its Expectation $\mathbb{E}Y$ is defined as a vector whose ith entry is the expectation of Y_i i.e., $\mathbb{E}Y = (\mathbb{E}Y_1, \mathbb{E}Y_2, \ldots, \mathbb{E}Y_n)^T$. 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 $Y_1, \ldots, Y_n$. 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 $Cov(Y_i, Y_j) = Cov(Y_j, Y_i)$.

One can also check:

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.

As a consequence of the second formula above, we get

$$var(a^TY) = a^TCov(Y)a = \sum_{i,j} a_i a_j Cov(Y_i, Y_j) \qquad \text{for every } n \times 1 \text{ vector } a.$$

17.3 Multivariate Normal and t-densities

We shall follow the same program as in the univariate case. We first define standard multivariate normal, then general multivariate normal followed by the multivariate t.

We say that a $p \times 1$ random vector Z has the standard p-variate normal distribution if its components $Z_1, \ldots, Z_p$ are independently distributed according to the standard normal distribution i.e., $Z_1, \ldots, Z_p \overset{\text{i.i.d}}{\sim} N(0,1)$. The joint density of $Z_1, \ldots, Z_p$ is then

$$\begin{aligned} f_{Z_1,\ldots,Z_p}(z_1, \ldots, z_p) &= \prod_{i=1}^{p} \left[\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z_i^2}{2}\right) \right] \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{\sum_{i=1}^p z_i^2}{2}\right) \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{\|z\|^2}{2}\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{z^T z}{2}\right) \end{aligned}$$

The mean vector of Z is simply the zero vector and the covariance matrix of Z is the $p \times p$ identity matrix I_p.

From the standard p-variate normal, we obtain a general p-variate normal distribution in the following way. Suppose μ is a fixed p-dimensional vector and suppose A is a fixed $p \times p$ **invertible** matrix. Define

$$X = \mu + AZ$$

Here AZ is the matrix-vector multiplication of the $p \times p$ matrix A with the $p \times 1$ vector Z. By the Jacobian formula, the joint density of the components $X_1, \ldots, X_p$ of X is given by

$$\begin{aligned} f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) &= f_{Z_1,\ldots,Z_p}(A^{-1}(x-\mu)) \left|\det(A^{-1})\right| \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{\left(A^{-1}(x-\mu)\right)^T \left(A^{-1}(x-\mu)\right)}{2}\right) \left|\det(A^{-1})\right| \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{(x-\mu)^T \left(A^{-1}\right)^T A^{-1}(x-\mu)}{2}\right) \left|\det(A^{-1})\right| \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{(x-\mu)^T \left(A^{T}\right)^{-1} A^{-1}(x-\mu)}{2}\right) \left|\det(A^{-1})\right| \\ &= \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{(x-\mu)^T \left(AA^{T}\right)^{-1} (x-\mu)}{2}\right) \left|\det(A^{-1})\right|. \end{aligned}$$

We now let

$$\Sigma := AA^T.$$

Because the determinant of a product of matrices equals the product of the determinants

$$\det(\Sigma) = \det(AA^T) = \det(A)\det(A^T) = (\det(A))^2$$

which implies, in particular, that $\det(\Sigma) > 0$. Using this (and the fact that the determinant of the inverse of a matrix equals the inverse of the determinant), we can write

$$\det(A^{-1}) = \frac{1}{\det A} = \frac{1}{\sqrt{\det \Sigma}}.$$

We can thus write

$$f_{X_1,\ldots,X_p}(x_1, \ldots, x_p) = \left(\frac{1}{\sqrt{2\pi}}\right)^p \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right) \frac{1}{\sqrt{\det \Sigma}}.$$

This density depends on the vector μ as well as on the matrix $\Sigma = AA^T$. It is therefore denoted by $N_p(\mu, \Sigma)$. It is easy to see (using $X = \mu + AZ$) that μ is the mean vector of X

and Σ is the covariance matrix of X:

$$\mathbb{E}(X) = \mathbb{E}(\mu + AZ) = \mu + A\mathbb{E}(Z) = \mu$$
$$\text{Cov}(X) = \text{Cov}(\mu + AZ) = A\text{Cov}(Z)A^T = A(I_p)A^T = AA^T = \Sigma.$$

Let us now define the multivariate t-density. This is obtained by changing the scale of the multivariate normal density via a chi-squared random variable. Specifically, let V denote a χ^2_k random variable that is independent of a p-variate standard normal vector Z. Let

$$T := \mu + \frac{1}{\sqrt{V/k}} AZ$$

for a $p \times 1$ vector μ and an invertible $p \times p$ matrix A. Note that T is given by

$$\begin{pmatrix} T_1 \\ \cdot \\ \cdot \\ \cdot \\ T_p \end{pmatrix} = \begin{pmatrix} \mu_1 + \frac{\sum_{j=1}^p A(1,j)Z_j}{\sqrt{\frac{V}{k}}} \\ \cdot \\ \cdot \\ \cdot \\ \mu_p + \frac{\sum_{j=1}^p A(p,j)Z_j}{\sqrt{\frac{V}{k}}} \end{pmatrix}. \tag{82}$$

Note that the scale change on each component of T is through the same scalar random variable V.

The distribution of this random vector T will be denoted by $t_{k,p}(\mu, \Sigma)$. Its density can be derived just as in the univariate case in the following way:

$$f_{T_1,\ldots,T_p}(y_1, \ldots, y_p) = \int_0^\infty f_{T_1,\ldots,T_p|V=x}(y_1, \ldots, y_p) f_V(x) dx.$$

Observe that, when V is fixed at x, the random vector T becomes

$$T = \mu + \frac{A}{\sqrt{x/k}} Z \sim N_p\left(\mu, \frac{A}{\sqrt{x/k}}\left(\frac{A}{\sqrt{v/k}}\right)^T\right) = N_p\left(\mu, \frac{k}{x}AA^T\right) = N_p\left(\mu, \frac{k}{x}\Sigma\right)$$

so that

$$\begin{aligned} f_{T_1,\ldots,T_p|V=x}(y) &= \frac{1}{(2\pi)^{p/2}\sqrt{\det(\frac{k}{x}\Sigma)}} \exp\left[-\frac{1}{2}(y-\mu)^T\left(\frac{k}{x}\Sigma\right)^{-1}(y-\mu)\right] \\ &= \frac{x^{p/2}}{(2\pi)^{p/2}k^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{x}{2v}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) \end{aligned}$$

where we used $\det(\frac{v}{x}\Sigma) = (v/x)^p \det(\Sigma)$. As a result

$$\begin{aligned} & f_{T_1,\ldots,T_p}(y_1, \ldots, y_p) \\ &= \int_0^\infty f_{T_1,\ldots,T_p|V=x}(y_1, \ldots, y_p) f_V(x) dx \\ &\propto \int_0^\infty \frac{x^{p/2}}{(2\pi)^{p/2}k^{p/2}\sqrt{\det(\Sigma)}} \exp\left(-\frac{x}{2k}(y-\mu)^T\Sigma^{-1}(y-\mu)\right) x^{\frac{k}{2}-1}e^{-x/2} dx \\ &\propto \int_0^\infty x^{\frac{p+k}{2}-1} \exp\left(-\frac{x}{2}\left[1 + \frac{1}{k}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]\right) dx. \end{aligned}$$

The change of variable

$$t = x\left[1 + \frac{1}{k}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]$$

leads to

$$\begin{aligned} f_T(y) &\propto \frac{1}{\left[1+\frac{1}{k}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{k+p}{2}}}\int_0^\infty t^{\frac{k+p}{2}-1}e^{-t/2}dt \\ &\propto \frac{1}{\left[1+\frac{1}{k}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{k+p}{2}}}. \end{aligned}$$

Therefore the density corresponding to $t_{k,p}(\mu,\Sigma)$ distribution is proportional to

$$y \mapsto \frac{1}{\left[1+\frac{1}{k}(y-\mu)^T\Sigma^{-1}(y-\mu)\right]^{\frac{k+p}{2}}}. \tag{83}$$

Note that, in the notation $t_{k,p}(\mu,\Sigma)$, k denotes degrees of freedom, p denotes dimension, μ and $\Sigma = AA^T$ denote the mean vector and covariance matrix of the corresponding normal random vector $\mu + AZ$.

As in the univariate case, when k (degrees of freedom) is large, $t_{k,p}(\mu,\Sigma)$ is very close to $N_p(\mu,\Sigma)$.

As an application involving the multivariate normal and t-densities, we shall look at Bayesian Linear Regression.

17.4 Bayesian Linear Regression

Here one observes data $(x_1,y_1),\ldots,(x_n,y_n)$. x_i denotes the explanatory variable value and y_i denotes the response variable value for the i^{th} individual. In usual linear regression analysis, we assume the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1,\ldots,n$ where

$$\epsilon_1,\ldots,\epsilon_n \overset{\text{i.i.d}}{\sim} N(0,\sigma^2).$$

There are three parameters in this model β_0,β_1 and σ^2. How to fit this model to the observed data $(x_1,y_1),\ldots,(x_n,y_n)$ i.e., how do we estimate the parameters β_0,β_1,σ and also characterize the uncertainty in the estimates.

As an alternative to the usual frequentist analysis, we shall apply probability theory to solve this problem. The first step is to select a prior for the unknown parameters β_0,β_1,σ. A reasonable prior reflecting ignorance is

$$\beta_0,\beta_1,\log\sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C,C)$$

for a large number C (the exact value of C will not matter in the following calculations). Note that as σ is always positive, we have made the uniform assumption on $\log\sigma$ (by the change of variable formula, the density of σ would be given by $f_\sigma(x) = f_{\log\sigma}(\log x)\frac{1}{x} = \frac{I\{-C<\log x<C\}}{2Cx} = \frac{I\{e^{-C}<x<e^C\}}{2Cx}$.

The joint posterior for all the unknown parameters β_0,β_1,σ is then given by (below we write the term "data" for $Y_1 = y_1,\ldots,Y_n = y_n$):

$$f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma) \propto f_{Y_1,\ldots,Y_n|\beta_0,\beta_1,\sigma}(y_1,\ldots,y_n) f_{\beta_0,\beta_1,\sigma}(\beta_0,\beta_1,\sigma).$$

The two terms on the right hand side above are

$$\begin{aligned} f_{Y_1,\ldots,Y_n|\beta_0,\beta_1,\sigma}(y_1,\ldots,y_n) &\propto \prod_{i=1}^{n} f_{Y_i|\beta_0,\beta_1,\sigma}(y_i) \\ &= \prod_{i=1}^{n} f_{\epsilon_i|\beta_0,\beta_1,\sigma}(y_i-\beta_0-\beta_1 x_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i-\beta_0-\beta_1 x_i)^2}{2\sigma^2}\right) \\ &\propto \sigma^{-n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right), \end{aligned}$$

and

$$\begin{aligned} f_{\beta_0,\beta_1,\sigma}(\beta_0,\beta_1,\sigma) &= f_{\beta_0}(\beta_0) f_{\beta_1}(\beta_1) f_{\sigma}(\sigma) \\ &\propto \frac{I\{-C<\beta_0<C\}}{2C}\frac{I\{-C<\beta_1<C\}}{2C}\frac{I\{e^{-C}<\sigma<e^{C}\}}{2C\sigma} \\ &\propto \frac{1}{\sigma} I\left\{-C<\beta_0,\beta_1,\log\sigma<C\right\}. \end{aligned}$$

We thus obtain

$$\begin{aligned} &f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma) \\ &\propto \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right) I\left\{-C<\beta_0,\beta_1,\log\sigma<C\right\}. \end{aligned}$$

The above is the joint posterior over β_0,β_1,σ. The posterior over only the main parameters β_0,β_1 can be obtained by integrating the parameter σ as follows:

$$\begin{aligned} f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) &= \int f_{\beta_0,\beta_1,\sigma|\text{data}}(\beta_0,\beta_1,\sigma)d\sigma \\ &\propto I\{-C<\beta_0,\beta_1<C\}\int_{e^{-C}}^{e^{C}} \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)d\sigma. \end{aligned}$$

When C is large, the above integral can be evaluated from 0 to ∞ which gives

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) \propto I\{-C<\beta_0,\beta_1<C\}\int_{0}^{\infty} \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)d\sigma.$$

The change of variable

$$s=\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2}}$$

allows us to write the integral as

$$\begin{aligned} &\int_0^{\infty}\sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)d\sigma \\ &=\left(\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)^{-n/2}\int_0^{\infty} s^{-n-1}\exp\left(-\frac{1}{2s^2}\right)ds \propto \left(\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2\right)^{-n/2}. \end{aligned}$$

The posterior density of (β_0, β_1) is thus

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) \propto I\{-C < \beta_0, \beta_1 < C\} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right)^{-n/2}. \tag{84}$$

A key role in the above posterior is played by the least squares criterion:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 t_i)^2 .$$

The usual point estimates of β_0 and β_1 are simply the minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ of the least squares criterion $S(\beta_0, \beta_1)$.

We can rewrite the posterior (84) as

$$\begin{aligned} f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) &\propto \left(\frac{1}{S(\beta_0,\beta_1)} \right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \\ &\propto \left(\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{S(\beta_0,\beta_1)} \right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \end{aligned} \tag{85}$$

Note that we have been able to bring in the term $(S(\hat{\beta}_0, \hat{\beta}_1))^{n/2}$ because it does not depend on β_0, β_1 and is thus a constant.

Generally, the density (89) will be quite sharply concentrated around the least squares estimator $(\hat{\beta}_0, \hat{\beta}_1)$ especially when n is large. This is because, when (β_0, β_1) is such that $S(\beta_0, \beta_1)$ is large compared to $S(\hat{\beta}_0, \hat{\beta}_1)$, the quantity

$$\left(\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{S(\beta_0,\beta_1)} \right)^{n/2}$$

would be quite negligible because of the large power $n/2$. As a result, the posterior density $f_{\beta_0,\beta_1|\text{data}}(\beta_0, \beta_1)$ will be concentrated around those values of (β_0, β_1) for which $S(\beta_0, \beta_1)$ is quite close to $S(\hat{\beta}_0, \hat{\beta}_1)$. For a concrete example, suppose $n = 762$ and (β_0, β_1) is such that $S(\beta_0, \beta_1) = (1.1)S(\hat{\beta}_0, \hat{\beta}_1)$. Then

$$\left(\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{S(\beta_0,\beta_1)} \right)^{n/2} = \left(\frac{1}{1.1} \right)^{381} \approx 1.7 \times 10^{-16}.$$

Such (β_0, β_1) will thus get negligible posterior probability. Even for (β_0, β_1) such that $S(\beta_0, \beta_1) = (1.01)S(\hat{\beta}_0, \hat{\beta}_1)$, we have

$$\left(\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{S(\beta_0,\beta_1)} \right)^{n/2} = \left(\frac{1}{1.01} \right)^{381} \approx 0.02$$

and so such (β_0, β_1) will also get fairly small posterior probability.

