EECS 126 Notes

Matthew Lacayo

April 11, 2021

These notes will follow along with the teachings of the book Introduction to Probability, Statistics, and Random Processes by Hossein Pishro-Nik. Additional notes will also be included that follow along with the lecture teaching of professor Tom Courtade from Spring 2021.

Contents

1 Multiple Random Variables

1.1 Joint Distributions and Independence

Definition 1.1: Joint PMF

Let $X_1, X_2, \cdots X_n$ be n discrete random variables. The joint pmf of v is defined as:

$$P_{X_1,X_2,\cdots X_n}(X_1, X_2, \cdots X_n) = P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n)$$

For continuous random variables, the joinf PDF can be defined similarly using integrals. The marginal PDF can be reconstructed by integrating out all other random variables from the joint PDF.

Definition 1.2: Joint CDF

The joint CDF of n random variables $X_1, X_2, \cdots X_n$ is defined as:

$$F_{X_1,X_2,\cdots X_n}(X_1, X_2, \cdots X_n) = F(X_1 \leq x_1, X_2 \leq x_2, \cdots, X_n \leq x_n)$$

Earlier results regarding independence of 2 random variables still hold for n random variables. Namely, if a group of n random variables are independent, then their joint PDF is the same as the product of the marginal PDFs, the joint CDF is the product of the marginal CDFs, the joint PMF is the product of the marginal PMFs, and the expected value of the product of the random variables is the same as the product of the expected values. Moreover, the results regarding sums of independent random variables still hold, and so we can still make use of the equations we developed for expected value and variance.

1.2 Moment Generating Functions

Definition 1.3: Moments

We define the nth moment of a random variable X to be $E[X^n]$.

Definition 1.4: Moment Generating Function (MGF)

The moment generating function of a random variable X is defined to be the function $M_X(s)$ where:

$$M_X(s) := E[e^{sX}]$$

The MGF of X exists if there exists a positive constant a such that $M_X(s)$ is finite for all $s \in [-a, a]$.

The MGF is useful because it allows us to extract all the moments of X simply by taking derivatives (follows by considering the taylor series expansion of $e^s X$). Namely,

$$E[X^k] = \frac{d^k}{ds^k} M_x(s)|_{s=0}$$

The MGF also uniquely defines the distribution of X. I.e. if two random variables X, Y have the same MGF, then they also have the same distribution.

Theorem 1.1

If $X_1, X_2, \cdots, X_n$ are n independent random variables, then:

$$M_{X_1,X_2,\cdots,X_n}(s) = M_{X_1}(s)M_{X_2}(s)\cdots M_{X_n}(s)$$

Example 1.1

Let $X \sim Binomial(n,p)$. Find the MGF of X.
Since X is binomial, we can think of it as the sum of n independent Bernoulli random variables:

$$X = X_1 + X_2 + \cdots + X_n$$

Since the X_i are i.i.d., we know that $M_X(s) = M_{X_1}(s)M_{X_2}(s)\cdots M_{X_n}(s)$.

$$M_{X_i}(s) = E[e^{sX_i}] = p(e^s) + (1-p)(e^0) = (1-p) + pe^s$$

Thus,

$$M_X(s) = (1 - p + pe^s)^n$$

Example 1.2

Use MGFs to prove that if $X \sim Binomial(m,p)$ and $Y \sim Binomial(n,p)$ are independent, then $X + Y \sim Binomial(m+n,p)$.

Solution: The MGF of $X+Y$ is:

$$M_X(s)M_Y(s) = (1-p+pe^s)^m \cdot (1-p+pe^s)^n = (1-p+pe^s)^{m+n}$$

Which is the MGF of $Z \sim Binomial(m+n,p)$.

1.3 Characteristic Functions

To be filled in.

1.4 Exercises Part 1

Example 1.3

Let X, Y, Z be three independent random variables with $X \sim N(\mu, \sigma^2)$, and $Y, Z \sim Uniform(0,2)$. We also know that:

- $E[X^2Y + XYZ] = 13$
- $E[XY^2 + ZX^2] = 14$

Find μ and σ.

