Contents

1 Elementary Probability **2**

1.1 Lecture 1 . . . 2

1.2 Lecture 2 . . . 3

1.2.1 Probability Fundamentals and Random Variables . . . 3

1.3 Lecture 3 . . . 5

1.3.1 Concentration Inequalities . . . 5

1.3.2 Covariance and Estimation . . . 6

2 Basic Probability **6**

2.1 Lecture 3, Continued . . . 6

2.1.1 Infinite Collections of Events and Borel-Cantelli . . . 6

2.2 Lecture 4 . . . 9

2.2.1 The Laws of Large Numbers, Revisited . . . 9

2.2.2 Independence . . . 9

2.2.3 Conditional Probability . . . 10

2.3 Lecture 5 . . . 13

2.3.1 Common Discrete Distributions . . . 13

2.3.2 Multiple Discrete Random Variables . . . 14

2.4 Lecture 6 . . . 16

2.5 Lecture 7 . . . 17

2.5.1 Multiple Continuous Random Variables . . . 17

3 Page Rank **18**

3.1 Lecture 7, Continued . . . 18

3.1.1 Discrete-Time Markov Chains . . . 18

3.2 Lecture 8 . . . 20

3.2.1 Hitting Times . . . 20

3.3 Lecture 9 . . . 21

3.3.1 Laws of Large Numbers . . . 21

4 Multiplexing **23**

4.1 Lecture 9, Continued . . . 23

4.2 Lecture 10 . . . 24

4.2.1 Central Limit Theorem . . . 24

4.3 Lecture 11 . . . 26

4.3.1 Characteristic and Moment-Generating Functions . . . 26

4.3.2 Proof of the CLT . . . 28

1 Elementary Probability

1.1 Lecture 1

This lecture I did not have a laptop yet, so I was unable to transcribe anything. Here is what I remember was discussed:

Definition 1.1 (Conditional Probability) For two events A, B, the probability of A given B is:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Definition 1.2 (MAP) We say the Most likely A Posteriori (MAP) estimate of a random variable X given $Y = y$, is:

$$\operatorname*{argmax}_x \mathbb{P}[X = x \mid Y = y]$$

Definition 1.3 (MLE) We say the Maximum Likelihood Estimate (MLE) of a random variable X given Y, is:

$$\operatorname*{argmax}_x \mathbb{P}[Y = y \mid X = x]$$

1.2 Lecture 2

1.2.1 Probability Fundamentals and Random Variables

We begin probability by defining a set Ω called the sample space. Elements of the sample space are termed outcomes. Subsets of Ω are termed as events.

For some event A, we can define the probability of A as follows:

$$\mathbb{P}[A] = \sum_{\omega \in A} \mathbb{P}[\omega]$$

Probability maps events to $[0, 1]$ in a consistent manner, satisfying the following axioms:

- $\mathbb{P}[\Omega] = 1$
- $\mathbb{P}[\emptyset] = 0$
- For two disjoint events A_1, A_2, we have $\mathbb{P}[A_1 \cup A_2] = \mathbb{P}[A_1] + \mathbb{P}[A_2]$

This is what we term a probability space. Often it is more helpful to work with events than with individual sample points (especially in the case of an uncountably infinite amount of sample points).

Definition 1.4 (Random Variable) A random variable $X : \omega \to B$ maps each outcome to elements of some other set (often $\mathbb{R}$). $X = x$ for some x is an event, with a well-defined probability.

Definition 1.5 (Independence) Two random variables X and Y are independent if

$$\mathbb{P}[X = x | Y = x] = \mathbb{P}[X = x]$$

i.e.

$$\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\,\mathbb{P}[Y = y]$$

Example 1.1 Suppose you flip a coin 10 times. We will show that X, the amount of heads in the first 4 flips, and Y, the amount of heads in the last 6 flips, are independent.
Let $a(x)$ be the amount of ways to get x heads in 4 flips and $b(y)$ be the amount of ways to get y heads in 6 flips. Then,

$$\mathbb{P}[X = x] = \frac{a(x) \cdot 2^6}{2^{10}} = \frac{a(x)}{2^4}$$
$$\mathbb{P}[Y = y] = \frac{b(y) \cdot 2^4}{2^{10}} = \frac{b(y)}{2^6}$$
$$\mathbb{P}[X = x, Y = y] = \frac{a(x) \cdot b(y)}{2^{10}} = \mathbb{P}[X = x] \cdot \mathbb{P}[Y = y]$$

Thus, the random variables are independent.

Definition 1.6 (Expectation) The expectation of a (discrete) random variable is:

$$\mathbb{E}[X] = \sum_x x\mathbb{P}[X = x]$$

This is often called the mean or the average value.

Theorem 1.1 Properties of expectation:

- $\mathbb{E}[a] = a$ for $a \in \mathbb{R}$
- If the space is uniform, then $\mathbb{E}[X] = \frac{1}{N}\sum_x x$
- $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$ for $\alpha, \beta \in \mathbb{R}$
- $X \leq Y \Rightarrow \mathbb{E}[X] \leq \mathbb{E}[Y]$
- If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] \not\Rightarrow X, Y$ independent

There are two ways of thus computing expectation. You can either sum over sample points, or take a lot of measurements of your random variable, then divide by the amount of measurements. The reason this works is because of property 2 above.

Definition 1.7 (Variance and Standard Deviation) The variance of a random variable X is defined as:

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X]]^2$$

The standard deviation of this random variable is:

$$\sigma_X = \sqrt{\text{Var}(X)}$$

The variance measures the spread away from the mean that a random variable may exhibit.

Theorem 1.2 Properties of variance:

- $\text{Var}(X) \geq 0$, with equality only if X is constant
- $\text{Var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$
- $\text{Var}(aX) = a^2\text{Var}(X)$ for constant a
- If X and Y are independent, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$
- In general, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$

Often we term $\mathbb{E}\left[X^k\right]$ as the kth moment of X, so the variance contains information about the second moment of X.

1.3 Lecture 3

1.3.1 Concentration Inequalities

Definition 1.8 (Indicator Random Variable) $\mathbb{1}_A$ is the indicator for event A, i.e. a random variable with the following values:

$$\mathbb{1}_A = \begin{cases} 1 & \text{if sample point in event } A \\ 0 & \text{otherwise} \end{cases}$$

Theorem 1.3 (Markov's Inequality) Consider random variable $X \geq 0$ and constant $a > 0$. Then,

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

Proof Let $Y = \mathbb{1}_{X \geq a}$. Then we know:

$$\begin{aligned} \mathbb{E}[Y] &= 0 \cdot \mathbb{P}[X < a] + 1 \cdot \mathbb{P}[X \geq a] = \mathbb{P}[X \geq a] \\ Y &\leq \frac{X}{a} \\ \mathbb{E}[Y] &\leq \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a} \\ \mathbb{P}[X \geq a] &\leq \frac{\mathbb{E}[X]}{a} \end{aligned}$$

Markov's inequality tends to be a coarse bound, and X, a have to be non-negative.