To sum up, when n is large, the posterior probability will be concentrated around those (β_0, β_1) for which $S(\beta_0, \beta_1)$ is very close to $S(\hat{\beta}_0, \hat{\beta}_1)$. Generally, this would imply that (β_0, β_1) would itself have to be close to $(\hat{\beta}_0, \hat{\beta}_1)$. For this reason, the indicator term in (89) has no effect when C is large. We can thus drop this indicator term and refer to the Bayesian posterior as simply

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0,\beta_1) \propto \left(\frac{S(\hat{\beta}_0,\hat{\beta}_1)}{S(\beta_0,\beta_1)} \right)^{n/2}. \tag{86}$$

We shall show in the next class that the above posterior density is simply the multivariate t-density.

18 Lecture Eighteen

18.1 Recap: Multivariate Normal and t Distributions

In the last class, we looked at the multivariate normal distribution $N_p(\mu, \Sigma)$ (p denotes dimension, μ denotes mean vector and Σ denotes covariance matrix) with density:

$$\left(\frac{1}{\sqrt{2\pi}}\right)^p \frac{1}{\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right).$$

and the multivariate t-distribution $t_{k,p}(\mu, \Sigma)$ (k is the degrees of freedom) with density proportional to

$$\left(\frac{1}{1+\frac{1}{k}(x-\mu)^T\Sigma^{-1}(x-\mu)}\right)^{\frac{k+p}{2}}. \tag{87}$$

One can generate random vectors having these distributions in the following way. First consider a $p \times 1$ random vector Z whose components $Z_1, \ldots, Z_p$ are i.i.d standard normal. Then

$$X = \mu + AZ$$

has the $N_p(\mu, \Sigma)$ distribution with $\Sigma = AA^T$. Also

$$T = \mu + \frac{1}{\sqrt{V/k}} AZ \tag{88}$$

has the $t_{k,p}(\mu, \Sigma)$ distribution. Here V has the χ^2_k distribution and we assume that V and Z are independent.

The following fact will be useful in the sequel.

Fact 18.1. *If $T \sim t_{k,p}(\mu, \Sigma)$ has components $T_1, \ldots, T_p$, then, for each $i = 1, \ldots, p$,*

$$T_i \sim t_k(\mu_i, \Sigma(i,i))$$

where μ_i is the i^{th} component of μ and $\Sigma(i,i)$ is the $(i,i)^{th}$ entry of Σ. In words, each T_i has the univariate t-distribution.

Proof. This fact follows directly from (88) because

$$T_i = \mu_i + \frac{1}{\sqrt{V/k}} \sum_{j=1}^p A(i,j) Z_j$$

Now $\sum_{j=1}^p A(i,j)Z_j$ has the normal distribution with mean zero and variance

$$\sum_{j=1}^p (A(i,j))^2 = \sum_{j=1}^p A(i,j)A^T(j,i) = (AA^T)(i,i) = \Sigma(i,i).$$

Therefore we can write

$$\sum_{j=1}^p A(i,j)Z_j = \sqrt{\Sigma(i,i)}W \qquad \text{where } W \sim N(0,1).$$

Thus

$$T_i = \mu_i + \frac{1}{\sqrt{V/k}} \sqrt{\Sigma(i,i)}W.$$

This has the same form as (88) except instead of AZ, we have the univariate product $\sqrt{\Sigma(i,i)}W$ where $W \sim N(0,1)$. Thus $T_i \sim t_k(\mu_i, \sqrt{\Sigma(i,i)})$. □

18.2 Application to Linear Regression

We considered the usual linear regression model in the last class. One observes data $(x_1, y_1), \ldots, (x_n, y_n)$. x_i denotes the explanatory variable value and y_i denotes the response variable value for the i^{th} individual. We consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, \ldots, n$ where

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

There are three parameters in this model β_0, β_1 and σ^2. The goal is to estimate the parameters β_0, β_1 and also characterize the uncertainty in the estimates.

We worked with the prior distribution

$$\beta_0, \beta_1, \log \sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C)$$

for a large number C and calculated the posterior density of β_0, β_1 (by integrating the full posterior of β_0, β_1, σ over σ) to be

$$f_{\beta_0,\beta_1|\text{data}}(\beta_0, \beta_1) \propto I\{-C < \beta_0, \beta_1 < C\} \left(\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \right)^{-n/2}.$$

We show below that this density is very closely related to the multivariate t-density (87) . To see this, let us use the notation

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 .$$

The usual point estimates of β_0 and β_1 are simply the minimizers $\hat{\beta}_0$ and $\hat{\beta}_1$ of the least squares criterion $S(\beta_0, \beta_1)$.

We can then rewrite the above posterior as

$$\begin{aligned} f_{\beta_0,\beta_1|\text{data}}(\beta_0, \beta_1) &\propto \left(\frac{1}{S(\beta_0, \beta_1)} \right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \\ &\propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1)}{S(\beta_0, \beta_1)} \right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \end{aligned} \tag{89}$$

Note that we have been able to bring in the term $(S(\hat{\beta}_0, \hat{\beta}_1))^{n/2}$ because it does not depend on β_0, β_1 and is thus a constant.

Using the notation

$$Y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & x_1 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & x_n \end{pmatrix} \quad \text{and} \quad \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \quad \text{and} \quad \hat{\beta} := \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix},$$

we can write

$$S(\beta_0, \beta_1) = S(\beta) = \|Y - X\beta\|^2 \quad \text{and} \quad S(\hat{\beta}_0, \hat{\beta}_1) = S(\hat{\beta}) = \|Y - X\hat{\beta}\|^2$$

We now use the following Pythagorean decomposition

$$S(\beta) = \|Y - X\beta\|^2 = \|Y - X\hat{\beta}\|^2 + \|X\beta - X\hat{\beta}\|^2 = S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta}).$$

We can thus write

$$\begin{aligned} f_{\beta_0,\beta_1|\text{data}}(\beta) &\propto \left(\frac{S(\hat{\beta})}{S(\beta)}\right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \\ &= \left(\frac{S(\hat{\beta})}{S(\hat{\beta}) + (\beta - \hat{\beta})^T X^T X(\beta - \hat{\beta})}\right)^{n/2} I\{-C < \beta_0, \beta_1 < C\} \\ &= \left(\frac{1}{1 + (\beta - \hat{\beta})^T \left(\frac{X^T X}{S(\hat{\beta})}\right)(\beta - \hat{\beta})}\right)^{n/2} I\{-C < \beta_0, \beta_1 < C\}. \end{aligned}$$

If we ignore the indicator above, the above density is simply the multivariate t-density with dimension $p = 2$, degrees of freedom $k = n - 2$, mean parameter $\hat{\beta}$ and covariance matrix parameter Σ where

$$\Sigma^{-1} = \frac{n-2}{S(\hat{\beta})}(X^T X) \quad \text{so that } \Sigma = \frac{S(\hat{\beta})}{n-2}(X^T X)^{-1}.$$

Therefore the posterior density is just the $t_{n-2,2}(\hat{\beta}, \Sigma)$ density truncated to the set $-C < \beta_0, \beta_1 < C$. When C is large, this truncation will have little practical effect so we can just treat the posterior density as $t_{n-2,2}(\hat{\beta}, \Sigma)$.

Let us now use some standard regression terminology:

$$\text{Residuals are } y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \text{ for } i = 1, \ldots, n.$$

$$S(\hat{\beta}) = S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2 \text{ is the Residual Sum of Squares (RSS)}$$

$$\hat{\sigma} := \sqrt{\frac{S(\hat{\beta})}{n-2}} \text{ is the Residual Standard Error.}$$

We have thus proved that

$$\beta \mid \text{data} \sim t_{n-2,2}(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}).$$

With this posterior density, one can do uncertainty quantification about the parameters β_0 and β_1. One can generate multiple samples from $t_{n-2,2}(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1})$ and plot the resulting lines to visualize the uncertainty in β_0 and β_1. One can also use Fact 18.1 to deduce that

$$\beta_0 \mid \text{data} \sim t_{n-2}(\hat{\beta}_0, \hat{\sigma}^2 (X^T X)^{11}) \quad \text{and} \quad \beta_1 \mid \text{data} \sim t_{n-2}(\hat{\beta}_1, \hat{\sigma}^2 (X^T X)^{22})$$

where $(X^T X)^{11}$ and $(X^T X)^{22}$ are the first and second diagonal entries of $(X^T X)^{-1}$ respectively. These univariate t-densities describe the marginal uncertainty in the intercept and slope parameters. When n is large, these will be close to the normal distributions $N(\hat{\beta}_0, \hat{\sigma}^2 (X^T X)^{11})$ and $N(\hat{\beta}_1, \hat{\sigma}^2 (X^T X)^{22})$ respectively. The quantities $\hat{\sigma}\sqrt{(X^T X)^{11}}$ and $\hat{\sigma}\sqrt{(X^T X)^{22}}$ are known as the standard errors of the intercept and the slope respectively.

18.3 Multiple Linear Regression

The analysis for multiple linear regression is very similar to analysis of the last section. Here one observes data $(y_i, x_{i1}, x_{i2}, \ldots, x_{im})$ for $i = 1, \ldots, n$. There are m explanatory variables $x_1, \ldots, x_m$ and one response variable. x_{ij} denotes the value of the j^{th} explanatory variable for the i^{th} individual and y_i is the value of the response variable for the i^{th} individual. The model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \epsilon_i$$

for $i = 1, \ldots, n$ where

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2).$$

The goal is to estimate the parameters $\beta_0, \ldots, \beta_m$ as well as σ from the data. σ is usually treated as a nuisance parameter and the main parameters of interest are $\beta_0, \ldots, \beta_m$. The model studied in the previous section is often called simple linear regression and it corresponds to $m = 1$ (i.e., there is only one explanatory variable).

We shall work with the prior distribution:

$$\beta_0, \beta_1, \ldots, \beta_m, \log \sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C).$$

Under this assumption, it can be easily seen that the posterior for $\beta_0, \ldots, \beta_m$ is given (just like in the last section) as

$$f_{\beta_0, \beta_1, \ldots, \beta_m | \text{data}} \propto \left(\frac{S(\hat{\beta}_0, \ldots, \hat{\beta}_m)}{S(\beta_0, \ldots, \beta_m)} \right)^{n/2} I\left\{ -C < \beta_0, \ldots, \beta_m < C \right\}$$

where

$$S(\beta_0, \ldots, \beta_m) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2$$

is the least squares criterion, and $\hat{\beta}_0, \ldots, \hat{\beta}_m$ are the least squares estimators (these are the minimizers of $S(\beta_0, \ldots, \beta_m)$). Using the matrix notation:

$$Y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & x_{11} & \cdot & \cdot & \cdot & x_{1m} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{i1} & \cdot & \cdot & \cdot & x_{im} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n1} & \cdot & \cdot & \cdot & x_{nm} \end{pmatrix} \quad \text{and} \quad \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix} \quad \text{and} \quad \hat{\beta} := \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\beta}_m \end{pmatrix},$$

it can be seen that

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

and

$$S(\beta) = \|Y - X\beta\|^2.$$

We can then show (using the same Pythagorean decomposition as in the last section) that

$$f_{\beta_0, \beta_1, \ldots, \beta_m | \text{data}}(\beta_0, \ldots, \beta_m) \propto \left(\frac{1}{1 + (\beta - \hat{\beta})^T \left(\frac{X^T X}{S(\hat{\beta})} \right) (\beta - \hat{\beta})} \right)^{n/2} I\{ -C < \beta_0, \ldots, \beta_m < C \}.$$

If we ignore the indicator above, the above density is simply the multivariate t-density with dimension $p = m+1$, degrees of freedom $k = n-p$, mean parameter $\hat{\beta}$ and covariance matrix parameter Σ where

$$\Sigma^{-1} = \frac{n-p}{S(\hat{\beta})}(X^T X) \quad \text{so that } \Sigma = \frac{S(\hat{\beta})}{n-p}(X^T X)^{-1}.$$

Therefore the posterior density is just the $t_{n-p,p}(\hat{\beta}, \Sigma)$ density truncated to the set $-C < \beta_0, \beta_1, \ldots, \beta_m < C$. When C is large, this truncation will have little practical effect so we can just treat the posterior density as $t_{n-p,p}(\hat{\beta}, \Sigma)$. We can then use the Fact 18.1 to obtain marginal t-distributions for each individual component $\beta_j, j = 0, 1, \ldots, m$.

Multiple Linear Regression can be used to fit even when there is only one explanatory variable to fit certain nonlinear functions of the explanatory variable. For example, one can fit quadratic functions via the model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i \tag{90}$$

with $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$ by the multiple linear regression methodology with

$$X = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & x_n & x_n^2 \end{pmatrix}.$$

The posterior density of $\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}$ will then be given by $t_{n-3,3}(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1})$ (note that the dimension now is 3 and the degrees of freedom is $n-3$).

A more general polynomial trend model (of degree k) can be fit analogously (the dimension of β will then be $k+1$ and the degrees of freedom of the posterior t-density will be $n-k-1$).

Other examples include the following. To capture seasonal trend in time series data (here the explanatory variable is time with values $t_1, \ldots, t_n$) with known period s (for example $s = 12$ in monthly data), one can use a model of the form

$$Y_i = \beta_0 + \sum_{f=1}^{r} \left(\beta_{f1} \cos \frac{2\pi f t_i}{s} + \beta_{f2} \sin \frac{2\pi f t_i}{s} \right) + \epsilon_i.$$

This is also a linear regression model with

$$\beta = (\beta_0, \beta_{11}, \beta_{12}, \beta_{21}\beta_{22}, \ldots, \beta_{r1}, \beta_{r2})^T$$

and the i^{th} row of the $n \times (2r+1)$ matrix X is given by

$$\left(1, \cos \frac{2\pi t_i}{s}, \sin \frac{2\pi t_i}{s}, \cos \frac{2\pi(2) t_i}{s}, \sin \frac{2\pi(2) t_i}{s}, \ldots, \cos \frac{2\pi(r) t_i}{s}, \sin \frac{2\pi(r) t_i}{s} \right).$$

Here the posterior t-density of β will have dimension $2r+1$ and degrees of freedom $n-(2r+1)$.

Time series datasets often have both trend and seasonality. These effects can be estimated by models of the form:

$$Y_i = \sum_{j=0}^{k} \beta_j^{(1)} t_i^j + \sum_{f=1}^{r} \left(\beta_{f1} \cos \frac{2\pi f t_i}{s} + \beta_{f2} \sin \frac{2\pi f t_i}{s} \right) + \epsilon_i.$$

Inference for this model can also be done through linear regression. The degrees of freedom for the posterior t-density of the coefficients will now be $n-(2r+k+1)$. Our methodology will work as long as $n > 2r+k+1$.

18.4 Models with Nonlinear Parameter Dependence

The Bayesian methodology can be used even to fit models with nonlinear parameter dependence such as:

$$Y_i = \beta_0 + \beta_1 \cos(2\pi f x_i) + \beta_2 \sin(2\pi f x_i) + \epsilon_i \tag{91}$$

with $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0,\sigma^2)$. The setting here is the usual simple linear regression setting (there is only one explanatory variable). The parameters are $\beta_0, \beta_1, \beta_2, \sigma$ as well as the frequency parameter f. One cannot use linear regression methodology directly here because the parameter f appears nonlinearly in the equation (91). We shall see how to handle this in the next class.