Solution: From equation 1, we get that $E[X^2] = 13 - \mu$. From equation 2, we get that $E[Y^2] = \frac{14 - E[X^2]}{\mu} = \frac{1+\mu}{\mu}$. We also know that $Var(Y) = Var(Z) = \frac{1}{3}$. Thus, $E[Y^2] = 1 + \frac{1}{3} = \frac{4}{3}$. This gives $\mu = 3$. Finally, $Var(X) = \sigma^2 = 13 - \mu - \mu^2 = 1$.

Example 1.4

Let X_1, X_2, X_3 be 3 i.i.d. $Bernoulli(p)$ random variables and:

- $Y_1 = \max(X_1, X_2)$
- $Y_2 = \max(X_1, X_3)$
- $Y_3 = \max(X_2, X_3)$
- $Y = Y_1 + Y_2 + Y_3$

Find $E[Y]$.

Solution: $E[Y] = 3E[Y_1]$. $E[Y_1] = (1 - (1-p)^2) = p(2-p)$. Thus, $E[Y] = 3p(2-p)$.

1.5 Markov and Chebyshev Inequalities

Theorem 1.2: Markov's Inequality

If X is a nonnegative random variable, then for any $a > 0$:

$$P(X \geq a) \leq \frac{E[X]}{a}$$

Theorem 1.3: Chebyshev's Inequality

If X is any random variable, then for any $b > 0$ we have:

$$P(|X - E[X]| \geq b) \leq \frac{Var(X)}{b^2}$$

Theorem 1.4

If X is a random variable, then for any $a \in \mathbb{R}$ we have:

- $P(X \geq a) \leq e^{-sa} M_X(s)$, for all $s > 0$
- $P(X \leq a) \leq e^{-sa} M_X(s)$, for all $s < 0$

Proof. We start by seeing that:

- $P(X \geq a) = P(e^{sX} \geq e^{sa})$, for all $s > 0$
- $P(X \leq a) = P(e^{sX} \geq e^{sa})$, for all $s < 0$

For $s > 0$, we can use Markov's inequality to write:

$$P(e^{sX} \geq e^{sa}) \leq \frac{E[e^{sX}]}{e^{sa}} = e^{-sa} M_X(s)$$

A similar result follows for $s < 0$.

□

This bound is useful because it depends on a new parameter: s. That is to say, we can optimize over s to find the tightest bound.

Theorem 1.5: Cauchy-Schwarz Inequality

For any two random variables X and Y, we have:

$$|E[XY]| \leq \sqrt{E[X^2]E[Y^2]}$$

Where equality holds if and only if $X = \alpha Y$ for some constant $\alpha \in \mathbb{R}$.

2 Limit Theorems and Convergence of Random Variables

2.1 Limit Theorems

Definition 2.1: Sample Mean

For i.i.d. random variables $X_1, X_2, \cdots, X_n$, the **sample mean**, denoted by $\overline{X}$ or M_n, is defined as:

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Note that since each X_i is a random variable, the sample mean is also a random variable. It can easily be shown that:

- $E[\overline{X}] = E[X_i]$
- $Var(\overline{X}) = \frac{Var(X)}{n}$

Theorem 2.1: Weak Law of Large Numbers (WLLN)

Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables with a finite expected value $E[X_i] = \mu < \infty$. Then, for any $\epsilon > 0$ we have:

$$\lim_{n\to\infty} P(|\overline{X} - \mu| \geq \epsilon) = 0$$

We can easily prove this if we make the assumption that the variance is finite, as we can then use chebyshevs inequality and show that the term on the right of the inequality goes to zero as n goes to infinity.

Definition 2.2: Normalized Variable

Suppose $X_1, X_2, \cdots, X_n$ are i.i.d. random variables with $E[X_i] = \mu < \infty$ and $Var(X_i) = \sigma^2 < \infty$. We define the normalized random variable Z_n as:

$$Z_n = \frac{\overline{X} - E[\overline{X}]}{\frac{\sigma}{\sqrt{n}}} = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma}$$

Also, note that $E[Z_n] = 0$ and $Var(Z_n) = 1$.

Theorem 2.2: The Central Limit Theorem (CLT)

Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables with $E[X_i] = \mu < \infty$ and $Var(X_i) = \sigma^2 < \infty$. Then, Z_n converges in distribution to the standard normal random variable as n goes to infinity. That is, for all $x \in \mathbb{R}$:

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x)$$

Where $\Phi(x)$ is the standard normal CDF.