Theorem 1.4 (Chebyshev's Inequality) For random variable X and $\epsilon > 0$:

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

Define $Z = |X - \mathbb{E}[X]|^2$, $a = \epsilon^2$, $\epsilon > 0$.

Proof Apply Markov's inequality:

$$\mathbb{P}\left[Z \geq \epsilon^2\right] \leq \frac{\mathbb{E}[Z]}{\epsilon^2}$$

Note that

$$\mathbb{E}[Z] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathrm{Var}(X)$$

This means that

$$\mathbb{P}\left[\sqrt{Z} \geq \epsilon\right] \leq \frac{\mathrm{Var}(X)}{\epsilon^2}$$

which is exactly the statement of Chebyshev's.

Chebyshev's is generally a tighter bound than Markov's.

Theorem 1.5 (Weak Law of Large Numbers) Assume $X_1, X_2, X_3, \ldots$ are independent random variables with the same expectation μ and the same variance σ^2, and define $Y_n = \frac{(X_1 + X_2 + \cdots + X_n)}{n}$ Then, we have that for any

constant $\epsilon > 0$.

$$\lim_{n\to\infty} \mathbb{P}\left[|Y_n - \mu| \geq \epsilon\right] \to 0$$

This can be shown by Chebyshev's inequality, namely note that the expression in the limit is bounded by $\frac{\text{Var}(Y_n)}{\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} \to 0$.
In words, this means the probability of the sample mean being within ϵ of the true mean approaches 1.

1.3.2 Covariance and Estimation

Definition 1.9 (Covariance)

$$\text{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

If X and Y are independent, then $\text{Cov}(X,Y) = 0$. If the latter is true, then X and Y are uncorrelated. We also define the coefficient of correlation:

$$\rho_{xy} = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

This is handy because $|\rho_{XY}| \leq 1$.

Suppose we want to estimate a random variable Y by $\hat{Y}$ given a correlated random variable Y.

We want to minimize $\mathbb{E}\left[\left(Y - \hat{Y}\right)^2\right]$ but have a linear relationship. This yields LLSE:

Theorem 1.6 (Linear Least Squares Estimate) The LLSE

$$\hat{Y} = \mathbb{E}[Y] + \frac{\text{Cov}(X,Y)}{\text{Var}(X)}(X - \mathbb{E}[X])$$

is the best linear estimate of Y given X.

2 Basic Probability

2.1 Lecture 3, Continued

2.1.1 Infinite Collections of Events and Borel-Cantelli

There are some important consequences of the axioms of probability.

Theorem 2.1 (Infinite Sub-Events) Consider some set A where $A = \bigcup_{n=1}^{\infty} A_n$ where:

$$A_1 \subseteq A_2 \subseteq \dots$$

Then, $\mathbb{P}[A_n] \to \mathbb{P}[A]$.
Furthermore, consider some set B where $B = \bigcap_{n=1}^{\infty} B_n$ where:

$$B_1 \supseteq B_2 \supseteq \dots$$

Then, $\mathbb{P}[B_n] \to \mathbb{P}[A]$.

Theorem 2.2 (Borel-Cantelli Theorem) Let $\{A_n\}_{n=1}^{\infty}$ be a collection of events such that $\sum_{n=1}^{\infty} \mathbb{P}[A_n] < \infty$, then $\mathbb{P}[A_n \text{ infinitely often}] = 0$.
$\{A_n \text{ infinitely often}\}$ is the following event:

$$\{\omega \mid \exists N(\omega) \text{ such that } \forall n > N(\omega), \omega \notin A_n\}$$

In English, the set describes all outcomes where you can assign a number to that outcome such that after A_N, you have no set membership.
The theorem claims that you CAN assign such a number (max event) to any outcome with nonzero probability.

Proof Define

$$B_n = \bigcup_{m \geq n} A_m$$

Note that $B_1 \supseteq B_2 \supseteq B_3 \ldots$
Calling,

$$B = \bigcap_n B_n = B_n$$

Note that $\omega \in \{A_n \text{ io}\}$ if and only if $\omega \in B_n$ for all n. But this means that $\omega \in B$. This means that $\{A_n \text{ io}\} = B$. So we must calculate $\mathbb{P}[B]$. However, note that $\mathbb{P}[B_n] \to \mathbb{P}[B]$ as $n \to \infty$. Thus, we must compute:

$$\begin{aligned}
\mathbb{P}[B] &= \lim_{n \to \infty} \mathbb{P}[B_n] \\
\mathbb{P}[B_n] &\leq \sum_{m=n}^{\infty} \mathbb{P}[A_m] \to 0 \\
\mathbb{P}[B] &= 0
\end{aligned}$$

The second step is justified by the following result from analysis. For non-negative sequence a_n, if $\sum_{i=1}^{\infty} a_n < \infty$, then $\lim_{n \to \infty} \sum_{m=n}^{\infty} a_m \to 0$.
Our result shows that $\mathbb{P}[A_n \text{ io}] = 0$ ∎

Consider the following example for coin flips:

Example 2.1 (Infinite Coin Flips) Consider the experiment of flipping a coin infinitely many times. Let

$$A_n = \{n\text{th flip is heads}\}$$

Then, in this experiment, the event $\{A_n \text{ infinitely often }\}$ (which we denote as $\{A_n \text{ io}\}$)

$$\{A_n \text{ io}\} = \{\omega \mid \text{heads never stop after some } N(\omega)\}$$

Here are some sequences that are in that event:

$$\begin{aligned}
\omega &= 0, 0, 1, 1, 1, 1, \ldots \\
\omega &= 0, 1, 0, 1, 0, 1, \ldots \\
\omega &= 0, 0, \underbrace{\ldots}_{\text{1 million 0's}}, 1, 0, 0, \underbrace{\ldots}_{\text{1 million 0's}}, 1, \ldots
\end{aligned}$$

Now consider the assigning the following probabilities to each heads (instead of the normal, uniform probability space):

$$\mathbb{P}[A_n] = \frac{1}{n^2}$$

$\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges, so by Borel-Cantelli, $\mathbb{P}[A_n \text{ io}] = 0$, i.e. the heads ALWAYS stop.

Now there is one more question. Does $\mathbb{P}[A_n \text{ i.o.}] = 0 \implies \{A_n \text{i.o.}\} = 0$? The answer is no. In this case, $\mathbb{P}[A_n \text{i.o.}] = 0$, but consider the outcome ω_n where the nth flip onwards is a heads; these are all in the infinitely often set, so it actually has infinite cardinality!

2.2 Lecture 4

2.2.1 The Laws of Large Numbers, Revisited

Here is a recap of the two different laws of large numbers.

First, we define two different types of convergence:

Definition 2.1 (Almost Sure Convergence) A random variable X_n almost surely converges to random variable X if

$$\mathbb{P}\left[X_n \to X\right] = 1$$

as $n \to \infty$.

Definition 2.2 (Convergence in Probability) A random variable X_n converges in probability to random variable X if

$$\mathbb{P}\left[|X_n - X| > \epsilon\right] \to 0$$

for any real number $\epsilon > 0$ as $n \to \infty$.