19 Lecture Nineteen

19.1 Last Class: Linear Regression

Last class, we used probability to perform inference in the usual linear regression model. Here one observes data $(y_i, x_{i1}, x_{i2}, \ldots, x_{im})$ for $i = 1, \ldots, n$. There are m explanatory variables $x_1, \ldots, x_m$ and one response variable. x_{ij} denotes the value of the j^{th} explanatory variable for the i^{th} individual and y_i is the value of the response variable for the i^{th} individual. The model is

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im} + \epsilon_i$$

for $i = 1, \ldots, n$ where

$$\epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0,\sigma^2).$$

The goal is to estimate the parameters $\beta_0, \ldots, \beta_m$ as well as σ from the data. σ is usually treated as a nuisance parameter and the main parameters of interest are $\beta_0, \ldots, \beta_m$.

We used the prior distribution:

$$\beta_0, \beta_1, \ldots, \beta_m, \log\sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C).$$

Under this assumption, we derived the posterior distribution for $\beta_0, \ldots, \beta_m$ to be

$$f_{\beta_0,\beta_1,\ldots,\beta_m|\text{data}} \propto \left(\frac{S(\hat{\beta}_0, \ldots, \hat{\beta}_m)}{S(\beta_0, \ldots, \beta_m)}\right)^{n/2} I\{-C < \beta_0, \ldots, \beta_m < C\}$$

where

$$S(\beta_0, \ldots, \beta_m) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_m x_{im})^2$$

is the least squares criterion, and $\hat{\beta}_0, \ldots, \hat{\beta}_m$ are the least squares estimators (these are the minimizers of $S(\beta_0, \ldots, \beta_m)$).

We also saw that the posterior density can also be written as

$$f_{\beta_0,\beta_1,\ldots,\beta_m|\text{data}}(\beta_0, \ldots, \beta_m) \propto \left(\frac{1}{1 + (\beta - \hat{\beta})^T \left(\frac{X^T X}{S(\hat{\beta})}\right)(\beta - \hat{\beta})}\right)^{n/2} I\{-C < \beta_0, \ldots, \beta_m < C\}.$$

using the notation:

$$Y = \begin{pmatrix} y_1 \\ \cdot \\ \cdot \\ \cdot \\ y_n \end{pmatrix} \quad \text{and} \quad X = \begin{pmatrix} 1 & x_{11} & \cdot & \cdot & \cdot & x_{1m} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{i1} & \cdot & \cdot & \cdot & x_{im} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_{n1} & \cdot & \cdot & \cdot & x_{nm} \end{pmatrix} \quad \text{and} \quad \beta := \begin{pmatrix} \beta_0 \\ \beta_1 \\ \cdot \\ \cdot \\ \cdot \\ \beta_m \end{pmatrix} \quad \text{and} \quad \hat{\beta} := \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \cdot \\ \cdot \\ \cdot \\ \hat{\beta}_m \end{pmatrix}.$$

If we ignore the indicator function, the above density is simply the multivariate t-density with dimension $p = m + 1$, degrees of freedom $k = n - p$, mean parameter $\hat{\beta}$ and covariance matrix parameter Σ where

$$\Sigma^{-1} = \frac{n-p}{S(\hat{\beta})}(X^T X) \quad \text{so that } \Sigma = \frac{S(\hat{\beta})}{n-p}(X^T X)^{-1}.$$

Therefore the posterior density is just the $t_{n-p,p}(\hat{\beta}, \Sigma)$ density truncated to the set $-C < \beta_0, \beta_1, \ldots, \beta_m < C$. When C is large, this truncation will have little practical effect so we can just treat the posterior density as $t_{n-p,p}(\hat{\beta}, \Sigma)$. Point estimates for β will just be the least squares estimator $\hat{\beta}$, and uncertainty is usually summarized by the standard errors which are simply the square roots of the diagonal entries of Σ. In other words, the standard error corresponding to $\hat{\beta}_j$ equals $\sqrt{\frac{S(\hat{\beta})}{n-p}}$ multiplied by the square-root of the corresponding diagonal entry of $(X^T X)^{-1}$.

19.2 Nonlinear Regression Models

In this framework, parameter inference in nonlinear regression models is handled in a very similar way. For a concrete example, consider the model:

$$Y_i = \beta_0 + \beta_1 \exp(-\beta_2 x_i) + \epsilon_i$$

for $i = 1, \ldots, n$ where, as in the previous section, $\epsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$. This is a nonlinear regression model because the parameter β_2 enters via the exponential function which is nonlinear. There are four unknown parameters $\beta_0, \beta_1, \beta_2, \sigma$. We can obtain parameter estimates and standard errors for them in a manner that is very similar to the analysis in linear regression. We work with the prior:

$$\beta_0, \beta_1, \beta_2, \log \sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C)$$

for a large $C > 0$. The posterior density of the parameters is then given by: The joint posterior for all the unknown parameters $\beta_0, \beta_1, \beta_2, \sigma$ is then given by (below we write the term "data" for $Y_1 = y_1, \ldots, Y_n = y_n$):

$$f_{\beta_0,\beta_1,\beta_2,\sigma|\text{data}}(\beta_0, \beta_1, \beta_2, \sigma) \propto f_{Y_1,\ldots,Y_n|\beta_0,\beta_1,\beta_2,\sigma}(y_1, \ldots, y_n) f_{\beta_0,\beta_1,\beta_2,\sigma}(\beta_0, \beta_1, \beta_2, \sigma).$$

The two terms on the right hand side above are the likelihood:

$$\begin{aligned} f_{Y_1,\ldots,Y_n|\beta_0,\beta_1,\beta_2,\sigma}(y_1,\ldots,y_n) &\propto \prod_{i=1}^{n} f_{Y_i|\beta_0,\beta_1,\beta_2,\sigma}(y_i) \\ &= \prod_{i=1}^{n} f_{\epsilon_i|\beta_0,\beta_1,\beta_2,\sigma}\left(y_i-\beta_0-\beta_1\exp(-\beta_2 x_i)\right) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(y_i-\beta_0-\beta_1\exp(-\beta_2 x_i))^2}{2\sigma^2}\right) \\ &\propto \sigma^{-n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1\exp(-\beta_2 x_i))^2\right), \end{aligned}$$

and

$$\begin{aligned} f_{\beta_0,\beta_1,\beta_2,\sigma}(\beta_0,\beta_1,\beta_2,\sigma) &= f_{\beta_0}(\beta_0)f_{\beta_1}(\beta_1)f_{\beta_2}(\beta_2)f_{\sigma}(\sigma) \\ &\propto \frac{I\{-C<\beta_0<C\}}{2C}\frac{I\{-C<\beta_1<C\}}{2C}\frac{I\{-C<\beta_2<C\}}{2C}\frac{I\{e^{-C}<\sigma<e^{C}\}}{2C\sigma} \\ &\propto \frac{1}{\sigma}I\left\{-C<\beta_0,\beta_1,\beta_2,\log\sigma<C\right\}. \end{aligned}$$

We thus obtain

$$\begin{aligned} &f_{\beta_0,\beta_1,\beta_2,\sigma|\text{data}}(\beta_0,\beta_1,\beta_2,\sigma) \\ &\propto \sigma^{-n-1}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1\exp(-\beta_2 x_i))^2\right) I\left\{-C<\beta_0,\beta_1,\beta_2,\log\sigma<C\right\}. \end{aligned}$$

Using the notation

$$S(\beta_0,\beta_1,\beta_2) := \sum_{i=1}^{n}(y_i-\beta_0-\beta_1\exp(-\beta_2 x_i))^2$$

for the sum of squares criterion, we can write the posterior as We thus obtain

$$f_{\beta_0,\beta_1,\beta_2,\sigma|\text{data}}(\beta_0,\beta_1,\beta_2,\sigma) \propto \sigma^{-n-1}\exp\left(-\frac{S(\beta_0,\beta_1,\beta_2)}{2\sigma^2}\right) I\left\{-C<\beta_0,\beta_1,\beta_2,\log\sigma<C\right\}.$$

Often our interest is only in the parameters β_0,β_1,β_2 (σ is a nuisance parameter). To obtain the posterior of β_0,β_1,β_2, we integrate the full posterior above with respect to σ. Assuming that C is large, we can do the integral from 0 to ∞ and this leads to (the calculation is the same as in the linear regression case)

$$f_{\beta_0,\beta_1,\beta_2|\text{data}}(\beta_0,\beta_1,\beta_2) \propto \left(\frac{1}{S(\beta_0,\beta_1,\beta_2)}\right)^{n/2} I\left\{-C<\beta_0,\beta_1,\beta_2<C\right\}.$$

This posterior density will take its largest value when $(\beta_0,\beta_1,\beta_2)$ minimizer the sum of squares $S(\beta_0,\beta_1,\beta_2)$. In other words, the maximu posterior density will be achieved by the least squares estimator:

$$(\hat{\beta}_0,\hat{\beta}_1,\hat{\beta}_2) = \text{minimizer of } S(\beta_0,\beta_1,\beta_2).$$

Unlike in the linear regression case where we can write the least squares estimator in closed form as $(X^TX)^{-1}X^TY$, we may not be able to write the least squares estimator in this nonlinear regression model in closed form. Nevertheless, generally there exists a unique least

squares estimator. The posterior distribution will assign nonnegligible probability only to those parameter values $\beta_0, \beta_1, \beta_2$ for which $S(\beta_0, \beta_1, \beta_2)$ is close to the smallest possible value $S(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$. This can be seen, for example, by rewriting the posterior density as

$$f_{\beta_0,\beta_1,\beta_2|\text{data}}(\beta_0, \beta_1, \beta_2) \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)}{S(\beta_0, \beta_1, \beta_2)}\right)^{n/2} I\left\{-C < \beta_0, \beta_1, \beta_2 < C\right\}.$$

For this reason, we can neglect the indicator function above (because the action will be very close to the least squares estimator $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$) and write

$$f_{\beta_0,\beta_1,\beta_2|\text{data}}(\beta_0, \beta_1, \beta_2) \propto \left(\frac{S(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)}{S(\beta_0, \beta_1, \beta_2)}\right)^{n/2}.$$

Unlike in the linear regression case, the right hand side above is not the (unnormalized) density of a multivariate t-distribution. However, we can approximate it by a multivariate t-distribution by a second Taylor expansion of $S(\beta_0, \beta_1, \beta_2)$ around the least squares estimator $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$. This Taylor expansion is justified because the posterior density will usually be quite concentrated around $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$. Writing β for the vector $(\beta_0, \beta_1, \beta_2)$ and $\hat{\beta}$ for the vector $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$, Taylor expansion is

$$S(\beta) \approx S(\hat{\beta}) + \left\langle \nabla S(\hat{\beta}), \beta - \hat{\beta} \right\rangle + \frac{1}{2}\left(\beta - \hat{\beta}\right)^T HS(\hat{\beta})\left(\beta - \hat{\beta}\right)$$

where $\nabla S(\hat{\beta})$ and $HS(\hat{\beta})$ are the gradient and Hessian of S at $\hat{\beta}$. Because $\hat{\beta}$ minimizes $S(\beta)$, we have $\nabla S(\hat{\beta}) = 0$ and so

$$S(\beta) \approx S(\hat{\beta}) + \frac{1}{2}\left(\beta - \hat{\beta}\right)^T HS(\hat{\beta})\left(\beta - \hat{\beta}\right)$$

Thus the posterior density is approximated by

$$f_{\beta_0,\beta_1,\beta_2|\text{data}}(\beta_0, \beta_1, \beta_2) \propto \left(\frac{S(\hat{\beta})}{S(\hat{\beta}) + \frac{1}{2}\left(\beta - \hat{\beta}\right)^T HS(\hat{\beta})\left(\beta - \hat{\beta}\right)}\right)^{n/2}.$$

Writing $p = 3$ for the dimension of the vector β, we get

$$\begin{aligned} f_{\beta|\text{data}}(\beta) &\propto \left(\frac{1}{1 + \frac{1}{2S(\hat{\beta})}\left(\beta - \hat{\beta}\right)^T HS(\hat{\beta})\left(\beta - \hat{\beta}\right)}\right)^{n/2} \\ &= \left(\frac{1}{1 + \frac{1}{n-p}\frac{n-p}{S(\hat{\beta})}\left(\beta - \hat{\beta}\right)^T \left(\frac{HS(\hat{\beta})}{2}\right)\left(\beta - \hat{\beta}\right)}\right)^{\frac{p+(n-p)}{2}} \end{aligned}$$

This is clearly a multivariate t-density. More specifically,

$$\beta \mid \text{data} \sim t_{n-p,p}\left(\hat{\beta}, \frac{S(\hat{\beta})}{n-p}\left(\frac{1}{2}HS(\hat{\beta})\right)^{-1}\right). \tag{92}$$

One can summarize this posterior distribution by simply reporting the point estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ and their standard errors which are the square roots of the diagonal entries of

$$\frac{S(\hat{\beta})}{n-p}\left(\frac{1}{2}HS(\hat{\beta})\right)^{-1}.$$

The earlier linear regression analysis is a special case of this analysis because we recover the earlier result by taking $S(\beta) = \|Y - X\beta\|^2$ (in this case $HS(\beta) = 2X^T X\beta$).

It should be noted that the result (92) is an approximation (in other words, the exact posterior is not t-distributed) obtained by the second order Taylor expansion of $S(\beta)$ around $\hat{\beta}$. An exact analysis of the posterior can be done in the following way. In this specific problem, there are three β parameters; $\beta_0, \beta_1, \beta_2$. The model is linear in β_0, β_1 for every fixed β_2. This means that the conditional posterior of β_0, β_1 for fixed β_2 is exactly t-distributed. So the marginal posterior density of β_2 can be calculated exactly. We shall do this analysis in more generality in the next section.

19.3 More on Nonlinear Regression Models

Consider the nonliear regression model written in vector-matrix notation as:

$$Y = X(\omega)\beta + \epsilon \tag{93}$$

where Y is $n \times 1$, ω is a $k \times 1$ vector of unknown parameters, $X(\omega)$ is an $n \times p$ matrix that depends in a known way on the unknown parameters in ω, β is a $p \times 1$ vector of unknown parameters and ϵ is a $n \times 1$ vector consisting of i.i.d $N(0, \sigma^2)$ errors. This model has $k+p+1$ parameters: k elements of ω, p elements of β and σ. The model depends linearly on the β parameters but possibly nonlinearly on the ω parameters. This setting includes the following special cases.