Example 2.1

In a communication system, each data packet consists of 1000 bits. Due to the noise, each bit may be received in error with probability 0.1. It is assumed bit errors occur independently. Find the probability that there are more than 120 errors in a certain data packet.

Solution: Let $Y = X_1 + X_2 + \cdots + X_1 000$ where each $X_i \sim Bernoulli(0.1)$. Thus, $E[X_i] = 0.1$ and $Var(X_i) = 0.09$. We want to find $P(Y \leq 120)$.

$$P(Y \leq 120) = P(\frac{Y - n\mu}{\sigma\sqrt{n}} \leq \frac{120 - \mu}{\sigma\sqrt{n}}) = P(Z_n \leq \frac{120 - 100}{\sqrt{0.09}\sqrt{1000}}) \approx \Phi(\frac{20}{\sqrt{90}})$$

Finally, $1 - \Phi(\frac{20}{\sqrt{90}}) = 0.0175$

2.2 Convergence of Random Variables

Definition 2.3: Convergence in Distribution

A sequence of random variables $X_1, X_2, X_3, \cdots$ converges **in distribution** to a random variable X, denoted as $X_n \xrightarrow{d} X$, if:

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

for all x at which $F_X(x)$ is continuous.

Example 2.2

Let $X_2, X_3, X_4, \cdots$ be a sequence of random variables such that:

$$F_{X_n}(x) = \begin{cases} 1 - (1 - \frac{1}{n})^{nx}, & \text{for } x > 0 \\ 0, & \text{otherwise} \end{cases} \Bigg\}$$

Show that $X_n \xrightarrow{d} Expo(1)$.

Solution:

$$\lim_{n\to\infty} 1 - (1 - \frac{1}{n})^{nx} = 1 - e^{-x}$$

This is the CDF of $Expo(1)$ as desired.

Definition 2.4: Convergence in Probability

A sequence of random variables $X_1, X_2, X_3, \cdots$ converges **in probability** to a random variable X, denoted as $X_n \xrightarrow{p} X$, if for all $\epsilon > 0$:

$$\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

Example 2.3

Let $X_n \sim Expo(n)$. Show that $X_n \xrightarrow{p} 0$ i.e. to the random variable that is always 0.

Solution:

$$\lim_{n\to\infty} P(|X_n| > \epsilon) = |e^{-n\epsilon}| = 0$$

as desired.

Example 2.4

Reread the WLLN. This is an example of convergence in probability.

Theorem 2.3: Almost Sure Convergence

A sequence of random variables $X_1, X_2, X_3, \cdots$ converges **almost surely** to a random variable X, denoted as $X_n \xrightarrow{a.s.} X$, if:

$$P(\{s \in S : \lim_{n\to\infty} X_n(s) = X(s)\}) = 1$$

This is saying that the set of samples for which $X_n = X$ forms an event of probability 1 as n goes to infinity.

Theorem 2.4: Strong Law of Large Numbers

Let $X_1, X_2, \cdots, X_n$ be i.i.d. random variables with $E[X_i] = \mu < \infty$. Let also:

$$M_n := \frac{X_1 + X_2 + \cdots + X_n}{n}$$

Then $M_n \xrightarrow{a.s.} \mu$.

3 Random Processes

3.1 Poisson Processes

Definition 3.1: Counting Process

A random process $\{N(t), t \in [0, \infty)\}$ is said to be a **counting process** if $N(t)$ is the number of events occurred from time 0 up to and including time t. For a counting process, we assume:

- $N(0) = 0$
- $N(t) \in \{0, 1, 2, \cdots\}$, for all $t \in [0, \infty)$
- For $0 \leq s < t$, $N(t) - N(s)$ shows the number of events that occur in the interval $(s, t]$

We typically refer to the occurrence of an event as an 'arrival'.

Definition 3.2: Poisson Process

Let $\lambda > 0$ be fixed. The counting process $\{N(t), t \in [0, \infty)\}$ is called a **Poisson Process** with rate λ if all the following conditions hold:

- $N(0) = 0$
- $N(t)$ has independent increments (disjoint intervals are independent)
- The number of arrivals in any interval of length $\tau > 0$ has $Poisson(\lambda\tau)$ distribution.

Note that the number of arrivals in any given interval depends only on the length of the interval, not on the location of the interval.

Example 3.1

The number of customers arriving at a grocery store can be modeled by a Poisson process with intensity $\lambda = 10$ customers per hour. Find the probability that there are 2 customers between 10:00 and 10:20.