Theorem 2.3 (Strong Law of Large Numbers) Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables. Define:

$$Y_n = \frac{X_1 + \cdots + X_n}{n}$$
$$Y = \mathbb{E}\left[X_1\right]$$

Y_n converges to Y almost surely.

Note the contrast with the weak law of natural numbers. The weak law had only convergence in probability. A key thing to note is that the strong law **implies** the weak law.

2.2.2 Independence

Let us now refine the notion of Independence.

Definition 2.3 (Pairwise Independence) Consider events A_j with $j \in J$.
The events are pairwise independent if for any $j, k \in J$,

$$\mathbb{P}\left[A_j \cap A_k\right] = \mathbb{P}\left[A_j\right]\mathbb{P}\left[A_k\right]$$

Definition 2.4 (Mutual Independence) Consider events A_j with $j \in J$.
The events are mutually independent if

$$\mathbb{P}\left[\bigcap_{j \in K} A_j\right] = \prod_{j \in K} \mathbb{P}\left[A_j\right], \forall K \subseteq J$$

Note that pairwise independence does not imply mutual independence. Here is an example of that edge case:

Example 2.2 Take probability space $\Omega = \{1, 2, 3, 4\}$, all equally likely. Consider the events: $A = \{1, 2\}, B = \{1, 3\}, C = \{1,4\}$.
Note that $\mathbb{P}[A \cap B] = \frac{1}{4} = \mathbb{P}[A]\mathbb{P}[B]$,
but $\mathbb{P}[A \cap B \cap C] = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}[A]\mathbb{P}[B]\mathbb{P}[C]$.

Now with independence, we can find that the converse of Borel-Cantelli is often true:

Theorem 2.4 (Converse of Borel-Cantelli Theorem) Let A_n be a collection of mutually indpendent events such that $\sum_{n=1}^{\infty} \mathbb{P}[A_n] = \infty$. Then, $\mathbb{P}[A_n \text{ infinitely often}] = 1$.

Let us use another example to understand this:

Example 2.3 Let A_n be the same event as the other example (the nth flip is heads) and assign:

$$\mathbb{P}[A_n] = \frac{1}{n}$$

where all the A_n are mutually independent.
Since $\sum_{n=1}^{\infty} \frac{1}{n} = \infty$ and thus by the converse of Borel-Cantelli: $\mathbb{P}[A_n \text{ io}] = 1$.

Example 2.4 (Glued Coins) Suppose you have n coins that are all glued together, i.e. the only two outcomes are $HHH\ldots$ or $TTT\ldots$. Then let A_n = the nth coin is heads. Note that

$$\mathbb{P}[A_n] = \frac{1}{2}$$

which means $\sum \mathbb{P}[A_n] = \infty$. Thus the probability of getting a sequence which has heads infinitely often is 1.

Theorem 2.5 (Kolmogorov's 0-1 theorem) If you have a set of events $\{A_n\}_{n=1}^{\infty}$ that all independent, then

$$\mathbb{P}[A_n \text{ infinitely often}] = 0 \text{ or } 1$$

2.2.3 Conditional Probability

Now we refine conditional probability for many events.

Definition 2.5 (Conditional Probability) Let A and B be two events, and assume $\mathbb{P}[B] > 0$. Then the conditional probability of A given B is:

$$\mathbb{P}[A \mid B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

Theorem 2.6 (Chain Rule) For two events we had $\mathbb{P}[A \cap B] = \mathbb{P}[A \mid B]\mathbb{P}[B]$. For n events A_i, we have:

$$\mathbb{P}[A_1 \cap A_2 \cap \cdots \cap A_n] = \mathbb{P}[A_1]\mathbb{P}[A_2 \mid A_1]\mathbb{P}[A_3 \mid A_1 \cap A_2]\ldots\mathbb{P}[A_n \mid A_1 \cap A_2 \cap \cdots \cap A_{n-1}]$$

if $\mathbb{P}[A_1 \cap A_2 \cap \ldots A_{n-1}] > 0$.

The generalized result above can be shown by induction, taking the case of two events as the base case and then inducting on n. Now we will bring in some of the most powerful tools.

Theorem 2.7 (Law of Total Probability) Let $A, B_1, \ldots B_n$ be events where B_i's are disjoint and $\bigcup_{i=1}^n B_i = \Omega$. Then,

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A \cap B_i]$$

Theorem 2.8 (Bayes' Rule) Let $A, B_1, \ldots B_n$ be events where B_i's are disjoint and $\bigcup_{i=1}^n B_i = \Omega$.

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^n \mathbb{P}\left[A \mid B_j\right]\mathbb{P}\left[B_j\right]}$$

Proof Note that we can use the initial definition to expand the left side:

$$\begin{aligned}\mathbb{P}[B_i \mid A] &= \frac{\mathbb{P}[A \cap B_i]}{\mathbb{P}[A]} \\ &= \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^n \mathbb{P}\left[A \cap B_j\right]} \\ &= \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^n \mathbb{P}\left[A \mid B_j\right]\mathbb{P}\left[B_j\right]}\end{aligned}$$

where the summation in the denominator comes from the law of total probability.

Often the B_j's are termed the prior probabilities, and A is considered the posterior probability.

For an event $B \subseteq R$, $\mathbb{P}[X \in B] = \mathbb{P}\left[(X^{-1}(B))\right]$ where

$$X^{-1}(B) = \{\omega \in \Omega \mid X(\omega) \in B\}$$

.

We can define the following for a random variable to the reals.

Definition 2.6 (Cumulative Distribution Function (CDF)) The Cumulative Distribution Function $F_X(x)$ of random variable X is defined by:

$$F_X(x) = \mathbb{P}[X \in (-\infty, x]] = \mathbb{P}[X \leq x]$$

Here are some properties of the CDF:

- F_X is non-decreasing.
- F_X is right-continuous.
- $F_X \to 0$ as $x \to -\infty$ and $F_X \to 1$ as $x \to \infty$.

Example 2.5 (CDF of an Indicator) Consider the following random variable:

$$I = \begin{cases} 0 \text{ with probability } 1-p \\ 1 \text{ with probability } p \end{cases}$$

Then the $F_I(i)$ is a step function: TODO Add figure

2.3 Lecture 5

Definition 2.7 (Discrete Random Variable) A discrete random variable X can be described fully by:

$$\{(x_n, p_n), n = 1, \ldots, N\}$$

where $p_n = \mathbb{P}[X = x_n]$. This is called the probability mass function (PMF) of X.

We can write the expectation as follows:

$$\mathbb{E}[X] = \sum_{n=1}^{N} x_n p_n$$

With $N = \infty$, the expectation may not be defined.