1. In the concrete example considered in the previous section, we can take $\omega = (\beta_2)$ and $\beta = (\beta_0, \beta_1)$. The matrix $X(\omega)$ is given by

$$X(\omega) = X(\beta_2) = \begin{pmatrix} 1 & \exp(-\beta_2 x_1) \\ 1 & \exp(-\beta_2 x_2) \\ . & . \\ . & . \\ . & . \\ 1 & \exp(-\beta_2 x_n) \end{pmatrix}$$

2. Consider the model

$$Y_i = \beta_0 + \beta_1 \cos(2\pi f x_i) + \beta_2 \sin(2\pi f x_i) + \epsilon_i.$$

   This is a nonlinear regression model where we are modeling the response as a sinusoidal function of the explanatory variable where the sinusoid has an unknown frequency f. This is a special case of (95) with $\omega = f$, β is the vector with components $\beta_0, \beta_1, \beta_2$ and the $X(\omega)$ matrix is

$$X(\omega) = X(f) = \begin{pmatrix} 1 & \cos(2\pi f x_1) & \sin(2\pi f x_1) \\ 1 & \cos(2\pi f x_2) & \sin(2\pi f x_2) \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & \cos(2\pi f x_n) & \sin(2\pi f x_n) \end{pmatrix}$$

3. Consider the model

$$Y_i = \beta_0 + \beta_1 I\{x_i > \omega\} + \epsilon_i$$

This is a changepoint in mean model where the response variable has mean β_0 when x is atmost ω and mean $\beta_0 + \beta_1$ when x exceeds ω. This is also a special case of (95) with

$$X(\omega) = \begin{pmatrix} 1 & I\{x_1 > \omega\} \\ 1 & I\{x_2 > \omega\} \\ . & . \\ . & . \\ . & . \\ 1 & I\{x_n > \omega\} \end{pmatrix}$$

4. Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \omega)_+ + \epsilon_i$$

This is a broken stick regression model where the regression line has slope β_1 when the covariate is at most ω and has slope $\beta_1 + \beta_2$ when the covariate exceeds ω. Here $(x - \omega)_+ = \max(x - \omega, 0)$. This is also a special case of (95) with

$$X(\omega) = \begin{pmatrix} 1 & x_1 & (x_1 - \omega)_+ \\ 1 & x_2 & (x_2 - \omega)_+ \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_n & (x_n - \omega)_+ \end{pmatrix}$$

The likelihood of the model (95) is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right)$$

To perform Bayesian analysis in the model (95), we assume as before that all the components of ω, all the components of β and $\log \sigma$ are all i.i.d uniformly distributed on $(-C, C)$ for a large C. The posterior density is then given by

$$f_{\omega,\beta,\sigma|\text{data}}(\omega, \beta, \sigma) \propto \sigma^{-n-1} \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right)$$

where we have ignored the indicator functions (assuming C is large). Often the main interest is in the ω parameters. So we integrate the posterior with respect to β and σ. This gives

$$f_{\omega|\text{data}}(\omega) \propto \int_0^\infty \sigma^{-n-1} \int_{\mathbb{R}^p} \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right) d\beta d\sigma \tag{94}$$

Now if $\hat{\beta}(\omega)$ is the least squares estimator for fixed ω:

$$\hat{\beta}(\omega) := \underset{\beta}{\text{argmin}} \ \|Y - X(\omega)\beta\|^2,$$

then using

$$\begin{aligned} \|Y - X(\omega)\beta\|^2 &= \|Y - X(\omega)\hat{\beta}(\omega)\|^2 + \|X(\omega)\beta - X(\omega)\hat{\beta}(\omega)\|^2 \\ &= \|Y - X(\omega)\hat{\beta}(\omega)\|^2 + \left(\beta - \hat{\beta}(\omega)\right)^T X(\omega)^T X(\omega) \left(\beta - \hat{\beta}(\omega)\right), \end{aligned}$$

the integral (96) then becomes

$$\int_0^\infty \int_{\mathbb{R}^p} \sigma^{-n-1} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) \exp\left(-\frac{(\beta - \hat{\beta}(\omega))X(\omega)'X(\omega)(\beta - \hat{\beta}(\omega))}{2\sigma^2}\right) d\beta d\sigma$$
$$= \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) \int_{\mathbb{R}^p} \exp\left(-\frac{(\beta - \hat{\beta}(\omega))X(\omega)'X(\omega)(\beta - \hat{\beta}(\omega))}{2\sigma^2}\right) d\beta d\sigma.$$

We shall now use the formula:

$$\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx_1 \dots dx_p = (2\pi)^{p/2}\sqrt{\det(\Sigma)}$$

where Σ is a $p \times p$ positive definite matrix and the integral is over $x = (x_1, \dots, x_p)$. This is basically the formula for the normalizing constant for the multivariate normal distribution.

This formula with $\Sigma^{-1} = X(\omega)'X(\omega)/(\sigma^2)$ (or equivalently $\Sigma = \sigma^2(X(\omega)'X(\omega))^{-1}$) gives

$$\int_{\mathbb{R}^p} \exp\left(-\frac{(\beta - \hat{\beta}(\omega))^T X(\omega)^T X(\omega)(\beta - \hat{\beta}(\omega))}{2\sigma^2}\right) d\beta$$
$$= (2\pi)^{p/2}\sqrt{\det\left(\sigma^2(X(\omega)'X(\omega))^{-1}\right)} = (2\pi)^{p/2}\sigma^p \left(\det(X(\omega)'X(\omega))\right)^{-1/2}.$$

The required integral is

$$(2\pi)^{p/2}(\det(X(\omega)'X(\omega)))^{-1/2} \int_0^\infty \sigma^{-n+p-1} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) d\sigma.$$

The change of variable

$$t = \frac{\sigma}{\|Y - X(\omega)\hat{\beta}(\omega)\|}$$

then gives

$$(2\pi)^{p/2}(\det(X'X))^{-1/2} \int_0^\infty \sigma^{-n+p-1} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) d\sigma$$
$$= (2\pi)^{p/2}(\det(X(\omega)'X(\omega)))^{-1/2}\|Y - X(\omega)\hat{\beta}(\omega)\|^{-n+p} \int_0^\infty t^{-n+p-1} \exp\left(-\frac{1}{2t^2}\right) dt$$
$$\propto (\det(X(\omega)'X(\omega)))^{-1/2}\|Y - X(\omega)\hat{\beta}(\omega)\|^{-n+p}.$$

Putting everything together, we have proved that

$$f_{\omega|\text{data}}(\omega) \propto (\det(X(\omega)'X(\omega)))^{-1/2}\|Y - X(\omega)\hat{\beta}(\omega)\|^{-n+p}.$$

This function of ω can be numerically understood when the dimension of ω is low. For example, if ω is one-dimensional, we can plot the posterior density on the computer and normalize it so the density integrates to one.

20 Lecture Twenty

20.1 Last Class: Nonlinear Regression Models with both linear and nonlinear parameter dependence

Consider the nonliear regression model written in vector-matrix notation as:

$$Y = X(\omega)\beta + \epsilon \tag{95}$$

where Y is $n \times 1$, ω is a $k \times 1$ vector of unknown parameters, $X(\omega)$ is an $n \times p$ matrix that depends in a known way on the unknown parameters in ω, β is a $p \times 1$ vector of unknown parameters and ϵ is a $n \times 1$ vector consisting of i.i.d $N(0, \sigma^2)$ errors. This model has $k+p+1$ parameters: k elements of ω, p elements of β and σ. The model depends linearly on the β parameters but possibly nonlinearly on the ω parameters. This setting includes the following special cases.

1. In the concrete example considered in the previous section, we can take $\omega = (\beta_2)$ and $\beta = (\beta_0, \beta_1)$. The matrix $X(\omega)$ is given by

$$X(\omega) = X(\beta_2) = \begin{pmatrix} 1 & \exp(-\beta_2 x_1) \\ 1 & \exp(-\beta_2 x_2) \\ . & . \\ . & . \\ . & . \\ 1 & \exp(-\beta_2 x_n) \end{pmatrix}$$

2. Consider the model

$$Y_i = \beta_0 + \beta_1 \cos(2\pi f x_i) + \beta_2 \sin(2\pi f x_i) + \epsilon_i.$$

   This is a nonlinear regression model where we are modeling the response as a sinusoidal function of the explanatory variable where the sinusoid has an unknown frequency f. This is a special case of (95) with $\omega = f$, β is the vector with components $\beta_0, \beta_1, \beta_2$ and the $X(\omega)$ matrix is

$$X(\omega) = X(f) = \begin{pmatrix} 1 & \cos(2\pi f x_1) & \sin(2\pi f x_1) \\ 1 & \cos(2\pi f x_2) & \sin(2\pi f x_2) \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & \cos(2\pi f x_n) & \sin(2\pi f x_n) \end{pmatrix}$$

3. Consider the model

$$Y_i = \beta_0 + \beta_1 I\{x_i > \omega\} + \epsilon_i$$

   This is a changepoint in mean model where the response variable has mean β_0 when x is atmost ω and mean $\beta_0 + \beta_1$ when x exceeds ω. This is also a special case of (95) with

$$X(\omega) = \begin{pmatrix} 1 & I\{x_1 > \omega\} \\ 1 & I\{x_2 > \omega\} \\ . & . \\ . & . \\ . & . \\ 1 & I\{x_n > \omega\} \end{pmatrix}$$

4. Consider the model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - \omega)_+ + \epsilon_i$$

   This is a broken stick regression model where the regression line has slope β_1 when the covariate is at most ω and has slope $\beta_1 + \beta_2$ when the covariate exceeds ω. Here

$(x-\omega)_+ = \max(x-\omega, 0)$. This is also a special case of (95) with

$$X(\omega) = \begin{pmatrix} 1 & x_1 & (x_1-\omega)_+ \\ 1 & x_2 & (x_2-\omega)_+ \\ . & . & . \\ . & . & . \\ . & . & . \\ 1 & x_n & (x_n-\omega)_+ \end{pmatrix}$$

The likelihood of the model (95) is

$$\left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right)$$

To perform Bayesian analysis in the model (95), we assume as before that all the components of ω, all the components of β and $\log\sigma$ are all i.i.d uniformly distributed on $(-C, C)$ for a large C. The posterior density is then given by

$$f_{\omega,\beta,\sigma|\text{data}}(\omega,\beta,\sigma) \propto \sigma^{-n-1} \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right)$$

where we have ignored the indicator functions (assuming C is large). Often the main interest is in the ω parameters. So we integrate the posterior with respect to β and σ. This gives

$$f_{\omega|\text{data}}(\omega) \propto \int_0^\infty \sigma^{-n-1} \int_{\mathbb{R}^p} \exp\left(-\frac{\|Y - X(\omega)\beta\|^2}{2\sigma^2}\right) d\beta d\sigma \tag{96}$$

Now if $\hat{\beta}(\omega)$ is the least squares estimator for fixed ω:

$$\hat{\beta}(\omega) := \underset{\beta}{\text{argmin}}\ \|Y - X(\omega)\beta\|^2,$$

then using

$$\begin{aligned} \|Y - X(\omega)\beta\|^2 &= \|Y - X(\omega)\hat{\beta}(\omega)\|^2 + \|X(\omega)\beta - X(\omega)\hat{\beta}(\omega)\|^2 \\ &= \|Y - X(\omega)\hat{\beta}(\omega)\|^2 + \left(\beta - \hat{\beta}(\omega)\right)^T X(\omega)^T X(\omega) \left(\beta - \hat{\beta}(\omega)\right), \end{aligned}$$

the integral (96) then becomes

$$\int_0^\infty \int_{\mathbb{R}^p} \sigma^{-n-1} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) \exp\left(-\frac{(\beta - \hat{\beta}(\omega))X(\omega)'X(\omega)(\beta - \hat{\beta}(\omega))}{2\sigma^2}\right) d\beta d\sigma$$
$$= \int_0^\infty \sigma^{-n-1} \exp\left(-\frac{\|Y - X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right) \int_{\mathbb{R}^p} \exp\left(-\frac{(\beta - \hat{\beta}(\omega))X(\omega)'X(\omega)(\beta - \hat{\beta}(\omega))}{2\sigma^2}\right) d\beta d\sigma.$$

We shall now use the formula:

$$\int_{\mathbb{R}^p} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right) dx_1 \ldots dx_p = (2\pi)^{p/2}\sqrt{\det(\Sigma)}$$

where Σ is a $p \times p$ positive definite matrix and the integral is over $x = (x_1, \ldots, x_p)$. This is basically the formula for the normalizing constant for the multivariate normal distribution.

This formula with $\Sigma^{-1} = X(\omega)'X(\omega)/(\sigma^2)$ (or equivalently $\Sigma = \sigma^2(X(\omega)'X(\omega))^{-1}$) gives

$$\int_{\mathbb{R}^p} \exp\left(-\frac{(\beta - \hat{\beta}(\omega))^T X(\omega)^T X(\omega)(\beta - \hat{\beta}(\omega))}{2\sigma^2}\right) d\beta$$
$$= (2\pi)^{p/2}\sqrt{\det\left(\sigma^2(X(\omega)'X(\omega))^{-1}\right)} = (2\pi)^{p/2}\sigma^p \left(\det(X(\omega)'X(\omega))\right)^{-1/2}.$$

The required integral is

$$(2\pi)^{p/2}(\det(X(\omega)'X(\omega)))^{-1/2}\int_0^\infty \sigma^{-n+p-1}\exp\left(-\frac{\|Y-X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right)d\sigma.$$

The change of variable

$$t=\frac{\sigma}{\|Y-X(\omega)\hat{\beta}(\omega)\|}$$

then gives

$$\begin{aligned}
&(2\pi)^{p/2}(\det(X'X))^{-1/2}\int_0^\infty \sigma^{-n+p-1}\exp\left(-\frac{\|Y-X(\omega)\hat{\beta}(\omega)\|^2}{2\sigma^2}\right)d\sigma \\
&= (2\pi)^{p/2}(\det(X(\omega)'X(\omega)))^{-1/2}\|Y-X(\omega)\hat{\beta}(\omega)\|^{-n+p}\int_0^\infty t^{-n+p-1}\exp\left(-\frac{1}{2t^2}\right)dt \\
&\propto (\det(X(\omega)'X(\omega)))^{-1/2}\|Y-X(\omega)\hat{\beta}(\omega)\|^{-n+p}.
\end{aligned}$$

Putting everything together, we have proved that

$$f_{\omega|\text{data}}(\omega)\propto(\det(X(\omega)'X(\omega)))^{-1/2}\|Y-X(\omega)\hat{\beta}(\omega)\|^{-n+p}.$$

This function of ω can be numerically analyzed when the dimension of ω is low. For example, if ω is one-dimensional, we can plot the posterior density on the computer and normalize it so the density integrates to one.