Solution: We know that the length of the interval $\tau = \frac{20}{60} = \frac{1}{3}$. Thus, the distribution of arrivals in the 20 minute interval is $X \sim Poisson(10 \cdot \frac{1}{3} = \frac{10}{3})$. Finally,

$$P(X = 2) = \frac{e^{-\frac{10}{3}}(\frac{10}{3})^2}{2!} \approx 0.2$$

Now, let us consider arrival and interarrival times for a Poisson process.

Example 3.2

Let $N(t)$ be a Poisson process with rate λ. Let X_1 be the time of the first arrival. Then,

$$\begin{aligned} P(X_1 > t) &= P(\text{no arrival in } (0,t]) \\ &= \frac{(\lambda t)^0}{0!} e^{-(\lambda t)} \\ &= e^{-\lambda t} \end{aligned}$$

Thus,

$$F_{X_1}(t) = 1 - e^{-\lambda t}$$

This is the CDF of Exponential (λ), and so $X_1 \sim$ Exponential (λ).

Let X_2 be the time elapsed between the first and the second arrival. Since the probability of an arrival in disjoint intervals is independent, we also have that $X_2 \sim$ Exponential (λ). This yields the following definition.

Definition 3.3: Interarrival Times for Poisson Processes

If $N(t)$ is a Poisson process with rate λ, then the interarrival times $X_1, X_2, \cdots$ are independent and for $i = 1, 2, 3, \cdots$:

$$X_i \sim \text{Exponential}\,(\lambda)$$

Definition 3.4: Arrival Times for Poisson Processes

If $N(t)$ is a Poisson process with rate λ, then the arrival times $T_1, T_2, \cdots$ have $Erlang(n, \lambda)$ distribution. I.e.

$$T_i = X_1 + X_2 + \cdots + X_i$$

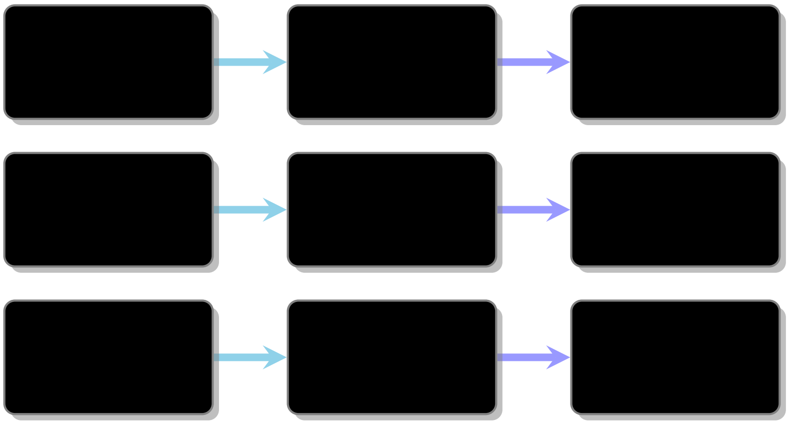
Thus,

- $E[T_i] = \frac{n}{\lambda}$
- $Var(T_i) = \frac{n}{\lambda^2}$

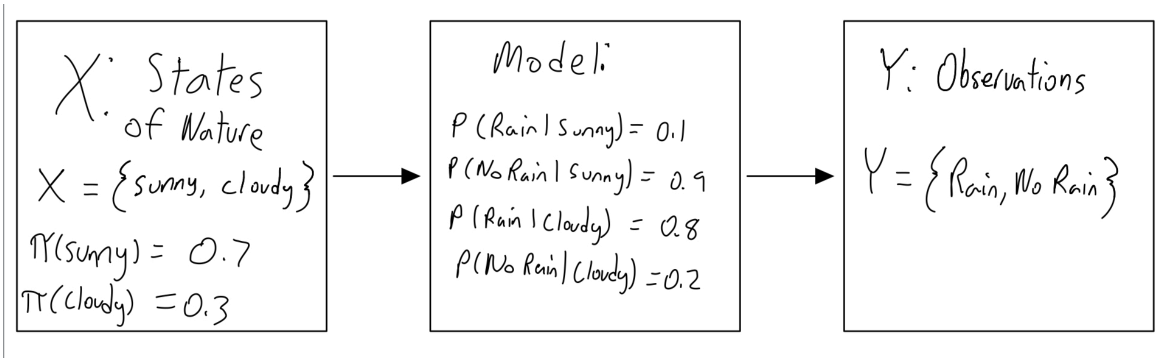
Now, we will discuss merging and splitting Poisson processes. That is, we will consider taking two independent Poisson processes and adding them together, or taking one poisson process and splitting its arrivals into multiple 'streams' according to some defined distribution.