Definition 2.8 (Function of a Random Variable) Calling $h(X)$ means changing to another random variable with the following PMF:

$$(h(x_n), p_n), n = 1, \ldots, N$$

The expectation of this is as follows:

$$\mathbb{E}[h(X)] = \sum_{n=1}^{N} h(x_n) p_n$$

Definition 2.9 (Coefficient of Variation) The coefficient of variation c of X is defined:

$$c = \sigma_X / \mathbb{E}[X]$$

2.3.1 Common Discrete Distributions

Bernoulli random variables model situations like individual coin flips.

Definition 2.10 (Bernoulli Random Variables) If $X =_D B(p)$ with $p \in [0, 1]$, then the PMF of X is:

$$\{(0, 1 - p), (1, p)\}$$

Furthermore, $\mathbb{E}[X] = p$ and $\text{Var}(X) = p(1 - p)$.

Geometric random variables model the situation where you count the number of coin flips until you get "heads".

Definition 2.11 (Geometric Random Variable) If $X =_D G(p)$ with $p \in [0, 1]$, then the PMF of X is:

$$\mathbb{P}[X = n] = (1 - p)^{n-1} p$$

Furthermore, $\mathbb{E}[X] = \frac{1}{p}$ and $\text{Var}(X) = \frac{1-p}{p^2}$.

The CDF can also be derived as $\mathbb{P}[X \leq n] = 1 - (1 - p)^n$, since it's the complement of failing n times. The CCDF (Complementary CDF) is thus $\mathbb{P}[X < n] = (1 - p)^n$.

Note 2.1 (Memoryless Property) The geometric distribution is memoryless, i.e. if $X =_D G(p)$, then

$$\mathbb{P}[X > m + n \mid X > m] = \mathbb{P}[X > n]$$

Binomial random variables model the situation of doing n coin flips and counting the heads, or the sum of n i.i.d. Bernoulli random variables.

Definition 2.12 (Binomial Random Variable) If $X =_D B(N, p)$ with $p \in [0, 1]$ and $N \geq 1$, then the PMF of X is:

$$\mathbb{P}[X = n] = \binom{N}{n} p^n (1-p)^{N-n}$$

Furthermore, $\mathbb{E}[X] = Np$ and $\text{Var}(X) = Np(1-p)$

The mode of the binomial distribution (the maximum probability) is at $n = \lfloor p(N+1) \rfloor$.

Poisson random variables are the limit of the binomials as the rate of coin flips goes to infinity. This represents the number of successes in an interval during a continuous process.

Definition 2.13 (Poisson Random Variable) If $X =_D P(\lambda)$ with $\lambda > 0$, then the PMF of X is:

$$\mathbb{P}[X = n] = \frac{e^{-\lambda}\lambda^n}{n!}$$

Furthermore, $\mathbb{E}[X] = \lambda$ and $\text{Var}(X) = \lambda$.

In fact, we can make this limit more precise.

Theorem 2.9 (Binomial Converges to Poisson) We have, setting $Np = \lambda$, where λ is fixed,

$$B(N, \lambda/N) \to P(\lambda)$$

2.3.2 Multiple Discrete Random Variables

Consider a pair of random variables (X, Y).

Definition 2.14 (Joint PMF) The joint distribution is given by:

$$p_{i,j} = \mathbb{P}\left[X = x_i, Y = y_j\right]$$

To find the PMF of one of the variables from the joint distribution, we can

Note 2.2 (Marginal PMF from JPMF)

$$\mathbb{P}[X = x_i] = \sum_j \mathbb{P}\left[X = x_i, Y = y_j\right]$$

Furthermore,

Theorem 2.10 (Independence for Random Variables) X and Y are independent if and only if

$$\mathbb{P}[X = x, Y = y] = \mathbb{P}[X = x]\,\mathbb{P}[Y = y]$$

If you have a function of multiple random variables, you can apply it similarly to the one variable case.

$$\mathbb{E}[h(X, Y)] = \sum_i \sum_j h(x_i, y_j)\mathbb{P}\left[X = x_i, Y = y_j\right]$$

First we extend the idea of conditioning to random variables.

Definition 2.15 (Conditional PMF) We call the conditional distribution of Y given X as:

$$\mathbb{P}\left[Y = y_j \mid X = x_i\right] = \frac{\mathbb{P}\left[X = x_i, Y = y_j\right]}{\mathbb{P}\left[X = x_i\right]}$$

Definition 2.16 (Conditional Expectation) The expectation of Y given X (i.e. the best guess of Y given X) is denoted $\mathbb{E}\left[Y \mid X\right]$ and is a function of X

Furthermore, if we want to use a function, we can compute it as follows:

$$\mathbb{E}\left[h(Y) \mid X = x_i\right] = \sum_j h(y_j)\mathbb{P}\left[Y = y_j \mid X = x_i\right]$$

Theorem 2.11 (Properties of Conditional Expectation) For two random variables X, Y,

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] &= \mathbb{E}\left[Y\right] \\
\mathbb{E}\left[h(X)Y \mid X\right] &= h(X)\mathbb{E}\left[Y \mid X\right] \\
\mathbb{E}\left[Y \mid X\right] &= \mathbb{E}\left[Y\right] \text{ if } X \text{ and } Y \text{ are independent} \\
\mathbb{E}\left[h_1(Y) + h_2(Y) \mid X\right] &= \mathbb{E}\left[h_1(Y) \mid X\right] + \mathbb{E}\left[h_2(Y) \mid X\right]
\end{aligned}$$

2.4 Lecture 6

Unfortunately, I didn't transcribe this lecture, as I was very tired. Here is one of the more important results.

In general $X_n \to X \not\Longrightarrow \mathbb{E}[X_n] \to \mathbb{E}[X]$. However, Dominated Convergence Theorem (DCT) and Monotone Convergence Theorem (MCT) provide sufficient conditions in the following form.

Theorem 2.12 (Continuous Tail Sum Formula) Let $X \geq 0$ be a non-negative random variable with $\mathbb{E}[X] < \infty$. Then,

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}[X > x] \, \mathrm{d}x$$

2.5 Lecture 7

2.5.1 Multiple Continuous Random Variables

Definition 2.17 (JCDF) For random variables X and Y, the joint CDF (JCDF) is given by:

$$F_{X,Y}(x,y) = \mathbb{P}[X \le x, Y \le y]$$

For continuous random variables, this is:

$$F_{X,Y}(x,y) = \int_{\infty}^{x}\int_{\infty}^{y} f_{X,Y}(x',y')\,\mathrm{d}y'\,\mathrm{d}x'$$

Then, note that to find $F_{\max X,Y}(k) = \mathbb{P}[\max X, Y \le k] = \mathbb{P}[X \le k, Y \le k]$, we can invoke the JCDF (and make simplifications if X and Y are independent). Furthermore, to find the probability density, we can simply differentiate with respect to k.

Similarly, $\mathbb{P}[\min X, Y > k] = \mathbb{P}[X > k, Y > k]$, which would be the JCCDF (which is the product of the CCDFs if X and Y are independent).

The sum of independent random variables is given by the convolution, *.

Theorem 2.13 (Convolution) Let $Z = X + Y$ where X and Y are independent. Then,

$$f_Z(z) = f_X * f_Y = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\,\mathrm{d}x$$

In addition, we have conditioning for continuous random variables as well.