20.2 Logistic Regression

Here is another regression model which can be handled in a straightforward fashion by probability theory. We are again in the usual regression setting where we observe data $(y_i, x_{i1}, x_{i2}, \ldots, x_{im})$ for $i = 1, \ldots, n$. There are m explanatory variables $x_1, \ldots, x_m$ and one response variable. x_{ij} denotes the value of the j^{th} explanatory variable for the i^{th} individual and y_i is the value of the response variable for the i^{th} individual. Suppose now that the response variable is binary i.e., $y_1, \ldots, y_n$ take values in $\{0, 1\}$. In this case, the logistic regression model assumes that:

$$Y_i \overset{\text{independent}}{\sim} \text{Bernoulli}\left(\frac{\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im})}{1+\exp(\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im})}\right) \qquad \text{for } i=1,\ldots,n.$$

Letting $x_i = (1, x_{i1}, \ldots, x_{im})^T$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_m)^T$, we can write the model also as

$$Y_i \overset{\text{independent}}{\sim} \text{Bernoulli}\left(\frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)}\right) \qquad \text{for } i=1,\ldots,n.$$

Observe that $x_1^T, \ldots, x_n^T$ form the rows of the $n \times p$ design matrix X (where $p = m+1$). The unknown parameters in the logistic regression model are $\beta_0, \ldots, \beta_m$ (note that, in contrast to the linear regression model, there is no σ parameter in logistic regression). In order to use probability theory for inference on $\beta_0, \ldots, \beta_m$, we assume the prior:

$$\beta_1, \ldots, \beta_p \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C) \tag{97}$$

for a large C. The posterior of β is then

$$\begin{aligned} f_{\beta|Y=y}(\beta) &\propto f_{Y|\beta}(y) f_\beta(\beta) \\ &\propto \left[\prod_{i=1}^{n} \left(\frac{\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)}\right)^{y_i} \left(1-\frac{\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)}\right)^{1-y_i}\right] I\{\beta_1,\ldots,\beta_p \in (-C,C)\} \\ &= \left[\prod_{i=1}^{n} \frac{\exp(y_i x_i^T\beta)}{1+\exp(x_i^T\beta)}\right] I\{\beta_1,\ldots,\beta_p \in (-C,C)\} \\ &= \left[\exp\left(\ell(\beta)\right)\right] I\{\beta_1,\ldots,\beta_p \in (-C,C)\} \end{aligned}$$

where

$$\ell(\beta) := \sum_{i=1}^{n} \left[y_i(x_i^T\beta) - \log\left(1+\exp(x_i^T\beta)\right)\right].$$

Note that $\ell(\beta)$ is simply the log-likelihood in this problem. The posterior density is not in standard form. If $p = 1$ or $p = 2$, then this can be plotted. One can use various MCMC techniques to obtain samples from this posterior. A closed form multivariate normal approximation that works quite well in practice will be described in the next class. Bayesian inference from this normal approximation to the posterior coincides with usual frequentist inference for logistic regression.

21 Lecture Twenty One

21.1 Logistic Regression

Here is another regression model which can be handled in a straightforward fashion by probability theory. We are again in the usual regression setting where we observe data $(y_i, x_{i1}, x_{i2}, \ldots, x_{im})$ for $i = 1, \ldots, n$. There are m explanatory variables $x_1, \ldots, x_m$ and one response variable. x_{ij} denotes the value of the j^{th} explanatory variable for the i^{th} individual and y_i is the value of the response variable for the i^{th} individual. Suppose now that the response variable is binary i.e., $y_1, \ldots, y_n$ take values in $\{0, 1\}$. In this case, the logistic regression model assumes that:

$$Y_i \overset{\text{independent}}{\sim} \text{Bernoulli}\left(\frac{\exp\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}\right)}{1+\exp\left(\beta_0+\beta_1 x_{i1}+\cdots+\beta_m x_{im}\right)}\right) \qquad \text{for } i = 1, \ldots, n.$$

Letting $x_i = (1, x_{i1}, \ldots, x_{im})^T$ and $\beta = (\beta_0, \beta_1, \ldots, \beta_m)^T$, we can write the model also as

$$Y_i \overset{\text{independent}}{\sim} \text{Bernoulli}\left(\frac{\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)}\right) \qquad \text{for } i = 1, \ldots, n$$

where β is the $(m+1) \times 1$ vector with components $\beta_0, \beta_1, \ldots, \beta_m$.

Suppose $x_1^T, \ldots, x_n^T$ form the rows of the $n \times p$ design matrix X (where $p = m + 1$). The unknown parameters in the logistic regression model are $\beta_0, \ldots, \beta_m$ (note that, in contrast to the linear regression model, there is no σ parameter in logistic regression). In order to use probability theory for inference on $\beta_0, \ldots, \beta_m$, we assume the prior:

$$\beta_0, \beta_1, \ldots, \beta_m \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C) \tag{98}$$

for a large C. The posterior of β is then

$$\begin{aligned}
f_{\beta|Y=y}(\beta) &\propto f_{Y|\beta}(y) f_\beta(\beta) \\
&\propto \left[\prod_{i=1}^n \left(\frac{\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)}\right)^{y_i} \left(1-\frac{\exp\left(x_i^T\beta\right)}{1+\exp\left(x_i^T\beta\right)}\right)^{1-y_i}\right] I\{\beta_0,\beta_1,\ldots,\beta_p \in (-C,C)\} \\
&= \left[\prod_{i=1}^n \frac{\exp(y_i x_i^T\beta)}{1+\exp(x_i^T\beta)}\right] I\{\beta_0,\beta_1,\ldots,\beta_p \in (-C,C)\} \\
&= \left[\exp\left(\ell(\beta)\right)\right] I\{\beta_0,\beta_1,\ldots,\beta_p \in (-C,C)\}
\end{aligned}$$

where

$$\ell(\beta) := \sum_{i=1}^n \left[y_i(x_i^T\beta) - \log\left(1+\exp(x_i^T\beta)\right)\right].$$

Note that $\ell(\beta)$ is simply the log-likelihood in this problem. The posterior density is not in standard form. If $p = 1$ or $p = 2$, then this can be plotted. One can use various MCMC techniques to obtain samples from this posterior. A closed form multivariate normal approximation that works quite well in practice can be found as follows. Bayesian inference from this normal approximation to the posterior coincides with usual frequentist inference for logistic regression. To get the normal approximation, first let us drop the indicator which will be irrelevant when C is large to get

$$f_{\beta|Y=y}(\beta) \propto \exp(\ell(\beta)).$$

The normal approximation will be obtained by a second-order Taylor expansion of $\ell(\beta)$. We shall do the Taylor expansion around the MLE $\hat{\beta}$ because the posterior is peaked at $\hat{\beta}$ and the high regions of the posterior are most likely very close to $\hat{\beta}$. Recall that the MLE $\hat{\beta}$ is defined as the maximizer of the likelihood (or log-likelihood):

$$\hat{\beta} := \operatorname*{argmax}_{\beta\in\mathbb{R}^p} \ell(\beta).$$

It is obtained by taking the gradient of the log-likelihood and solving the equation obtained by setting the gradient to zero. It is easy to check that the gradient of the log-likelihood is

$$\nabla\ell(\beta) = \sum_{i=1}^n \left(y_i - \frac{\exp\left(x_i'\beta\right)}{1+\exp\left(x_i'\beta\right)}\right) x_i. \tag{99}$$

To get the maximum likelihood estimator $\hat{\theta}$, we need to set the gradient above to zero and solve the resulting equation for θ. This cannot be done in closed form and the usual method is to use an iterative scheme such as Newton's algorithm. The answer can be obtained from inbuilt functions in R or Python. More details behind the Newton algorithm will be provided a bit later.

Coming back to the posterior $\exp(\ell(\beta))$, Taylor expansion of $\ell(\beta)$ around the MLE $\hat{\beta}$ gives

$$f_{\beta|Y=y}(\beta) \propto \exp(\ell(\beta)) \approx \exp\left(\ell(\hat{\beta}) + \left\langle \nabla\ell(\hat{\beta}), \beta-\hat{\beta}\right\rangle + \frac{1}{2}(\beta-\hat{\beta})^T H\ell(\hat{\beta})(\beta-\hat{\beta})\right)$$

where $H\ell(\beta)$ denotes the Hessian of $\ell(\beta)$:

$$H\ell(\beta) = -\sum_{i=1}^n \frac{\exp(x_i'\beta)}{\left(1+\exp(x_i'\beta)\right)^2} x_i x_i'.$$

Because the $\ell(\hat{\beta})$ term is a constant, it can be ignored in proportionality. Also $\nabla \ell(\hat{\beta})$ equals zero. We thus have

$$f_{\beta|Y=y}(\beta) \propto \exp\left(\frac{1}{2}(\beta-\hat{\beta})^T H\ell(\hat{\beta})(\beta-\hat{\beta})\right) = \exp\left(-\frac{1}{2}(\beta-\hat{\beta})^T\left(-H\ell(\hat{\beta})\right)(\beta-\hat{\beta})\right)$$

We have switched above to $-H\ell(\hat{\beta})$ because this matrix is positive semi-definite as $\hat{\beta}$ maximizes $\ell(\beta)$. The above term is simply the unnormalized density of the multivariate normal distribution with mean $\hat{\beta}$ and covariance matrix $-H\ell(\hat{\beta})$. Observe that

$$-H\ell(\hat{\beta}) = \sum_{i=1}^{n} \frac{\exp(x_i'\hat{\beta})}{\left(1+\exp(x_i'\hat{\beta})\right)^2} x_i x_i'.$$

Now let W denote the $n \times n$ diagonal matrix whose i^{th} diagonal entry is

$$\frac{\exp(x_i'\hat{\beta})}{\left(1+\exp(x_i'\hat{\beta})\right)^2}$$

and also recall again that X is the $n \times p$ matrix with rows $x_1', \ldots, x_n'$. It is then easy to check that

$$-H\ell(\hat{\beta}) = X'WX. \tag{100}$$

The posterior normal approximation is thus

$$N(\hat{\beta}, (X'WX)^{-1}). \tag{101}$$

The standard errors corresponding to $\beta_0, \ldots, \beta_m$ can then be obtained by the square roots of the diagonal entries of $(X'WX)^{-1}$.

It turns out that Bayesian inference done with the above normal posterior approximation (101) coincides with the frequentist inference in the logistic regression model. It is easy to check this, say, in R (construct a 95% credible interval for, say, one of the components of β and then compare it with the frequentist interval). Thus the usual frequentist inference for the logistic regression model can be viewed from a Bayesian perspective. Note that the above analysis relies on two assumptions: (a) the prior for β is assumed to be uniform on the large cube $(-C, C)^p$, and (b) the posterior is approximated by a normal distribution. These assumptions may be of course not reasonable in a particular application. In such a situation, it is conceptually very clear as to how one would proceed: if the normal approximation to the posterior is not accurate, one needs to work with the actual posterior. If the uniform prior is not reasonable, one can do the full posterior analysis (or by taking a normal approximation to the posterior) for a more appropriate prior.

21.2 Details behind the Newton Algorithm for computing the MLE

The MLE $\hat{\beta}$ of β is the maximizer of $\ell(\beta)$. The maximizer of $\ell(\beta)$ cannot be computed in closed form. Newton's method is commonly used for maximizing $\ell(\beta)$. Newton's method uses the iterative scheme

$$\beta^{(m+1)} = \beta^{(m)} - \left(H\ell(\beta^{(m)})\right)^{-1} \nabla \ell(\beta^{(m)}) \tag{102}$$

As was saw in (99),

$$\nabla \ell(\beta) = \sum_{i=1}^{n} (y_i - \pi_i)\, x_i$$

where π_i is given by

$$\pi_i = \frac{\exp(x_i'\beta)}{1+\exp(x_i'\beta)}$$

Letting π be the $n \times 1$ vector with entries $\pi_1, \ldots, \pi_n$, we can write $\nabla \ell(\beta)$ in matrix notation as (note that X has rows $x_1^T, \ldots, x_n^T$ or, equivalently, X^T has columns $x_1, \ldots, x_n$):

$$\nabla \ell(\beta) = X^T (Y - \pi)$$

where, as usual in regression, Y denotes the $n \times 1$ vector of response values. Also from (100), we can write

$$H\ell(\beta) = -X^T W X$$

where W is the $n \times n$ diagonal matrix whose i^{th} diagonal entry is $\pi_i(1-\pi_i)$. Newton's iterative scheme (102) therefore becomes

$$\beta^{(m+1)} = \beta^{(m)} + (X^T W X)^{-1} X^T (Y - \pi).$$

This can be rewritten as

$$\beta^{(m+1)} = (X^T W X)^{-1} X^T W Z \tag{103}$$

where

$$Z = X\beta^{(m)} + W^{-1}(Y - \pi). \tag{104}$$

The method of obtaining the MLE $\hat{\beta}$ therefore proceeds iteratively as follows. First have an initial estimate of β. Call this initial estimator $\hat{\beta}^{(0)}$. Use this estimator to calculate p_i via

$$\pi_i = \frac{\exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \cdots + \hat{\beta}_p^{(0)} x_{ip})}{1 + \exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} \cdots + \hat{\beta}_p^{(0)} x_{ip})}.$$

Use these values of π_i to create the response variable values Z_i via (104) and also use values of π_i to construct the matrix W. With Z and W, we can estimate β via

$$\hat{\beta}^{(1)} = (X^T W X)^{-1} X^T W Z.$$

Now replace the initial estimator $\hat{\beta}^{(0)}$ by $\hat{\beta}^{(1)}$ and repeat this process. Keep repeating this until two successive estimates $\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m+1)}$ do not change much. At that point, stop and report the estimate of β in the logistic regression model as $\hat{\beta}^{(m)}$.

The expression $(X^T W X)^{-1} X^T W Z$ is reminiscent of the usual $(X^T X)^{-1} X^T Y$ which is the usual estimate of β in the linear model. In fact, this is the least squares estimate in a weighted least squares model.

22 Lecture Twenty Two

22.1 Linear Regression Recap

So far we have studied the linear regression model:

$$Y = X\beta + \epsilon \qquad \text{with } \epsilon_1, \ldots, \epsilon_n \overset{\text{i.i.d}}{\sim} N(0, \sigma^2)$$

under the prior that the components of β are i.i.d Unif$(-C, C)$ for a large C. We have seen (Problem 1(a) in Homework Five) that

$$\beta \mid \text{data}, \sigma \sim N_p\left(\hat{\beta}, \sigma^2 (X^T X)^{-1}\right) \tag{105}$$

where p is the dimension of β. This is the posterior distribution of β conditional on σ. This cannot be used for inference on β because σ is unknown. The posterior of β (without any conditioning on σ) is given by (under the prior $\log \sigma \sim \text{Unif}(-C, C)$)

$$\beta \mid \text{data} \sim t_{n-p,p}\left(\hat{\beta}, \hat{\sigma}^2 (X^T X)^{-1}\right) \tag{106}$$

where $\hat{\sigma}$ is the residual standard error. When $n - p$ is large, the t-distribution will be quite close to normal, so we can write

$$\beta \mid \text{data is approximately } N_p\left(\hat{\beta}, \sigma^2 (X^T X)^{-1}\right). \tag{107}$$

Inference on β is done either using (106) or (107).