4 Hypothesis Testing

We now turn our attention to the methods to best utilize models, as well as how we can make models.



The goal of a "model" is to define the transition probabilities between some state of nature (or hypothesis), and some observation that we see potentially resulting from the underlying state of nature. For example, given that it is cloudy (state of nature), what is the probability that it rains (observation)? Note that we will call these transition probabilities **likelihoods**, as they represent the likelihood of seeing some observation given an underlying state of nature. We refer to the probability distribution of the states of nature as the **prior**, and we usually denote this probability distribution with the symbol π. In our example, the prior would be the probability distribution of the different states of the sky. For example, we would have $\pi(\text{cloudy}) = P(\text{cloudy})$, $\pi(\text{sunny}) = P(\text{sunny})$, etc.



Note that our discussion has been using the phrase "state of nature", however it will be helpful to begin thinking of this as a "hypothesis" instead. The justification for this choice will be more clear after reading the coming sections. However, for now, whenever you see the phrase hypothesis, try thinking of it as a "state of nature" instead if that makes more intuitive sense.

4.1 MAP and MLE

Say that we are given some observation, and we would like to know under what conditions would the likelihood of observing this outcome be maximized. In other words, after seeing some evidence, we now want to know which hypothesis we believe is the most likely candidate for producing the evidence. This could be helpful if we are not able to observe the inner working of a system and are only able to observe some output of that system, but we still want to be able to make some predictions about what is going on behind the scenes. Let us now formalize our goal. Assume that the possible hypotheses lie in a set X, and the possible observations lie in a set Y. Given some observation $y \in Y$, we want to find:

$$\arg\max_{x \in X} p_{x|y}(x|y)$$

How can we interpret these probabilities? We know that $\pi(x)$ is the *prior* of x, and it represents the probability of the hypothesis being x (no evidence observed). We can also interpret $p_{x|y}(x|y)$ as being our updated belief about the probability of the hypothesis being x now that we have observed an observation y. In other words, observing y gave us new information, and so we now want to update our beliefs on the hypotheses based on this information. Using this way of thinking is critical to understanding the utility of Bayes Law. Remember that the model gives us access to the transition probabilities between the hypotheses and the possible observations: $p_{y|x}(y|x)$, and we know that we can relate this to $p_{x|y}(x|y)$ using Bayes Law, and so we can can convert our goal into the following:

$$\arg\max_{x \in X} p_{x|y}(x|y) = \arg\max_{x \in X} \frac{p_{y|x}(y|x)p(x)}{p(y)} = \arg\max_{x \in X} \frac{p_{y|x}(y|x)\pi(x)}{p(y)}$$

Note that since our goal is to find $x \in X$ that maximizes the expression, we can ignore the denominator. This is because the denominator depends only on y, and so maximizing the entire expression amounts to maximizing what we can control, which in this case is just the numerator. Therefore, to find our guess for which hypothesis caused the evidence that we saw, we define:

Definition 4.1: MAP Estimate

$$\hat{X}_{MAP}(y) := \underbrace{\arg\max_{x \in X} p_{x|y}(x|y)}_{\text{"A Posteriori Probability"}} = \underbrace{\arg\max_{x \in X} p_{y|x}(y|x)\pi(x)}_{\text{Computable with model}}$$

This is called the **MAP Estimate**, and it is a function of the observation. MAP stands for **Maximum A Posteriori Probability**. A Posteriori Probability refers to the "after probability", meaning the updated probability distribution of the hypothesis given we observe some piece of evidence. The above equation shows how we can think about MAP. Our goal is to maximize the A Posteriori Probability, and we reduced this problem into one that we can solve by utilizing the model.