Definition 2.18 (Conditional PDF) Consider two random variables, X and Y such that $f_X(x) \neq 0$ at the point we are considering. Then:

$$f_{X|Y}(x \mid y) = f_{XY}(x,y)/f_Y(y)$$

Definition 2.19 (Conditional Variance) Let X and Y be random variables. Then we define conditional variance as:

$$\mathrm{Var}(Y \mid X) = \mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2$$

Similar to the law of iterated expectation, there is a similar formula for variance:

Theorem 2.14 (Law of Iterated Variance) Let X and Y be random variables. Then,

$$\mathrm{Var}(Y) = \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])$$

Proof

$$\begin{aligned}
\mathrm{Var}(Y \mid X) &= \mathbb{E}\left[Y^2 \mid X\right] - (\mathbb{E}[Y \mid X])^2 \\
\mathbb{E}[\mathrm{Var}(Y \mid X)] &= \mathbb{E}\left[Y^2\right] - \mathbb{E}\left[(\mathbb{E}[Y \mid X])^2\right] \\
&= \mathrm{Var}(Y) + (\mathbb{E}[Y])^2 - \mathbb{E}\left[(\mathbb{E}[Y \mid X])^2\right] \\
&= \mathrm{Var}(Y) + (\mathbb{E}[\mathbb{E}[Y \mid X]])^2 - \mathbb{E}\left[(\mathbb{E}[Y \mid X])^2\right] \\
&= \mathrm{Var}(Y) - \mathrm{Var}(\mathbb{E}[Y \mid X]) \\
\mathrm{Var}(Y) &= \mathbb{E}[\mathrm{Var}(Y \mid X)] + \mathrm{Var}(\mathbb{E}[Y \mid X])
\end{aligned}$$

An interpretation of this is to think about dividing Ω into disjoint sets where S_i where $X = x_i$. The conditional expectation, The first term is saying that taking the variance of replacing each set with the average over the set, is the first-order approximation. Then, to correct, we average the variances across each and add those in.

3 Page Rank

3.1 Lecture 7, Continued

Page rank is an algorithm originally used by Google for ranking the pages from a keyword search.

It tries to look at the problem as a Markov chain; the weight of each page p is sum over all pages linking to p, times the probability of visiting p from that other page. Symbolically, this is:

$$\pi(i) = \sum_{j \in \mathcal{X}} \pi(j) P(j,i), \forall i \in \mathcal{X}$$

where π is a (row) vector of the weights, and P is a matrix whose (j,i) entry corresponds to the probability of transitioning from j to i. So, we can rewrite the above equations as

$$\pi = \pi P$$

We also add the normalization constraint

$$\sum_{i \in \mathcal{X}} \pi(i) = 1$$

So, π is a probability distribution.

3.1.1 Discrete-Time Markov Chains

Definition 3.1 (Discrete Time Markov Chain (DTMC)) The DTMC $\{X(n), n \geq 0\}$ over state space $\mathcal{X}$ with $P = [P(i,j)]$ as the transition matrix with $P(i,j) = \mathbb{P}[X(n+1) = j \mid X(n) = i]$.
Markov chains have the memoryless property:

$$\mathbb{P}[X(n+1) = j \mid X(n) = i, X(m), m < n] = P(i,j) \forall i, j$$

Generally, we have P constant with respect to time, i.e. it's time-homogenous. Note that we have not shown that defining probabilities like this is consistent with the axioms of probability; let us assume such a choice exists for now.

Then, let $\pi_n(i)$ be the probability that the Markov chain is in state i at time n. Then, for every time step, using vector notation, we get the recurrence:

$$\pi_{n+1} = \pi_n P \implies \pi_n = \pi_0 P^n$$

Definition 3.2 (Stationary Distribution) Let P be the transition matrix of a markov chain. We call π a stationary distribution of the Markov chain if:

$$\pi = \pi P \text{ and } \sum_{i=1}^{n} \pi(i) = 1$$

We call these equations the balance equations.

One thing to note is that we can rewrite the equations as $(P - I)\pi = 0$. However, since every row sums to 0, this is not a full rank matrix, so we need the extra constraint about the sum of all the elements of π.

Next, we have some ways to classify Markov chains.

Definition 3.3 (Irreducible) A Markov Chain is irreducible if it can reach any state from any other state (possibly in multiple steps).

Definition 3.4 (Aperiodic) Let $d(i) = \gcd\{n \geq 1 \mid P^n(i,i) > 0\}$. An irreducible DTMC is aperiodic if $d(i) = 1$ for all i (in fact in an irreducible DTMC, $d(i)$ is the same as i).

Note that if a Markov chain has a self loop, it is aperiodic because we can get from i to i in a single step, so $d(i) = 1$ always.

Example 3.1 Consider the following Markov chain, which is irreducible and aperiodic. (DIAGRAM NEEDED)

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0.6 & 0 & 0.4 \\ 0 & 0.9 & 0.1 \end{bmatrix}$$

Note that $P(i,j)$ is the probability of going from i to j. Note that all the rows sum to 1 (by the total probability rule). Now let us solve the balance equations:

$$\pi = \pi P$$

$$\begin{bmatrix} \pi_0 & \pi_1 & \pi_2 \end{bmatrix} = \begin{bmatrix} \pi_0 & \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ 0.6 & 0 & 0.4 \\ 0 & 0.9 & 0.1 \end{bmatrix}$$

This yields the equations:

$$\begin{cases} 0.6\pi_1 = \pi_0 \\ \pi_0 + 0.9\pi_2 = \pi_1 \\ 0.4\pi_1 + 0.1\pi_2 = \pi_2 \\ \pi_0 + \pi_1 + \pi_2 = 0 \end{cases}$$

Note that some of these equations are dependent, so we needed the extra constraint. This can then be solved to get:

$$\pi = \begin{bmatrix} 0.294 & 0.489 & 0.217 \end{bmatrix}$$

With these notions, we have the following results.

Theorem 3.1 (Big Theorem for Finite DTMC) Consider an irreducible DTMC over a finite state space. Then,

- There is a unique invariant distribution π
- The long-term fraction of time spent in state n is given by

$$\lim_{N \to \infty} \frac{1}{N} \sum_{n=0}^{N-1} \mathbb{1}_{\{X(n)=i\}} = \pi(i)$$

- If the DTMC is aperiodic, $\pi_n \to \pi$ as $n \to \infty$, independent of π_0

Another way to state the last result is that each row in P^n converges to π as $n \to \infty$. Why? Consider $\pi_0 = \begin{bmatrix} 1 & 0 & \dots & 0 \end{bmatrix}$. Then $\pi_n = \pi_0 P^n = p_1 \to \pi$, so the first row approaches π. The argument is similar for any row.