22.2 Linear Regression with Gaussian prior

It is common to consider other priors for linear regression. A general class of priors is given by

$$\beta \sim N_p(m_0, Q_0) \tag{108}$$

for a mean vector m_0 and covariance matrix Q_0. In this case, it can be shown (left as exercise) that the posterior distribution of β conditional on σ is given by

$$\beta \mid \text{data}, \sigma \sim N_p\left(m_1, Q_1\right) \tag{109}$$

where

$$m_1 = \left(Q_0^{-1} + \frac{1}{\sigma^2} X^T X\right)^{-1} \left(Q_0^{-1} m_0 + \frac{1}{\sigma^2} X^T Y\right) \quad \text{and} \quad Q_1 = \left(Q_0^{-1} + \frac{1}{\sigma^2} X^T X\right)^{-1}.$$

(109) is the analogue of (105) for the Gaussian prior (108). Actually, the result (105) can be seen as a special case of (109) when the prior covariance Q_0 becomes large (think of the setting where the smallest eigenvalue of Q_0 goes to ∞). Because when Q_0 approaches infinity, it is easy to see that

$$\begin{aligned} m_1 &= \left(Q_0^{-1} + \frac{1}{\sigma^2} X^T X\right)^{-1} \left(Q_0^{-1} m_0 + \frac{1}{\sigma^2} X^T Y\right) \\ &\to \left(\frac{1}{\sigma^2} X^T X\right)^{-1} \left(\frac{1}{\sigma^2} X^T Y\right) = (X^T X)^{-1} X^T Y = \hat{\beta} \end{aligned}$$

and also

$$Q_1 = \left(Q_0^{-1} + \frac{1}{\sigma^2} X^T X\right)^{-1} \to \sigma^2 \left(X^T X\right)^{-1}.$$

As one concrete example of Q_0 being large, think of $Q_0 = CI_p$ when C is large. The posterior for this prior is the same as (105) (i.e., there is no difference between the $\text{Unif}(-C, C)$ and $N_p(0, CI_p)$ priors) . The Gaussian prior result (109) can thus be seen as a generalization of (105).

In many applications, it makes sense to work with the prior (108) as opposed to the $Unif(-C, C)$ prior. We shall illustrate in a real data setting in the next section.

22.3 Linear Regression on an Earnings Dataset

Consider the dataset `ex1029` from the R package `Sleuth3` (this package is written by Ramsey and Schafer to accompany their introductory statistics book *Statistical Sleuth*). This dataset contains weekly wages in 1987 for a sample of $n = 25682$ males between the ages of 18 and 70 who worked full-time along some covariates including theiry years of experience. We shall work with the two variables:

$$y = \text{response variable} = \log(\text{weekly earnings})$$

and

$$x = \text{years of experience}.$$

We shall fit linear regression models of y on x. The reason for working with log(earnings) as opposed to earnings directly is for better interpretation.

The most basic model between y and x is the usual linear model:

$$y = \beta_0 + \beta_1 x + \epsilon.$$

From a visual examination of the scatterplot between y and x, it can be easily seen that this simple linear regression model is not adequate as the relationship between y and x is clearly nonlinear (y increases with x for small values of x and decreses with x for large values of x). A more suitable model is

$$y = \beta_0 + \beta_1 x + \beta_2 (x - s)_+ + \epsilon \tag{110}$$

where s is also an unknown parameter. This model fits two lines which are connected at the point s. The rate of change of y with x is β_1 for $x \leq s$ and $\beta_1 + \beta_2$ for $x > s$. We have previously seen how to fit models of the form (110) to data.

For this specific dataset, the model (110) is not suitable either because there is no reason to just have one change of slope for the regression function. A more suitable model would be

$$y = \beta_0 + \beta_1 x + \beta_2 (x - s_1)_+ + \beta_3 (x - s_2)_+ + \cdots + \beta_{k+1} (x - s_k)_+ + \epsilon$$

for a k that is not too small. There are two problems with working with this model:

1. It is difficult to fit it to the data unless k is very small. The methodology that we studied in Lecture 20 involved marginalizing the β's and σ to get a posterior only for $s_1, \ldots, s_k$. This is a k-dimensional posterior and a grid-based method for selecting the posterior mode or mean would be quite computationally challenging if k is not small.

2. It is difficult to select a suitable value of k.

One way to circumvent these problems is to introduce a change of slope term $(x - s)_+$ at every possible value of s. In this dataset, the variable x takes the values $0, 1, \ldots, 63$. So we consider the model:

$$y = \beta_0 + \beta_1 x + \beta_2 (x - 1)_+ + \beta_3 (x - 2)_+ + \cdots + \beta_{64} (x - 63)_+ + \epsilon \tag{111}$$

This is the model that we shall work with. It is a linear regression model with 65 coefficients. The intercept β_0 is interpreted as the log(earnings) for someone who is starting their career ($x = 0$). Also, for $j = 1, \ldots, 63$, the term $100\beta_j$ is interpreted as the percent change in earnings when someone moves from $(j-1)$ years of experience to j years in experience. [**This interpretation is actually incorrect; see the notes for the next lecture (Lecture 23) for the correct interpretation**]

How to fit the model (111) to the observed data. The first approach is to work with the uniform prior $\beta_j \sim \text{Unif}(-C, C)$ for $j = 0, 1, \ldots, 63$. This is equivalent to just doing the usual linear regression (using the R function `lm` for instance). This gives the least squares estimates $\hat{\beta}_0^{ls}, \ldots, \hat{\beta}_{63}^{ls}$ and the function that we use for explaining the relationship between earnings and experience is $y = \hat{f}^{ls}(x)$ where

$$\hat{f}^{ls}(x) := \hat{\beta}_0^{ls} + \hat{\beta}_1^{ls} x + \hat{\beta}_2^{ls}(x-1)_+ + \hat{\beta}_3^{ls}(x-2)_+ + \cdots + \hat{\beta}_{64}^{ls}(x-63)_+. \tag{112}$$

In this particular dataset, this function turns out to be somewhat wiggly and not smooth, which is not very interpretable. The situation is more pronounced when the number of observations n not large. One can reduce the size of this dataset to, say, $n = 500$ by sampling 500 observations (rows) at random from this dataset. One can then refit the least squares estimate of y on $x, (x-1)_+, \ldots, (x-63)_+$ to this smaller dataset. Here the fitted function (112) will be much more wiggly.

In order to obtain a smooth function fit to the data, one can use the prior

$$\beta_0 \sim N(0, C) \quad \text{and} \quad \beta_1, \ldots, \beta_{64} \overset{\text{i.i.d}}{\sim} N(0, \tau^2)$$

for a small τ. Here, if we take τ to be small, we are insisting on $\beta_1, \ldots, \beta_{64}$ to be small which will lead to a smoother fit. The assumption $\beta_0 \sim N(0, C)$ on β_0 is very similar to $\beta_0 \sim \text{Unif}(-C, C)$ and it just says that we do not enforce anything on β_0 a priori. We can write this prior as

$$\beta \sim N_{65}(m_0, Q_0)$$

where β is the 65×1 vector with components $\beta_0, \beta_1, \ldots, \beta_{64}$, m_0 is the 65×1 vector of zeros, and Q_0 is the 65×65 diagonal matrix with diagonal entries $C, \tau^2, \tau^2, \ldots, \tau^2$. The posterior distribution of β can then be calculated using (109) as:

$$\beta \mid \text{data}, \sigma \sim N_{65}\left(\left(Q_0^{-1} + \frac{1}{\sigma^2}X^T X\right)^{-1} \frac{1}{\sigma^2} X^T Y, \left(Q_0^{-1} + \frac{1}{\sigma^2}X^T X\right)^{-1}\right).$$

This posterior can be used for inference on β. Note that it depends on τ and σ. One can take a small value for τ if smooth function fit is desired. For σ, one can take a prior such as $\log \sigma \sim \text{Unif}(-C, C)$ and calculate the marginal posterior of β given the data alone (another method is described in the next section). The posterior mean is

$$\tilde{\beta}^\tau = \left(Q_0^{-1} + \frac{1}{\sigma^2}X^T X\right)^{-1} \frac{1}{\sigma^2} X^T Y$$

which can be used to get the function fit:

$$\tilde{f}^\tau(x) := \tilde{\beta}_0^\tau + \tilde{\beta}_1^\tau x + \tilde{\beta}_2^\tau (x-1)_+ + \tilde{\beta}_3^\tau (x-2)_+ + \cdots + \tilde{\beta}_{64}^\tau (x-63)_+.$$

When τ is small, it can be checked that $\tilde{f}^\tau(x)$ will be a very smooth function of x. If τ is really really small, then $\tilde{f}^\tau(x)$ will be essentially a constant.

This is a pretty straightforward methodology for fitting a smooth function of experience to the log(earnings) data. However, the key issue is the choice of the tuning parameter τ. Here is where probability gives a very nice solution. This is discussed next.

22.4 Choosing the tuning parameter τ

The choice of τ is quite crucial to this analysis. If τ is large, then the estimator $\tilde{\beta}^\tau$ will be very similar to the least squares estimator $\hat{\beta}^{ls}$ so that the fitted function $\tilde{f}^\tau$ will be quite

wiggly. On the other hand, if τ is extremely small, then the fitted function $\tilde{f}^{\tau}$ will be basically constant which would not be useful. So we our ideal choice for τ should be neither too large nor too small. How do we make this choice?

Here is how probability theory solves this problem. We shall discuss choices for τ as well as for σ which is also an unknown parameter that needs to be chosen in order to calculate the estimate $\tilde{\beta}^{\tau}$ and $\tilde{f}^{\tau}$. The idea is simply to treat τ and σ as unknown parameters and put priors on them. We shall use the prior:

$$\log \tau \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C) \quad \text{and} \quad \log \sigma \overset{\text{i.i.d}}{\sim} \text{Unif}(-C, C)$$

Note that this prior implies that we are allowing essentially (because C is large) all possible values of τ and σ. In particular, we are not *a priori* ruling out large τ because we don't like wiggly solutions. We then compute the posterior of τ and σ as:

$$f_{\tau,\sigma|\text{data}}(\tau, \sigma) \propto f_{\tau,\sigma}(\tau, \sigma) f_{\text{data}|\tau,\sigma}(\text{data}). \tag{113}$$

We need to calculate the likelihood term above which is the conditional distribution of the data given τ and σ alone. As the model specifies the distribution of the data $Y_1 \ldots, Y_n$ in terms of β, we need to integrate out β to obtain the likelihood in terms of τ and σ. Fortunately, this integral can be obtained in closed form because of the following result:

$$\beta \sim N_p(m_0, Q_0) \text{ and } Y \mid \beta \sim N_n(X\beta, \sigma^2 I_n) \implies Y \sim N\left(Xm_0, XQ_0X^T + \sigma^2 I_n\right).$$

Therefore (note that for us m_0 is the zero vector)

$$f_{\text{data}|\tau,\sigma}(\text{data}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n (\det \Sigma)^{-1/2} \exp\left(-\frac{1}{2} Y^T \Sigma^{-1} Y\right) \qquad \text{with } \Sigma := XQ_0X^T + \sigma^2 I_n.$$

Recall that Q_0 is the diagonal matrix with diagonal entries $C, \tau^2, \ldots, \tau^2$. Throughout C is a large constant (in the analysis of the Earnings data, I took $C = 10^6$). We can use this likelihood in (113) to calculate the posterior of τ and σ. We can get a grid based discrete appriximation of this posterior. Generally, this posterior will be peaked around the maximum likelihood values $\hat{\tau}$ and $\hat{\sigma}$ so one can obtain a simpler procedure by just taking τ and σ to be the maximum likelihood values.

In the Earnings dataset, this procedure can be applied to the dataset of size 500 (randomly sampled from the original dataset). One can also apply this to the full dataset but the matrix inversions appearing above can be somewhat slow (some linear algebra tricks can be used to make this implementable for larger n). This analysis leads to a fairly small value of $\hat{\tau}$ that leads to a smooth function fit $\tilde{f}^{\hat{\tau}}$. There is something quite interesting here. There is a big difference between the two likelihoods:

$$f_{\text{data}|\beta,\sigma}(\text{data}) \quad \text{and} \quad f_{\text{data}|\tau,\sigma}(\text{data}).$$

Indeed, maximizing $f_{\text{data}|\beta,\sigma}(\text{data})$ leads to the least squares estimate $\hat{\beta}^{ls}$ which would be quite wiggly. On the other hand, maximizing $f_{\text{data}|\tau,\sigma}(\text{data})$ leads to a fairly small estimate of $\hat{\tau}$ leading to a smooth function fit $\tilde{f}^{\hat{\tau}}$. The reason for this discrepancy can be understood by noting that

$$f_{\text{data}|\tau,\sigma}(\text{data}) = \int f_{\text{data}|\beta,\sigma}(\text{data}) f_{\beta|\tau}(\beta) d\beta.$$

When τ is large, the term $f_{\beta|\tau}(\beta)$ will be small simply because the normal density with variance τ^2 will be flat for large τ. On the other hand, when τ is too small, the weight $f_{\beta|\tau}(\beta)$ will be significant only for very smooth βs but these βs will have poor values for $f_{\text{data}|\beta,\sigma}(\text{data})$.

Let me stress once more that this method for choosing τ does not *a priori* prefer small values of τ. The marginal or integrated likelihood automatically selects a value of τ that is small because it gives the best likelihood for the observed data.

It should be emphasized that the integrated likelihood $f_{\text{data}|\tau,\sigma}(\text{data})$ really does not exist in frequentist statistics so this method of tuning parameter selection is largely Bayesian.

22.5 Additional Comments and References

The method for regression with the prior $N(0, \tau^2)$ is very similar to Ridge Regression (`https://en.wikipedia.org/wiki/Ridge_regression`). Usually, the tuning parameter in Ridge regression is selected via Cross-Validation which is a method that is quite different from the above approach using the integrated or marginal likelihood. For a somewhat nuanced discussion on the benefits of Bayesian tuning parameter selection and Cross Validation, see `http://www.inference.org.uk/mackay/Bayes_FAQ.html#cv`.

If you want to read more into this approach for high-dimensional models, I strongly recommend the 1992 paper titled *Bayesian Interpolation* by David MacKay (MacKay gives a short summary of this paper in this blog post: `https://statmodeling.stat.columbia.edu/2011/12/04/david-mackay-and-occams-razor/`).

23 Lecture Twenty Three

23.1 Comments on the Coefficient Interpretation in Last Class's Regression Model

In the last class, we studied a regression model for $y = \log(\text{Earnings})$ in terms of $x =$ Years of Experience. The interpretation I described for the coefficient parameters in that model was incorrect. The correct interpretation is given below. The simplest model for y in terms of x is the linear model which corresponds to the equation (without the error term):

$$y = \beta_0 + \beta_1 x. \tag{114}$$

The interpretation of β_0 and β_1 in this model are very clear. β_0 is the simply the value of $y = \log(\text{Earnings})$ when $x = 0$ (i.e., for someone just joining the workforce). To obtain the interpretation for β_1, first plug in $x = 1$ in (114) and then $x = 0$ and subtract the second equation from the first. This gives

$$\beta_1 = y_1 - y_0 = \log(E_1) - \log(E_0) = \log\left(\frac{E_1}{E_0}\right) \approx \frac{E_1 - E_0}{E_0}.$$

HEre E_0 and E_1 are Earnings for $x = 0$ and $x = 1$ respectively, and in the last equality, we used $\log(u) \approx u - 1$ for $u \approx 1$. Thus $100\beta_1$ represents the increment (in percent) in salary from year 0 to year 1. For example, $\beta_1 = 0.05$ means that the salary increases by 5% from year 0 to year 1.