Example 4.1: Computing MAP

Assume that we have a coin that flips heads with probability 1/4. Your friend (or imaginary friend) flips the coin behind your back and tells you the result. Unfortunately, your friend is very nervous (perhaps they have a 5 page essay due in 2 hours yet they chose not to begin until the last minute because they were too busy watching youtube videos all day) and occasionally tells you the opposite of the true result with probability 2/5. Let y represent the result that your friend tells you such that $y = 0$ represents tails and $y = 1$ represents heads. Compute $\hat{X}_{MAP}(y = 0)$ and $\hat{X}_{MAP}(y = 1)$.

We will begin with $\hat{X}_{MAP}(y = 0)$. We want to compute $\arg\max_{x \in X} p_{y|x}(y|x)\pi(x)$. We have:

$$\hat{X}_{MAP}(y = 0) =$$

$$\arg\max_{x\in X} \begin{cases} p_{y|x}(hear\ tails|true\ tails)\pi(true\ tails) = 3/5(3/4) = 9/20 : x = true\ tails \\ p_{y|x}(hear\ tails|true\ heads)\pi(true\ heads) = 2/5(1/4) = 2/20 : x = true\ heads \end{cases}$$

$$= true\ tails$$

Similarly,

$$\hat{X}_{MAP}(y = 1) =$$

$$\arg\max_{x\in X} \begin{cases} p_{y|x}(hear\ heads|true\ tails)\pi(true\ tails) = 2/5(3/4) = 6/20 : x = true\ tails \\ p_{y|x}(hear\ heads|true\ heads)\pi(true\ heads) = 3/5(1/4) = 3/20 : x = true\ heads \end{cases}$$

$$= true\ tails$$

This means that no matter what our nervous friend tells us, we will guess that the hypothesis (the true result of the coin flip) was actually tails. This makes some intuitive sense, since we know that our friend gives inaccurate information almost half the time, and so our best bet is to just guess the outcome that was more likely to occur. If we instead assume that our friend gives the wrong result with probability 1/5 instead, then we will get that the best estimate for the hypothesis is to believe what our friend says. This also makes intuitive sense, as decreasing the probability of hearing an incorrect result makes the evidence more valuable. This leads us to wonder if perhaps there is some threshold for this lying probability where we should believe our friend vs disregarding their evidence.

We call MAP a type of **estimator**. In particular, MAP is useful when we know the prior (the probability distribution of the hypotheses). However, what if we do not have access to this distribution? One possible solution is to assume that the prior is uniform among (us) all hypotheses: i.e. $\pi(x) = 1/|X|$. If we try to now compute the MAP estimate using this assumption, we will be able to treat $\pi(x)$ as a constant (since it is the same for any x), and so we will be able to pull it out of the argmax expression entirely. We will get a different type of estimate called the **MLE** or **Maximum Likelihood Estimate**:

Definition 4.2: MLE

$$\hat{X}_{MLE}(y) := \arg\max_{x \in X} p_{y|x}(y|x)$$

This has a very nice interpretation: we will estimate the hypothesis by choosing the hypothesis that has greatest probability of producing the observed outcome, regardless of the prior distribution of the hypotheses. Note that the term we noted earlier, *likelihood*, is now coming into play. With MLE, we want to maximize the likelihood (transition probability between hypothesis and evidence), which is the same thing as our "model".

Example 4.2: Computing MLE

Using the same example as we did for MAP, compute the MLE.

We can ignore the prior distribution, and just consider our model. In this problem, our model is the conditional probability distribution of whether or not our friend tells us the correct result. Since our friend lies with probability $2/5$, they are more likely to be telling the truth. Thus, our MLE is to believe what our friend says since $3/5 > 2/5$. If our friend lied with probability $3/5$, then our MLE would become believing the opposite of what our friend tells us.

At the end of the first example, an idea of potentially finding thresholds for probabilities was mentioned. This will be a very useful technique in the coming section, so we will walk through examples of this. First, it will be useful to define a **likelihood ratio**. The purpose of this will be made more clear in the coming example.

Definition 4.3: Likelihood Ratio

We define the likelihood ratio in situations where we are considering two possible hypotheses (a binary situation). We can think of this as meaning $X = \{0, 1\}$. We define:

$$L(y) := \frac{p_{y|x}(y|x=1)}{p_{y|x}(y|x=0)}$$

The choice of having the likelihood for $x = 1$ in the numerator instead of $x = 0$ is just by convention.