3.2 Lecture 8

3.2.1 Hitting Times

Note 3.1 Consider a Markov Chain X_i. Suppose we wanted to find the time it takes to hit any state in some set A, starting at some other state i. Formally, we are asking,

$$\beta(i) := \mathbb{E}\left[T_A \mid X_0 = i\right], i \in \mathcal{X}, A \subseteq \mathcal{X}$$

Since Markov chains are memoryless, we only need to consider the ways to step from edges out of us, i.e. one step. We call these equations first-step equations (FSEs). The FSEs for this problem are:

$$\begin{cases} 1 + \sum_j P(i,j)\beta(j) & i \notin A \\ 0 & i \in A \end{cases}$$

The reasoning for this is as follows. If we are already in set A, it takes us 0 steps to get into it. Otherwise, we take one step to another state j, which has probability $P(i,j)$ of happening, and then the value of $\beta(i)$ will just be $1 + \beta(j)$.

This way of solving for hitting times produces n^2 linear equations to solve.

Here are some other problems we can solve with FSEs/Hitting times.

Note 3.2 Suppose you want to find the probability of hitting a state in one set A before another B,

$$\alpha(i) := \mathbb{P}\left[T_A < T_B \mid X_0 = i\right], i \in \mathcal{X}, A,B \subseteq \mathcal{X}, A \cap B = \emptyset$$

Then, the FSEs are the following:

$$\begin{cases} \sum_j P(i,j)\alpha(j) & i \notin A \cup B \\ 1 & i \in A \\ 0 & i \in B \end{cases}$$

The reasoning is similar: once you hit a bad state, your probability drops to 0 and if you hit a good state, your probability is 1. Otherwise, it is the sum of the probabilities of of moving to an adjacent state and that state ending up as something good (the sum is due to total probability rule).

Note 3.3 Consider a discounted reward for visiting states earlier. Define $h(i)$ as the reward for being in state i. Then we define

$$Z = \sum_{n=0}^{T_A} \beta^n h(X(n)), A \subseteq \mathcal{X}, 0 < \beta \leq 1$$

as the discounted reward (from a gambling run, perhaps). You accrue wealth until you hit A, but take too long and your reward is reduced. We want to find:

$$\mathbb{E}\left[Z \mid X_0 = i\right], i \in X$$

The FSEs then become:

$$\begin{cases} h(i) + \beta \sum_j P(i,j)\delta(j) & i \in A \\ h(i) & i \in A \end{cases}$$

3.3 Lecture 9

3.3.1 Laws of Large Numbers

We now discuss the law(s) of large numbers in detail, and how they relate to Markov chains.

Theorem 3.2 (Laws of Large Numbers) Let $\{X(i), i \geq 1\}$ be a sequence of independent and identically distributed (IID) random variables, with mean μ and let $S(n) = \sum_{i=1}^{n} X(i)$. Assume $\mathbb{E}[|X(i)|] < \infty$.
The Strong Law of Large Numbers (SLLN) states that:

$$\mathbb{P}\left[\lim_{n\to\infty} S(n)/n = \mu\right] = 1$$

i.e. $S(n)/n$, the sample mean, converges to the true mean μ **almost surely**.
The Weak Law of Large Numbers (WLLN) states that fixing $\varepsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left[\left|\frac{S(n)}{n} - \mu\right| > \varepsilon\right] = 0$$

i.e. $S(n)/n$, the same mean, converges to the true mean μ in **probability**.

Note that SLLN implies WLLN, so SLLN is stronger. Furthermore,

$$\mathbb{E}[S_n/n] = \frac{\mathbb{E}[\sum_i X_i]}{n} = \frac{n\mathbb{E}[X_i]}{n} = \mathbb{E}[X_i]$$

So, convergence in expectation is even weaker than WLLN (convergence in probability).

We will show the first implication by showing something stronger; that almost sure convergence implies convergence in probability.

Proof 3.1 (A.S. implies Convergence in Probability) Fix $\varepsilon > 0$ and let X_n converge to X almost surely. Define the following events:

$$\begin{aligned} A_n &= \{|X_n - X| > \varepsilon\} \\ B_n &= \cup_{m\geq n} A_m \\ B &= \cap_{n=1}^{\infty} B_n \\ B &= \{A_n \text{i.o.}\} \end{aligned}$$

However, by definition of almost sure convergence, we must have $\mathbb{P}[A_n \text{i.o.}] = 0$. However:

$$\begin{aligned} &\mathbb{P}[|X_n - X| > \varepsilon] \leq \mathbb{P}[B_n] \to 0 \\ &\mathbb{P}[|X_n - X| > \varepsilon] \to 0 \end{aligned}$$

However, convergence in probability does not always imply almost sure convergence. Here is a counter-example:

Example 3.2 (Not all convergence is made equal) Pick ω uniformly in $[0, 1]$.
Define $X_1(\omega) = 1$ and then for $n \geq 2$, make it piecewise, with value 1 for a width $1/n$ and value 0 for a width $1 - 1/n$. Furthermore, we stack the start of the pulse at the end of the interval of the last one; wrapping it around as needed (modulo 1).
For any ε, the width of X_n being 1 goes down to 0, as we continue along.

$$\mathbb{P}[|X_n - 0| > \varepsilon] \to 1$$

However,

$$\mathbb{P}\left[X_n \to 0\right] = 0$$

since for some ω, we can always find an X_n later which is 1 (there will be another pulse containing ω).

Armed with the machinery, now we can finally prove the Big Theorem.

Proof 3.2 Let us focus on part a). We define:

$$m_j = \mathbb{E}\left[T_j \mid X(0) = j\right]$$

where $T_j = \min\{n > 0 \mid X(n) = j\}$ (i.e. first time visiting j).
Then, we claim the following equality holds (for large N).

$$N \sum_j \frac{1}{m_j} P(j,i) = \frac{N}{m_i}$$

Each term in the left hand side is the long term fraction of time you've spent in visiting j and then i, while the term on the right hand side is the long term fraction of time you've spent visiting i.
Dividing through by N, we have that:

$$\sum_j \frac{1}{m_j} P(j,i) = \frac{1}{m_i}$$

calling $\pi = \begin{bmatrix} \frac{1}{m_1} & \frac{1}{m_2} & \dots \end{bmatrix}$, then we essentially have:

$$\pi(j)P(j,i) = \pi(i)$$
$$\pi P = \pi$$

Proving uniqueness is simple and was left out of lecture.

Next let us prove part b). Let $A(n)$ be the number of visits to state i by time n. Let T_j^ℓ be the difference between ℓth time you came back to j and the $\ell - 1$th time. Then:

$$A(n)/n \sim \frac{k}{T_i^1 + T_i^2 + \cdots + T_i^k}$$

since the LHS is frequency. However, note that T_i^ℓ are i.i.d. because the Markov chain is memoryless, so coming back to yourself the next time Then:

$$\frac{1}{n/A(n)} = \frac{T_i^1 + T_i^2 + \cdots + T_i^k}{k}$$

However, the RHS almost surely goes to the average value, m_i. This means

$$\frac{1}{n/A(n)} \to \frac{1}{m_i} = \pi(i)$$

4 Multiplexing

4.1 Lecture 9, Continued

Multiplexing relates to the sharing of a common resource. Consider some link or channel with some rate capacity C. Suppose it is being shared by some group of connections $x_1, x_2, \ldots, x_n$, but not all connections are active at the same time. Think about a phone line–everyone doesn't have to use the phone at the same time. Then, the rate for each connection is C/v, where v is a random variable the represents the number of active connections.