Now let us consider the model

$$y = \beta_0 + \beta_1 x + \beta_2 (x - 1)_+ \tag{115}$$

Here the interpretation of β_0 and β_1 are exactly the same as for the model (114). β_0 again represents log(Earnings) for $x = 0$ and β_1 represents the increment (in percent) from year 0

to year 1. What is the interpretation for β_2? It is easy to see that:

$$\log E_2 = \beta_0 + 2\beta_1 + \beta_2 \quad \log E_1 = \beta_0 + \beta_1 \quad \log E_0 = \beta_0.$$

Thus

$$\beta_2 = \log E_2 - 2\log E_1 + \log E_0 = \log \frac{E_2}{E_1} - \log \frac{E_1}{E_0} \approx \frac{E_2 - E_1}{E_1} - \frac{E_1 - E_0}{E_0}.$$

Thus $100\beta_2$ represents the change in the percent increment between years 1 and 2 compared to the percent increment between years 0 and 1. For example, $\beta_2 = 0.0003$ means that the percent increment decreases by 0.03 after year 2. If $\beta_1 = 0.05$, we would have a 5% increment after year 1 and a $5 - 0.03 = 4.97\%$ increment after year 2.

Now consider the model that we actually used last time:

$$y = \beta_0 + \beta_1 x + \beta_2 (x-1)_+ + \beta_3 (x-2)_+ + \cdots + \beta_{64}(x-63)_+. \tag{116}$$

Here the interpretation for $\beta_0, \beta_1, \beta_2$ are just the same as in Model (115). More generally, the interpretation for $\beta_j, j \geq 2$ is as follows: $100\beta_j$ is the change in the percent increment between years $j-1$ and j compared to the percent increment between years $j-2$ and $j-1$. For a concrete example, suppose

$$\beta_0 = 5.74 \quad \beta_1 = 0.05 \quad \beta_2 = -0.0003 \quad \beta_3 = -0.0008 \quad \beta_4 = -0.001 \quad \ldots$$

then

1. weekly earnings for someone just joining the workforce is $\exp(5.74) = \$311.06$,
2. increment after year 1 is 5%,
3. increment after year 2 is $(5 - 0.03) = 4.97\%$,
4. increment after year 3 is $(4.97 - 0.08) = 4.89\%$,
5. increment after year 4 is $(4.89 - 0.1) = 4.79\%$, and so on.

If all $\beta_j, j \geq 2$ are negative, then, after a while, the increments may become negative which means that the salary actually starts decreasing after a certain number of years of experience.

It should be clear from the above that β_0, β_1 and $\beta_j, j \geq 2$ are different kinds of parameters (they have different units for instance). In particular, we would expect $\beta_j, j \geq 2$ to be quite small. In the last class, we analyzed model (116) with the prior

$$\beta_0 \sim N(0, C) \quad \text{and} \quad \beta_1, \ldots, \beta_{64} \overset{\text{i.i.d}}{\sim} N(0, \tau^2)$$

for a parameter $\tau > 0$ (and large C). This analysis seems to treat β_1 as well as $\beta_j, j \geq 2$ in the same way. A better prior would be:

$$\beta_0 \sim N(0, C) \;\; \beta_1 \sim N(0, C) \quad \text{and} \quad \beta_2, \ldots, \beta_{64} \overset{\text{i.i.d}}{\sim} N(0, \tau^2) \tag{117}$$

23.2 Comments on Regularization

Parameter estimation for the model (116) using the prior (117) is an example of regularization. Regularization is one of the most important ideas in statistics and machine learning in the past 30 years. The reason for regularization is that, in models with lots of parameters such as (116), standard (unregularized) estimation procedures give nonsensical answers. For

example, for the model (116) applied to 500 randomly selected observations from the full `ex1029` dataset from the R package `Sleuth3`, the usual least squares estimates are given by

$$\beta_0 = 5.55 \quad \beta_1 = 0.045 \quad \beta_2 = 0.312 \quad \beta_3 = -0.431 \quad \beta_4 = 0.299 \quad \ldots$$

The interpretation would then become

1. weekly earnings for someone just joining the workforce is $\exp(5.55) = \$257.24$,
2. increment after year 1 is 4.5%,
3. increment after year 2 is $(4.5 + 31.2) = 35.7\%$,
4. increment after year 3 is $(35.7 - 43.1) = -7.4\%$,
5. increment after year 4 is $(-7.4 + 29.9) = 22.5\%$, and so on.

These increments fluctuate wildly so as to make these numbers nonsensical. This is the reason why one regularizes while dealing with many parameters. In the context of the model (116), regularization is done in the following ways:

1. **The common approach**: The common approach estimates $\beta_0, \ldots, \beta_m$ (here $m = 64$) by minimizing the least squares criterion plus a penalty term which encourages $\beta_2, \ldots, \beta_m$ to be small. One way of doing this is via the minimization of

$$\sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i - \beta_2 (x_i - 1)_+ - \beta_3 (x_i - 2)_+ - \cdots - \beta_m (x_i - (m-1))_+)^2 + \lambda \left(\beta_2^2 + \cdots + \beta_m^2\right)$$

   for a suitable tuning parameter λ. When λ is large, the minimizer of the above criterion will have $\beta_j, 2 \leq j \leq m$ small. But if λ is too large, then $\beta_j, 2 \leq j \leq m$ will be very close to zero. On the other hand, if λ is too small, then $\beta_j, 2 \leq j \leq m$ will be close to the least squares estimates. The choice of λ is obviously quite crucial. For this, one uses the idea that "Regularized estimates often lead to better predictions". To use this idea in practice for selecting λ, the original dataset is divided into two subsets: training and test datasets. One would fit a regularized model to the training dataset by minimizing the above criterion for each value of λ. The value of λ which minimizes average prediction error on the test dataset would then be selected. This is the basic idea underlying methods such as cross-validation. This approach is quite common although there are some issues such as: (a) there are no principled approaches for doing the training-test splits, and often different splits lead to different answers, and (b) the value of λ leading to best predictions on the test dataset might lead to estimates of β_j's that still fluctuate quite a bit (although not as much as the unpenalized least squares estimates).

2. **The Probability/Bayesian approach**: This is the approach that we discussed in the last lecture. The starting point is the observation that the usual least squares estimate can be seen as Bayesian estimates corresponding to the prior:

$$\beta_0, \beta_1, \beta_2, \ldots, \beta_m \overset{\text{i.i.d}}{\sim} N(0, C) \tag{118}$$

   for a large C. To achieve regularization, one changes the above model to

$$\beta_0, \beta_1 \overset{\text{i.i.d}}{\sim} N(0, C) \quad \text{and} \quad \beta_2, \ldots, \beta_m \overset{\text{i.i.d}}{\sim} N(0, \tau^2) \tag{119}$$

   where $\tau > 0$ is an unknown parameter. The modeling assumption (119) is more flexible than (118). Indeed, (119) includes (118) as a special case when $\tau = \sqrt{C}$. Using the

prior assumption (119), one can calculate the (marginal) likelihood of σ and τ by integrating the conditional density of the data given β with respect to the prior (119):

$$f_{\text{data}|\tau,\sigma}(\text{data}) = \int f_{\text{data}|\beta}(\text{data}) f_{\beta|\tau,\sigma}(\beta) d\beta.$$

The best value of τ would then be obtained by maximizing this integrated likelihood over τ and σ.

These two methods for performing regularization are actually quite different. There is no training-test split in the Bayesian approach. On the other hand, there is no such thing as integrated or marginal likelihood in the common approach. For a comparative discussion on the benefits of the Bayesian approach, see `http://www.inference.org.uk/mackay/Bayes_FAQ.html#cv`.

If you want to read more into the Bayesian approach for high-dimensional models, I strongly recommend the 1992 paper titled *Bayesian Interpolation* by David MacKay (MacKay gives a short summary of this paper in this blog post: `https://statmodeling.stat.columbia.edu/2011/12/04/david-mackay-and-occams-razor/`).

24 Lecture Twenty Four

24.1 Central Limit Theorem (CLT)

The following is the simplest version of the CLT.

Theorem 24.1 (Central Limit Theorem)**.** *Suppose $X_i, i = 1, 2, \ldots$ are i.i.d with $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2 < \infty$. Then, with $\bar{X}_n = (X_1 + \cdots + X_n)/n$,*

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

converges in distribution to $N(0, 1)$. Convergence in distribution here means that

$$\mathbb{P}\{a \leq \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \leq b\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \qquad \text{for all } -\infty \leq a < b \leq +\infty.$$

Informally, the CLT says that for i.i.d observations $X_1, \ldots, X_n$ 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. $S_n = X_1 + \cdots + X_n$ 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. 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 (proved later). 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 previously 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^2_n). The CLT therefore implies that the χ^2_n 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. A sketch of this proof is given later in this lecture.

24.2 CLT Proof strategy

To prove the CLT, the natural idea is to write down some sort of formula for

$$\mathbb{P}\left\{a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq b\right\}$$

and then see what happens to the formula as $n \to \infty$. The problem with this approach is that the formula is a bit tricky to write (it depends on whether the random variables

are discrete or continuous, for example). Even if we assume that the random variables are discrete, the formula is a bit messy. For example, suppose that $X_1, X_2, \ldots$ are i.i.d discrete random variables taking the values $0, 1, 2, \ldots$. Then

$$\begin{aligned}
&\mathbb{P}\left\{a \leq \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq b\right\} \\
&= \mathbb{P}\left\{n\mu + a\sigma\sqrt{n} \leq X_1 + \cdots + X_n \leq n\mu + b\sigma\sqrt{n}\right\} \\
&= \sum_{x_1,\ldots,x_n : x_j \in \{0,1,2,\ldots\}} p_{x_1} p_{x_2} \ldots p_{x_n} I\left\{n\mu + a\sigma\sqrt{n} \leq x_1 + \cdots + x_n \leq n\mu + b\sigma\sqrt{n}\right\}
\end{aligned}$$

where $p_j := \mathbb{P}\{X_1 = j\}$. Understanding the behaviour of this as n gets large is a bit tricky.

For this reason, while dealing with sums of independent random variables, people usually work with certain transforms of distributions.

24.3 Transforms

There are three commonly used transforms: z-transform (also known as Probability Generating Function) for random variables taking values in $\{0, 1, 2, \ldots\}$, Laplace transform (also known as Moment Generating Function), and the Fourier transform (also known as the Characteristic Function).

24.3.1 z-Transform (Probability Generating Function)

Suppose X is a discrete random variable taking the values $0, 1, 2, \ldots$. The z-transform of X is defined as

$$G_X(z) := \mathbb{E}\left(z^X\right) = \sum_{j=0}^{\infty} z^j \mathbb{P}\{X = j\}.$$

Here is a simple example.

Example 24.2. *Suppose X takes the values $8, 13, 20, 29, 35$ with probabilities $0.3, 0.2, 0.15, 0.3, 0.05$. Then the z-transform of X is given by*

$$G_X(z) = 0.3z^8 + 0.2z^{13} + 0.15z^{20} + 0.3z^{29} + 0.05z^{35}.$$

In general $G_X(z)$ will be a complicated power series with many terms. In some cases however, the series corresponding to $G_X(z)$ can be summed explicitly to yield a simpler formula for $G_X(z)$ as in the next example.

Example 24.3. *Suppose X has the Poisson distribution with mean λ. Then*

$$G_X(z) = \sum_{j=0}^{\infty} z^j \mathbb{P}\{X = j\} = \sum_{j=0}^{\infty} z^j e^{-\lambda} \frac{\lambda^j}{j!} = e^{-\lambda} \sum_{j=0}^{\infty} \frac{(\lambda z)^j}{j!} = e^{-\lambda} e^{\lambda z} = \exp\left(\lambda(z-1)\right).$$

Note that $G_X(z)$ uniquely determines the distribution of X because

$$\mathbb{P}\{X = k\} = \frac{d^k}{dz^k} G_X(z)\Big|_{z=0}$$

for every $k = 0, 1, 2, \ldots$.

The z-transform (and all other transforms) have the important property that the transform for the sum of n i.i.d random variables is simply the transform of the individual random variable raised to power n. This is because:

$$\begin{aligned} G_{X_1+\ldots,+X_n}(z) &= \mathbb{E}\left(z^{X_1+\cdots+X_n}\right) \\ &= \mathbb{E}\left(z^{X_1}z^{X_2}\ldots z^{X_n}\right) = \left(\mathbb{E}z^{X_1}\right)\left(\mathbb{E}z^{X_2}\right)\ldots\left(\mathbb{E}z^{X_n}\right) = \left(G_{X_1}(z)\right)^n. \end{aligned}$$

The utility of this result for dealing with sums of i.i.d random variables can be found in the following two examples.

Example 24.4. *Suppose $X_1, \ldots, X_{100}$ are i.i.d with common distribution giving probabilities $0.3, 0.2, 0.15, 0.3, 0.05$ to the values $8, 13, 20, 29, 35$ with probabilities. What is the distribution of $X_1 + \cdots + X_{100}$?*

$X_1 + \cdots + X_{100}$ will be a discrete random variable taking many possible values. Specifying its distribution via the probability mass function will be tedious. But it is very easy to write down its z-transform as

$$G_{X_1+\cdots+X_{100}}(z) = \left(G_{X_1}(z)\right)^{100} = \left(0.3z^8 + 0.2z^{13} + 0.15z^{20} + 0.3z^{29} + 0.05z^{35}\right)^{100}.$$

Example 24.5. *The following is a fundamental fact. The sum of n independent random variables $X_1, \ldots, X_n$ with $X_i \sim Poi(\lambda_i)$ for $i = 1, \ldots, n$ equals $Poi(\lambda_1 + \cdots + \lambda_n)$. This can be very easily proved using z-transforms (and the fact that the z-transform of $Poi(\lambda)$ equals $\exp(\lambda(z-1))$) because*

$$\begin{aligned} G_{X_1+\cdots+X_n}(z) &= G_{X_1}(z)G_{X_2}(z)\ldots G_{X_n}(z) \\ &= \exp\left(\lambda_1(z-1)\right)\exp\left(\lambda_2(z-1)\right)\ldots\exp\left(\lambda_n(z-1)\right) \\ &= \exp\left((\lambda_1 + \ldots\lambda_n)(z-1)\right). \end{aligned}$$

Thus the z-transform of $X_1+\cdots+X_n$ coincides with that of $Poi(\lambda_1+\cdots+\lambda_n)$ which implies that $X_1 + \cdots + X_n$ has the $Poi(\lambda_1 + \cdots + \lambda_n)$ distribution.

While the z-transform makes dealing with independent sums convenient, it is not a general tool as it is only defined for discrete random variables taking the values $0, 1, 2, \ldots$. This is the reason for considering Laplace and Fourier transforms.