Example 4.3: Thresholding Intro: MLE

Consider a situation our friend flips a coin (potentially biased) and reads us the result. If the result is heads, they lie about it with probability p. If the result is tails, they lie about it with probability q. We seek to compute the MLE as a function of p and q.

We have:

$$\hat{X}_{MLE}(y=0) = \\ \arg\max_{x \in X} \begin{cases} p_{y|x}(hear\ tails|true\ tails) = 1-p : x = true\ tails \\ p_{y|x}(hear\ tails|true\ heads) = p : x = true\ heads \end{cases}$$

And:

$$\hat{X}_{MLE}(y=1) = \\ \arg\max_{x \in X} \begin{cases} p_{y|x}(hear\ heads|true\ tails) = q : x = true\ tails \\ p_{y|x}(hear\ heads|true\ heads) = 1-q : x = true\ heads \end{cases}$$

This gives us:

$$\hat{X}_{MLE}(y=0) = \begin{cases} true\ tails : p \leq 1/2 \\ true\ heads : p > 1/2 \end{cases}$$

And:

$$\hat{X}_{MLE}(y=1) = \begin{cases} true\ tails : q \geq 1/2 \\ true\ heads : q < 1/2 \end{cases}$$

We can also write this in terms of the likelihood ratio by noticing that $L(y=0) = p/(1-p)$, and similarly $L(y=1) = (1-q)/q$. This gives us the equivalent formulation:

$$\hat{X}_{MLE}(y=0) = \begin{cases} true\ tails : L(y=0) \leq 1 \\ true\ heads : L(y=0) > 1 \end{cases}$$

$$\hat{X}_{MLE}(y=1) = \begin{cases} true\ tails : L(y=1) \leq 1 \\ true\ heads : L(y=1) > 1 \end{cases}$$

And finally, we can reduce this to:

$$\hat{X}_{MLE}(y) = \begin{cases} 0 : L(y) \leq 1 \\ 1 : L(y) > 1 \end{cases}$$

If this formulation is not immediately clear, try plugging in some values for p and q to convince yourself that these are two equivalent formulations.

Example 4.4: Thresholding Intro: MAP

Consider a situation our friend flips a coin (potentially biased) and reads us the result. If the result is heads, they lie about it with probability p. If the result is tails, they lie about it with probability q. We seek to compute the MLE as a function of p and q.

Note that since we did not specify the bias of the coin, we will have to give a name to the bias. For this, we notice that the bias of the coin is the prior distribution, and so we will denote the bias by $\pi(0)$ and $\pi(1)$. Since this is Bernoulli, we have that $\pi(1) = 1 - \pi(0)$. Note that $x = 0$ corresponds to the case *true tails*, and $x = 1$ corresponds to the case *true heads*.

$$\hat{X}_{MAP}(y=0) =$$

$$\arg\max_{x \in X} \begin{cases} p_{y|x}(\textit{hear tails}|\textit{true tails})\pi(\textit{true tails}) = (1-p)\cdot\pi(0) : x = \textit{true tails} \\ p_{y|x}(\textit{hear tails}|\textit{true heads})\pi(\textit{true heads}) = p\cdot\pi(1) : x = \textit{true heads} \end{cases}$$

Similarly,

$$\hat{X}_{MAP}(y=1) =$$

$$\arg\max_{x \in X} \begin{cases} p_{y|x}(\textit{hear heads}|\textit{true tails})\pi(\textit{true tails}) = q\cdot\pi(0) : x = \textit{true tails} \\ p_{y|x}(\textit{hear heads}|\textit{true heads})\pi(\textit{true heads}) = (1-q)\cdot\pi(1) : x = \textit{true heads} \end{cases}$$

Thus, we have:

$$\hat{X}_{MAP}(y=0) = \begin{cases} \textit{true tails} : (1-p)\cdot\pi(0) \geq p\cdot\pi(1) \\ \textit{true heads} : (1-p)\cdot\pi(0) < p\cdot\pi(1) \end{cases}$$

And:

$$\hat{X}_{MAP}(y=1) = \begin{cases} \textit{true tails} : q\cdot\pi(0) \geq (1-q)\cdot\pi(1) \\ \textit{true heads} : q\cdot\pi(0) < (1-q)\cdot\pi(1) \end{cases}$$