Let us model link sharing with each channel being active as a Bernoulli random variable with probability p of the link being active. Then,

$$v \sim B(n, p)$$

We also discuss the following:

Definition 4.1 (PPF) The percent point function (PPF) of a random variable is the "inverse" CDF. Note that a lot of CDFs are not bijections, so instead we say the following.
Suppose $F_X(x) = p_1$ for all $x \in [x_1, x_2)$ and then it jumps to p_2 at x_2, the PPF as follows:

$$PPF(p) = x_2 \qquad \forall p \in (p_1, p_2]$$
$$PPF(p) = x_1 \qquad p = p_1$$

Then suppose we wanted to find the smallest m such that $\mathbb{P}[v > m] \leq \delta$ or $\mathbb{P}[v \leq m] \geq 1 - \delta$. If δ is small, this means that each active user will get at least a rate of C/m with probability $1 - \delta$ or higher (think of this rate as the speed of your internet connection, for example). To do this, just take $PPF(1 - \delta)$ to get the correct m to be confident (typically $\delta = 0.05$).

Now, we return to Normal/Gaussian random variables, as they help us investigate the sums of random variables. Here's a few useful facts:

Note 4.1 For $X \sim \mathcal{N}(\mu, \sigma^2)$:

- $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$, $-\infty < x < \infty$.
- $X = \mu + \sigma W$, where $W \sim \mathcal{N}(0, 1)$ (the standard normal). You can show this by change of variables on the CDF.
- $\mathbb{P}[W > 1.65] \approx 0.05$, $\mathbb{P}[W > 1.96] \approx 0.025$, $\mathbb{P}[W > 2.32] \approx 0.01$
- The above facts holds for general Gaussian RV (i.e. if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\mathbb{P}[X > \mu + 1.65\sigma] = 0.05$)
- The normal is symmetric about its mean.

4.2 Lecture 10

4.2.1 Central Limit Theorem

First, we talk about another type of convergence.

Definition 4.2 (Convergence in Distribution) Let $\{X(n), n \geq 1\}$ and X be random variables. We say $X(n)$ converges in distribution to X (or it weakly converges), and write $X(n) \xrightarrow{d} X$, if

$$\mathbb{P}[X(n) \leq x] \to \mathbb{P}[X \leq x] \; \forall x \text{ s.t. } \mathbb{P}[X = x] = 0$$

Let us see why the $\mathbb{P}[X = x] = 0$ is necessary.

Example 4.1 Consider $X(n) = 3 + \frac{1}{n}$ and $X = 3$. We want these to converge in distribution, but consider that

$$\mathbb{P}[X(n) \leq 3] = 0$$

for all n, but $\mathbb{P}[X \leq 3] = 1$, which would lead a lack of convergence. However, we stipulate that we should only consider points x where there is no discrete mass; $x = 3$ does not qualify because $\mathbb{P}[X = 3] = 1 > 0$. So, the convergence in distribution still happens!

This is the weakest type of convergence we have discussed. If we know convergence in probability of $X(n)$ to X, then we can conclude that $X(n) \xrightarrow{d} X$. Let us see a counter example of the converse.

Example 4.2 Take $X_n = X \sim \text{Bernoulli}(1/2)$. Note that $X_n \xrightarrow{\text{d}} 1 - X$ because the CDFs are identical. However, take $0 < \varepsilon < 1$. Then:

$$\begin{aligned}\mathbb{P}[|X_n - (1 - X)| > \varepsilon] &= \mathbb{P}[|X - (1 - X)| > \varepsilon] \\ &= \mathbb{P}[|2X - 1| > \varepsilon] \\ &\to 1\end{aligned}$$

We now have an important result, armed with our Gaussian knowledge.

Theorem 4.1 (Central Limit Theorem) Let $\{X(n), n \geq 1\}$ be a set of iid random variable with mean $\mathbb{E}[X(n)] = \mu$ and variance $\text{Var}(X(n)) = \sigma^2$. Define $S(n) = \sum_{i=1}^{n} X(i)$. Then,

$$\frac{S(n) - n\mu}{\sigma\sqrt{n}} \xrightarrow{\text{d}} \mathcal{N}(0, 1)$$

In practice, we can think of CLT in another way. Let X_i be iid random variables with mean μ and variance σ^2. Then

$$X_1 + X_2 + \cdots + X_n \approx \mathcal{N}\left(n\mu, n\sigma^2\right)$$

Here is a quick result, based on the fact that a Binomial RV is the sum of many Bernoulli RVs.

Theorem 4.2 If $X_N \sim \text{Binomial}(N, p)$, then:

$$X_N \approx \mathcal{N}(Np, Np(1 - p))$$

for large N.

Now let us see how to apply CLT to find bounds. Let $Y(N) \sim \text{Binomial}(N, p)/N$. Then

$$\sigma_{Y(N)} = \sqrt{p(1-p)N/N^2} = \sqrt{p(1-p)/N}$$

and the mean of $Y(N)$ is p. Define

$$A_1 = \{Y(N) \geq p + 1.65\sqrt{p(1-p)/N}\}$$
$$A_2 = \{Y(N) \leq p - 1.65\sqrt{p(1-p)/N}\}$$

Due to CLT, $\mathbb{P}[A_1 \cup A_2] \approx 0.1$ (from the tail identities we covered last lecture) and thus $\mathbb{P}\left[A_1^C \cap A_2^C\right] \approx 0.9$, i.e.

$$\mathbb{P}\left[p - 1.65\sqrt{p(1-p)/N} \leq Y(N) \leq p + 1.65\sqrt{p(1-p)/N}\right] = 0.9$$
$$\mathbb{P}\left[Y(N) - 1.65\sqrt{p(1-p)/N} \leq p \leq Y(N) + 1.65\sqrt{p(1-p)/N}\right] = 0.9$$

This statement states that $Y(N)$ as an estimator for p is within this interval with confidence 90Replacing the argument with 1.96 gives 95% confidence. Sometimes, we argue that $p(1-p) < 1/4$, so

$$\mathbb{P}\left[Y(N) - 1.65\sqrt{0.25/N} \leq p \leq Y(N) - 1.65\sqrt{0.25/N}\right] = 0.9$$
$$\mathbb{P}\left[Y(N) - \frac{0.83}{\sqrt{N}} \leq p \leq Y(N) - \frac{0.83}{\sqrt{N}}\right] = 0.9$$

However, when we only have observations and do not know the underlying distribution, we can do the following.