24.3.2 Laplace Transform (Moment Generating Function)

The Laplace transform of a random variable X is defined as the function:

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

for all $t \in (-\infty, \infty)$ for which $\mathbb{E}(e^{tX}) < \infty$. Note that $M_X(0) = 1$.

Example 24.6 (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 Laplace transform is defined for every random variable X (although for some random variables $M_X(t)$ can be $+\infty$ for many values of t) unlike the z-transform which is only defined for random variables taking values in $\{0, 1, 2, \dots\}$. Just like the z-transform, the Laplace transform also factorizes for independent random variables. Indeed, if $X_1, \dots, X_n$ are independent, then

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

This is a consequence of the fact that

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

the last equality being a consequence of independence.

The Laplace transform is known as the Moment Generating Function because it allows one to easily read off the moments of the random variable. For $k \geq 1$, the number $\mathbb{E}(X^k)$ is called the k^{th} moment of X. Knowledge of $M_X(t)$ allows one to easily read off the moments of X because the power series expansion of $M_X(t)$ 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 because

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

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 application, we can deduce the moments of the standard normal distribution from the fact that its Laplace Transform 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 Central Limit Theorem can be established using Laplace transforms in the following way.

Proof of the CLT with Laplace Transforms. 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 converges in distribution to $N(0, 1)$. We shall show that the Laplace transform of Y_n converges of the Laplace transform of $N(0, 1)$ which is $e^{t^2/2}$:

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

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,

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

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

$$\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a \qquad \text{provided } \lim_{n\to\infty} a_n = a \tag{120}$$

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 (120). It remains to prove (120). 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 (120). □

The above proof of the CLT has two deficiencies:

1. It tacitly assumes that the moment generating function of $X_1, \ldots, X_n$ exists for all t. This is much stronger than the existence of the variance of X_i (which is all the CLT needs). Indeed if $M_X(t)$ exists for all t in any open interval containing zero, then moments of all orders (not just the variance) exist.

2. We have proved that the Laplace transform of $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ converges to that of $N(0,1)$. It is not clear though as to how this implies that

$$\mathbb{P}\left\{a \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq b\right\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

   for $-\infty \leq a < b \leq \infty$.

To fix these two deficiencies, one works with the Fourier transform for proving the CLT.

24.3.3 Fourier Transform (Characteristic Function)

The Fourier transform of a random variable X is defined as the function:

$$\phi_X(t) := \mathbb{E}\left(e^{itX}\right) = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX)$$

for all $t \in (-\infty, \infty)$. Here $i = \sqrt{-1}$. The Fourier transform is defined for every random variable and it is finite for all $t \in (-\infty, \infty)$. This is because $\cos(tX)$ and $\sin(tX)$ are always bounded by 1 so the expectation will obviously be finite. For example, suppose X has the Cauchy distribution with density:

$$f_X(x) := \frac{1}{\pi(1+x^2)}.$$

Then it is easy to check that the Laplace transform $M_X(t)$ will equal $+\infty$ for all $t \neq 0$. On the other hand, the Fourier transform of X is

$$\phi_X(t) = \mathbb{E}e^{itX} = \int_{-\infty}^{\infty} \frac{e^{itx}}{\pi(1+x^2)} dx.$$

It turns out that the above integral equals $e^{-|t|}$ so that

$$\phi_X(t) = e^{-|t|}.$$

You should find a proof of the above online. Just like the other two transforms, the Fourier transform factorizes for independent random variables. Indeed, if $X_1, \ldots, X_n$ are independent, then

$$\phi_{X_1+\cdots+X_n}(t) = \phi_{X_1}(t)\phi_{X_2}(t)\ldots\phi_{X_n}(t).$$

This is a consequence of the fact that

$$\mathbb{E}e^{it(X_1+\cdots+X_n)} = \mathbb{E}\left(\prod_{j=1}^{n} e^{itX_j}\right) = \prod_{j=1}^{n} \mathbb{E}e^{itX_j},$$

the last equality being a consequence of independence.

Example 24.7. *The following is a standard fact. If $X_1, X_2, \ldots, X_n$ are i.i.d random variables having the Cauchy distribution. Then their mean $\bar{X}_n := (X_1 + \cdots + X_n)/n$ also has the Cauchy distribution. This can be proved using Fourier transforms as follows. The Fourier transform of $\bar{X}_n$ equals*

$$\begin{aligned}\phi_{\bar{X}_n}(t) &= \mathbb{E}e^{it\bar{X}_n} \\ &= \mathbb{E}\exp\left(i\frac{t}{n}\left(X_1 + \cdots + X_n\right)\right) = \phi_{X_1+\cdots+X_n}\left(\frac{t}{n}\right) = \left(\phi_{X_1}\left(\frac{t}{n}\right)\right)^n\end{aligned}$$

Because the Fourier transform of the Cauchy distribution equals $e^{-|t|}$, we get

$$\phi_{\bar{X}_n}(t) = \left(\exp\left(-|t|/n\right)\right)^n = \exp(-|t|).$$

Thus the Fourier transform of $\bar{X}_n$ also equals $e^{-|t|}$ which implies that $\bar{X}_n$ also has the Cauchy distribution.

In the next class, we shall study the proof of the CLT using the Fourier transform.

25 Lecture Twenty Five

25.1 Recap: Last Class

In the last lecture, we looked at the Central Limit Theorem:

Theorem 25.1 (Central Limit Theorem)**.** *Suppose $X_i, i = 1, 2, \ldots$ are i.i.d with $\mathbb{E}(X_i) = \mu$ and $var(X_i) = \sigma^2 < \infty$. Then, with $\bar{X}_n = (X_1 + \cdots + X_n)/n$,*

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

converges in distribution to $N(0,1)$. Convergence in distribution here means that

$$\mathbb{P}\{a \leq \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma} \leq b\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \qquad \text{for all } -\infty \leq a < b \leq +\infty.$$

We also discussed the usefulness of transforms for proving the CLT, and proved it using the Laplace Transform (MGF) by showing that

$$M_{Y_n}(t) \to M_{N(0,1)}(t) = e^{t^2/2} \qquad \text{where } Y_n := \frac{\sqrt{n}\left(\bar{X}_n - \mu\right)}{\sigma}$$

This proof suffers from the following two deficiencies:

1. It tacitly assumes that the moment generating function of $X_1, \ldots, X_n$ exists for all t. This is much stronger than the existence of the variance of X_i (which is all the CLT needs). Indeed if $M_X(t)$ exists for all t in any open interval containing zero, then moments of all orders (not just the variance) exist.

2. We have proved that the Laplace transform of $\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma}$ converges to that of $N(0,1)$. It is not clear though as to how this implies that

$$\mathbb{P}\left\{a \leq \sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \leq b\right\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx$$

   for $-\infty \leq a < b \leq \infty$.

To fix these two deficiencies, the Fourier transform (or Characteristic Function) is used to prove the CLT.

25.2 CLT proof via the Fourier Transform

Recall, from the last lecture, that the Fourier transform of a random variable X is defined as the function:

$$\varphi_X(t) := \mathbb{E}\left(e^{itX}\right) = \mathbb{E}\cos(tX) + i\mathbb{E}\sin(tX)$$

for all $t \in (-\infty, \infty)$. Here $i = \sqrt{-1}$. The Fourier transform is defined for every random variable and it is finite for all $t \in (-\infty, \infty)$. This is because $\cos(tX)$ and $\sin(tX)$ are always bounded by 1 so the expectation will obviously be finite. Just like the Laplace transform, the Fourier transform also factorizes for independent random variables. Specifically, if $X_1, \ldots, X_n$ are independent, then

$$\varphi_{X_1+\cdots+X_n}(t) = \varphi_{X_1}(t)\varphi_{X_2}(t)\ldots\varphi_{X_n}(t).$$

As a consequence, if $X_1, \ldots, X_n$ are i.i.d,

$$\varphi_{X_1+\cdots+X_n}(t) = \left(\varphi_{X_1}(t)\right)^n.$$

A sketch of the proof of the CLT via the Fourier transform is provided below. I will skip some technical details and provide only the high level ideas. Full details can be found, for example, in Chapter 6 of the book *A Course in Probability Theory* by Kai Lai Chung.

Fourier Proof of CLT. We have i.i.d random variables $X_1, X_2, \ldots$ which have mean μ and finite variance σ^2. Let $Y_n := \sqrt{n}(\bar{X}_n - \mu)/\sigma$. We shall first prove that the Fourier transform of Y_n converges to that of $N(0,1)$:

$$\varphi_{Y_n}(t) \to \varphi_{N(0,1)}(t) \qquad \text{for every } t.$$

The Fourier transform of $N(0,1)$ equals

$$\varphi_{N(0,1)}(t) = \int_{-\infty}^{\infty} e^{itx}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx = e^{-t^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(x-it)^2\right)dx = e^{-t^2/2}.$$

We will show that

$$\varphi_{Y_n}(t) \to e^{-t^2/2} \qquad \text{for every } t.$$

Because

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

we can write,

$$\begin{aligned}\varphi_{Y_n}(t) &= \varphi_{\sum_i (X_i-\mu)/(\sqrt{n}\sigma)}(t)\\ &= \varphi_{\sum_i (X_i-\mu)/\sigma}(tn^{-1/2}) = \prod_{i=1}^{n}\varphi_{(X_i-\mu)/\sigma}(tn^{-1/2}) = \left(\varphi(tn^{-1/2})\right)^n\end{aligned}$$

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

$$\varphi(tn^{-1/2}) \approx \varphi(0) + \frac{t}{\sqrt{n}}\varphi'(0) + \frac{t^2}{2n}\varphi''(0).$$

It is easy to check that, for every random variable R, we have $\varphi_R(0) = \mathbb{E}e^{i(0)R} = 1$. Also

$$\begin{aligned}\varphi_R'(t) &= \frac{d}{dt}\mathbb{E}e^{itR} = \mathbb{E}\frac{de^{itR}}{dt} = \mathbb{E}\left(iRe^{itR}\right)\\ \varphi_R''(t) &= \frac{d}{dt}\varphi_R'(t) = \mathbb{E}\frac{d(iRe^{itR})}{dt} = \mathbb{E}\left((iR)^2e^{itR}\right).\end{aligned}$$

Plugging in $t = 0$, we get

$$\varphi'_R(0) = \mathbb{E}(iR) = i\mathbb{E}R \quad \text{and} \quad \varphi''_R(0) = \mathbb{E}(iR)^2 = i^2\mathbb{E}(R^2) = -\mathbb{E}(R^2).$$

Using $R = (X_1 - \mu)/\sigma$ (which has mean zero and unit variance) and $\varphi = \varphi_R$, we get

$$\varphi(0) = 1 \quad \text{and} \quad \varphi'(0) = 0 \quad \text{and} \quad \varphi''(0) = -1.$$

Therefore

$$\varphi_{Y_n}(t) = \left(\varphi(tn^{-1/2})\right)^n \approx \left(\varphi(0) + \frac{t}{\sqrt{n}}\varphi'(0) + \frac{t^2}{2n}\varphi''(0)\right)^n = \left(1 - \frac{t^2}{2n}\right)^n \to e^{-t^2/2}$$

as $n \to \infty$. This proves that the Fourier transform of Y_n converges to that of $N(0,1)$. From here, we now need to deduce that

$$\mathbb{P}\{a \le Y_n \le b\} \to \mathbb{P}\{a \le N(0,1) \le b\} \tag{121}$$

This statement is equivalent to

$$\mathbb{E}I_{[a,b]}(Y_n) \to \mathbb{E}I_{[a,b]}(N(0,1))$$

where $I_{[a,b]}(x) := I\{a \le x \le b\}$ is the indicator function of the interval $[a,b]$. How does this follow from the convergence of the Fourier transforms:

$$\mathbb{E}e^{itY_n} \to \mathbb{E}e^{it(N(0,1))}.$$

The idea is that Fourier analysis guarantees that the indicator function can be represented as a linear combination of the complex functions e^{itx}. This representation is of the form:

$$I_{[a,b]}(x) = \int_{-\infty}^{\infty} e^{itx} g(t)dt$$

for some function g. From here, one can write

$$\begin{aligned}
\mathbb{E}I_{[a,b]}(Y_n) &= \mathbb{E}\left(\int_{-\infty}^{\infty} e^{itY_n} g(t)dt\right) \\
&= \int_{-\infty}^{\infty} \mathbb{E}\left(e^{itY_n}\right) g(t)dt \\
&= \int_{-\infty}^{\infty} \varphi_{Y_n}(t)g(t)dt \\
&\to \int_{-\infty}^{\infty} \varphi_{N(0,1)}(t)g(t)dt \\
&= \int_{-\infty}^{\infty} \mathbb{E}\left(e^{itN(0,1)}\right) g(t)dt = \mathbb{E}\left(\int_{-\infty}^{\infty} e^{itN(0,1)} g(t)dt\right) = \mathbb{E}I_{[a,b]}\left(N(0,1)\right)
\end{aligned}$$

which proves (121). For a rigorous version of this argument, see Chapter 6 of the book *A Course in Probability Theory* by Kai Lai Chung. □

25.3 Closing Thoughts

In this class, we took the view that probability is a general and principled method of reasoning under uncertainty. Here is a quote by Jaynes (from one of his papers in 1957): *the purpose of any application of probability theory is simply to help us in forming reasonable judgements*

in situations where we do not have complete information. Hopefully, this course convinced you that probability theory applies to many problems that are commonly studied in the fields of statistics and machine learning (what is usually called "Bayesian Statistics" is just Probability Theory). I would like to leave you with a couple of quotes by Laplace (from the book *A Philosophical Essay on Probabilities* by Pierre Simon Laplace) glorifying probability theory. Laplace was one of the founders of probability theory and Bayesian statistics:

1. *It is remarkable that a science, which commenced with a consideration of the games of chance, should be elevated to the rank of the most important subjects of human knowledge.*

2. *If we consider*

   a) the analytical methods to which this theory has given birth;

   b) the truth of the principles which serve as a basis;

   c) the fine and delicate logic which their employment in the solution of problems requires;

   d) the establishments of public utility which rest upon it;

   e) the extension which it has received and which it can still receive by its application to the most important questions of natural philosophy and the moral science;

   if we consider again

   a) that, even in the things which cannot be submitted to calculus, it gives the surest hints which can guide us in our judgements, and

   b) that it teaches us to avoid the illusions which offtimes confuse us,

   then we shall see that there is no science more worthy of our meditations, and that no more useful one could be incorporated in the system of public instruction.

References

[1] Jaynes, E. T, (2003) *Probability theory: the logic of science.* Cambridge University Press, 2003.

[2] Sinai, Y. G, (1992) *Probability theory: an introductory course.* Springer Verlag, 1992.

[3] DeGroot, M. H. (1986) A conversation with David Blackwell. *Statistical Science*, **1**, 40–53.

[4] Mosteller, F, (1987) *Fifty challenging problems in probability with solutions.* Courier Corporation, 1987.

[5] MacKay, D. J. C., (2003) *Information theory, inference, and learning algorithms.* Cambridge University Press, 2003.