Rearranging the inequalities so that we can rewrite in terms of $L(y=0)$ and $L(y=1)$ gives:

$$\hat{X}_{MAP}(y=0) = \begin{cases} \textit{true tails} : L(y=0) \leq \pi(0)/\pi(1) \\ \textit{true heads} : L(y=0) > \pi(0)/\pi(1) \end{cases}$$

$$\hat{X}_{MAP}(y=1) = \begin{cases} \textit{true tails} : L(y=1) \leq \pi(0)/\pi(1) \\ \textit{true heads} : L(y=1) > \pi(0)/\pi(1) \end{cases}$$

This can finally be reduced to:

$$\hat{X}_{MAP}(y) = \begin{cases} 0 : L(y) \leq \pi(0)/\pi(1) \\ 1 : L(y) > \pi(0)/\pi(1) \end{cases}$$

As we can see from the last two examples, for a given problem we are able to find the thresholds for which our estimates change based entirely on the likelihood ratio. We will continue to make use of the likelihood ratio in the next section.

4.2 Binary Hypothesis Testing

Binary hypothesis testing handles situations where $|X| = 2$, which means we can equivalently view our set of hypotheses as being $X = \{0, 1\}$. By convention, we say that $x = 0$ is the **null hypothesis**, and $x = 1$ is the **alternate hypothesis**. Our goal is to create a "decision rule" or "test" that tells us which hypothesis we should guess if we are given an observation. In other words, a decision rule is a function $\hat{X} : Y \to \{0, 1\}$ (a function that maps an observation to either 0 or 1).

Definition 4.4: Decision Rule

If $X = \{0, 1\}$ is our set of hypotheses, and Y is our set of evidence, then a function $\hat{X} : Y \to \{0, 1\}$ is called a **decision rule**.

With any decision rule, there are two possible errors that can occur. We define them below:

Definition 4.5: Type I and Type II Errors

Type I Error = False Positive: $P(\hat{X} = 1|x = 0)$ (Probability our decision rule outputs 1, but the true hypothesis is 0)

Type II Error = False Negative: $P(\hat{X} = 0|x = 1)$ (Probability our decision rule outputs 0, but the true hypothesis is 1)

An optimal test would be able to minimize both of these errors, however minimizing two quantities simultaneously is difficult to quantify (how do we weight the value of each error type? What if minimizing type I is more important than minimizing type II? etc.). Thus, to search for an "optimal" test we will rephrase our optimization objective. We will first pick some $\beta \geq 0$ and say that we will minimize the rate of type II errors subject to the constraint that the rate of type I errors is less than or equal to β. In other words, once we have picked some constant $\beta \geq 0$, we limit our space of all possible decision rules to a smaller space of all decision rules that achieve a type I error rate that is smaller than or equal to β. Of this smaller set of decision rules, we wish to find the one that minimizes the rate of type II errors. Mathematically, let S denote the set of all possible decision rules. Let S_β be the set of decision rules that achieve a type I error rate less than or equal to beta, i.e.: $S_\beta = \{\hat{X} \in S : \underbrace{P(\hat{X} = 1|x = 0)}_{Type\ I\ Error\ Rate} \leq \beta\}$. Then, we define our optimal test:

Definition 4.6: Optimal Decision Rule

$$\hat{X}^* = \arg\min_{\hat{X} \in S_\beta} \underbrace{P(\hat{X} = 0|x = 1)}_{Type\ II\ Error\ Rate}$$

The question remains if there is a way to more easily find this optimal decision rule in S_β, as it doesn't seem like a good time to consider every single possible decision rule. The following result gives us a very elegant characterization:

Theorem 4.1: Neyman-Pearson Lemma

Given $\beta \in [0, 1]$, the optimal decision rule is a (randomized) threshold test:

$$\hat{X}_\beta(y) = \begin{cases} 1 \text{ if } L(y) > \lambda \\ 0 \text{ if } L(y) < \lambda \\ \text{Bernoulli}\,(\gamma) \ \text{ if } L(y) = \lambda \end{cases}$$

Where λ, γ are chosen so that:

$$\underbrace{P(\hat{X}_\beta(y) = 1 | x = 0)}_{Type\ I\ Error\ Rate} = \beta$$

$\hat{X}_\beta(y)$ is called the “Neyman-Pearson Rule”, and it is the most “powerful” test (i.e. the test that minimized the type II error) subject to the constraint that $P(\text{Type I Error}) \leq \beta$