Note 4.2 Consider any iid random variables $\{X(n), n \geq 1\}$ without knowledge of variance.

$$\mu_n = \frac{X(1) + X(2) + \cdots + X(n)}{n}$$
$$\sigma_n^2 = \frac{\sum_{m=1}^{n}(X(m) - \mu_n)^2}{n-1}$$

where the $n-1$ comes from the fact that we want σ_n^2 in expectation to be the same as the true variance of the $X(i)$'s. (see Walrand for derivation).
Then, the following is a 90% confidence interval for $\mu = \mathbb{E}[X(i)]$:

$$[\mu_n - 1.65\frac{\sigma_n}{\sqrt{n}} < \mu < \mu_n + 1.65\frac{\sigma_n}{\sqrt{n}}]$$

and this is a 95% confidence interval for μ:

$$[\mu_n - 2\frac{\sigma_n}{\sqrt{n}} < \mu < \mu_n + 2\frac{\sigma_n}{\sqrt{n}}]$$

4.3 Lecture 11

4.3.1 Characteristic and Moment-Generating Functions

Definition 4.3 (Characteristic Function) The characteristic function of a random variable X is the function:

$$\phi_X(u) = \mathbb{E}\left[e^{iuX}\right]$$

with domain $u \in \mathbb{R}$ and $i = \sqrt{-1}$.

Definition 4.4 (Moment-Generating Function) The moment-generating function (MGF) of a random variable X is the function:

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

with domain $t \in \mathbb{R}$.

The characteristic function and moment generating function uniquely determines a random variable. For example, you can determine the associated PDF/CDF from it.

Note that the MGF does not always exist (if X is large with high probability, it may blow up), but when it does the relationship between it and the characteristic function can be summarized as:

$$\phi_X(t) = M_{iX}(t) = M_X(it)$$

Furthermore, we have the following:

Note 4.3 (Moment Generation)

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \mathbb{E}\left[\sum_{i=0}^{\infty} \frac{(tX)^i}{i!}\right] \\ &= \sum_{i=0}^{\infty} \frac{t^i \mathbb{E}\left[X^i\right]}{i!} \end{aligned}$$

So,

$$\mathbb{E}\left[X^n\right] = M_X^{(n)}(0)$$

i.e. the nth derivative of M_X evaluated at 0.

We work through an example.

Example 4.3 (Characteristic of $\mathcal{N}(0,1)$) We apply the definition with LOTUS directly.

$$\begin{aligned} \phi_X(u) &= \int_{-\infty}^{\infty} e^{iux} \frac{1}{2\pi} e^{-x^2/2}\,\mathrm{d}x \\ \frac{\mathrm{d}}{\mathrm{d}u}\phi_X(u) &= \int_{-\infty}^{\infty} ixe^{iux} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\,\mathrm{d}x \\ &= -\int_{-\infty}^{\infty} ie^{iux} \frac{1}{2\sqrt{\pi}}\,\mathrm{d}e^{-x^2/2} \end{aligned}$$

Now we apply integration by parts:

$$\begin{aligned}
\phi_X'(u) &= -ie^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} i\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,\mathrm{d}e^{iux} \\
\phi_X'(u) &= -u\int_{\infty}^{\infty} e^{iux}\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\,\mathrm{d}x \\
\phi_X'(u) &= -u\phi_X(u) \\
\frac{\phi_X'(u)}{\phi_X(u)} &= -u \\
(\log(\phi_X(u)))' &= -u \\
\log(\phi_X(u)) &= -u^2/2 + c \\
\phi_X(u) &= Ae^{-u^2/2}
\end{aligned}$$

Furthermore, we know that $\phi_X(0) = 1$ (since this is just the integral of the Normal over $\mathbb{R}$). Thus,

$$\phi_X(u) = e^{-u^2/2}$$

Now, let us use this form to find the moments of $\mathcal{N}(0,1)$.

$$\begin{aligned}
\phi_X(u) &= \mathbb{E}\left[e^{iuX}\right] \\
&= \sum_{n=0}^{\infty}\frac{1}{n!}i^n u^n \mathbb{E}\left[X^n\right] \\
&= e^{-u^2/2} \\
&= \sum_{m=0}^{\infty}\frac{1}{m!}\left(\frac{-u^2}{2}\right)^m
\end{aligned}$$

Now, we match coefficients of u^{2m}.

$$\begin{aligned}
\frac{1}{(2m)!}i^{2m}\mathbb{E}\left[X^{2m}\right] &= \frac{1}{m!}\left(\frac{-1}{2}\right)^m \\
\mathbb{E}\left[X^{2m}\right] &= \frac{(2m)!}{m!\cdot 2^m}
\end{aligned}$$

Note that the summation has no terms for u^{2m+1}, so the coefficients are all zero. This means $\mathbb{E}\left[X^{2m+1}\right] = 0$. Altogether:

$$\mathbb{E}\left[X^n\right] = \begin{cases} \frac{n!}{(n/2)!\cdot 2^n} & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd} \end{cases}$$

Here is a useful formula for MGFs:

Theorem 4.3 (Independent MGFs) Consider indepdent random variables X_1 and X_2. Then the moment-generating function of their sum is given by:

$$M_{X_1+X_2}(t) = M_{X_1}(t)M_{X_2}(t)$$

4.3.2 Proof of the CLT

We give a brief sketch of the proof of the Central Limit Theorem. This is not a full argument, as we do not show that the characteristic function uniquely determines a random variable.

Proof 4.1 Define

$$Y(n) = \frac{X(1) + X(2) + \cdots + X(n) - n\mu}{\sigma\sqrt{n}}$$

Then, let us compute its characteristic function.

$$\begin{aligned}
\phi_{Y(n)}(u) &= \mathbb{E}\left[e^{iuY(n)}\right] \\
&= \mathbb{E}\left[\prod_{m=1}^{n} \exp\left\{\frac{iu(X(m)-\mu)}{\sigma\sqrt{n}}\right\}\right] \\
&= \prod_{m=1}^{n} \mathbb{E}\left[\exp\left\{\frac{iu(X(m)-\mu)}{\sigma\sqrt{n}}\right\}\right] && \text{(Independence)} \\
&= \mathbb{E}\left[\exp\left\{\frac{iu(X(1)-\mu)}{\sigma\sqrt{n}}\right\}\right]^n && \text{(Identically Distributed)} \\
&= \mathbb{E}\left[1 + \frac{iu(X(1)-\mu)}{\sigma\sqrt{n}} - \frac{u^2(X(1)-\mu)^2}{2\sigma^2 n} + o\left(\frac{1}{n}\right)\right]^n \\
&= \left(1 - \frac{iu(\mathbb{E}\left[X(1)-\mu\right])}{\sigma\sqrt{n}} - \frac{u^2}{2}\cdot\frac{(X(1)-\mu)}{\sigma^2} + o\left(\frac{1}{n}\right)\right)^n \\
&= \left(1 - \frac{u^2}{2} + o\left(\frac{1}{n}\right)\right)^n \\
&\to e^{-u^2/2} && \text{(Limit Definition of } e\text{)}
\end{aligned}$$

Since the characteristic function converges to that of a standard normal, this means that the distribution converges as well.