Everything You Need To Know About MAT377

Nathanael Chwojko-Srawley

December 14, 2021

Contents

> This course introduces students to various topics in mathematical probability theory. Topics include basic concepts (such as probability, random variables, expectations, conditional probability) from a mathematical point of view, examples of distributions and stochastic processes and their properties, convergence results (such as the law of large numbers, central limit theorem, random series, etc.); various inequalities, and examples of applications of probabilistic ideas beyond statistics (for example, in geometry and computer science).

So far, my intuition on probability is that:

1. Define probability rigorously (we define a probability space Ω with the possible events, and the probability of the sum of the probabilities of all the events must be 1)

2. capture events quantitatively (ex. $\{\text{heads}, \text{tails}\}$, we can make a function $f : \{\text{heads}, \text{tails}\} \to \{0, 1\}$ that gives a quantitative values). This function will represent an arbitrary choice, and so we call it a *random variable*

3. We'll take a lot of time on how to analyze the average value this function will return, or more precisely the *expected value* (given you know the probability of all the events, and you have quantified what each even means with your function, what value do you "expect" to happen")

4. Then, we might ask how many of the events land close to the expected value, that will define the *variance*.

5. Switching gears a bit, we will want to learn how to simplify really complicated problems by learning how to *approximate* them. Since we're working over $\mathbb{R}$, we have an order, and so bounding will usually mean that we can bound value from above and bellow, and so we will define many *inequalities*. For this, I highly recommend you look at the proofs, since they contain many bounding tricks.

Preliminaries

In this section, I will put down a varying number of results that will be referenced in the book, sometimes only once, sometimes multiple times. These results are put here so that the reader can, if they so wish, be “ready in advance” for these results. I found that knowing these in advanced helped with the flow of my reading.

I will put a word on where the particular results in this chapter will be used, and if it will be used once, multiple times, or be a recurring property.

Inclusion Exclusion principle

Often, will be dealing with terms like $(x_1 + x_2 + \cdots x_n)^k$, and we will want to simplify further. Here is my intuition on how to remember the following formula:

$$(x_1 + x_2 + \cdots + x_n)^k = \sum_{k_1,\ldots,k_n} \binom{n}{k_1, k_2, \ldots k_n} (x_1)^{k_1}(x_2)^{k_2}\cdots(x_n)^{k_n} \qquad \sum_{i=1}^{n} k_i = k$$

or in the special case of 2 variables:

$$(x+y)^n = \sum_{k_x,k_y} \binom{n}{k_x, k_y} x^{k_x} y^{k_y} = \sum_{k=1}^{n} \binom{n}{k} x^k y^{n-k}$$

Start with $(x_1 + x_2 \cdots + x_n)(x_1 + x_2 \cdots + x_n)$. The variable x_1 can multiply any of the variables on the left, producing $x_1x_1, x_1x_2, ..., x_1, x_n$. TBD

Stirling’s Formula:

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

Cauchy-Shwartz inequality

a little bit of big-O

A little bit of Taylor series for some inequalities.

Recall the trick for compution $\int e^{x^2} dx$ from multivariable calculus: compute$I^2 = \int e^{x^2+y^2} dxdy$ by doing change of variable to polar coordiantes, and then take the square-root.

A trick to remembering the number of functions from finite sets $f : A \to B$: Let's say $|A| = 1$ and $|B| = 2$. Then the single element of A can map to one of 2 positions, so there are 2^1 possibilites. Dually, if $|A| = 2$ and $|B| = 1$, then both elements must only map to 1 position, so there are 1^2 possibilites. In general, there are $|B|^{|A|}$ possibilities.

1

Introduction

1.1 Motivation and Basic Terminology

The following example if from Pachenko as a motivation to why we should study probability. The essence of this example is that we can sometimes answer existential questions using statistical means:

Example 1.1: Motivation

Let $n \geq 2$, and fix a choice of n vector on the unit sphere

$$v_1, v_2, ..., v_n \in S^n$$

or equivalently n vectors from $\mathbb{R}^n$ where $\|v_i\| = 1$ for all i. Then among all linear combinations of the form

$$v = v_1 \pm v_2 \pm \cdots \pm v_n$$

does there exists one such that $\|v\| \leq \sqrt{n}$? What about $\|v\| \geq \sqrt{n}$?

Notice that if all the vectors are orthogonal, then we will form a "n"-triangle in $\mathbb{R}^n$, and os $\|v\| = \sqrt{n}$, but it is not necessarily the case that it will be.

The way we solve this is the following: Consider the vector space $\{-1, 1\}^n$ (i.e. $(\mathbb{Z}/2\mathbb{Z})^n$ as an n dimensional $\mathbb{Z}/2\mathbb{Z}$ vector-space). Consider any $\varepsilon \in \{-1, 1\}^n$ and define the function

$$v(\varepsilon) = \varepsilon_1 v_1 + \varepsilon_2 v_2 + \cdots + \varepsilon_2 v_2$$

note that the choice of the v_i's are fixed, and so v is fixed, meaning we only need to keep track of the choice of ε (equivalently, the thing over which we have a choice is ε). The question now becomes if we can find an ε such that $\|v(\varepsilon)\| \leq \sqrt{n}$. In fact, let's modify this question a bit: it is equivalent to ask where there exists an ε such that $\|v(\varepsilon)\|^2 \leq n$. We do this since working with a square-root can be cumbersome, and getting rid of it does not change the question. To proceed with the question, remember that if we compute the average of a set of numbers then there must exist at least one number that's higher than or equal to the average (equivalently lower than or equal to the average). To this end, take

$$\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\|v(\varepsilon)\|^2 = \frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\|\varepsilon_1 v_1 + \cdots + \varepsilon_n v_n\|^2 = \frac{1}{2^n}\sum_{\varepsilon\in\{-1,1}^n\sum_{i,j=1}^n \varepsilon_i\varepsilon_j v_i v_j$$

then we can split the equation in the following way

$$\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\sum_{i,j=1}^n \varepsilon_i\varepsilon_j v_i v_j = \sum_{i,j=1}^n v_i v_j \frac{1}{2^n}\left(\sum_{\varepsilon\in\{-1,1\}^n}\varepsilon_i\varepsilon_j\right)$$

at this point, we will split the some in when $i = j$ and $i \neq j$. If $i = j$, then $\varepsilon_i\varepsilon_j = \varepsilon_i\varepsilon_i = 1$, so we get

$$\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\varepsilon_i\varepsilon_j = \frac{2^n}{2^n} = 1$$

if $i \neq j$, we can (rest here, lot's of things that are tedious to type out p.3)

Thus, we get

$$\frac{1}{2^n}\sum_{\varepsilon\in\{-1,1\}^n}\|v(\varepsilon)\|^2 = n$$

Thus, since the average is n, the average of $\|v(\varepsilon)\|$ is $\sqrt{n}$, and by the logic earlier it must be that at least one element is of the form $\|v(\varepsilon)\| \leq \sqrt{n}$ or $\|v(\varepsilon)\| \geq \sqrt{n}$.

We can then ask a more interesting question: Will *most* linear combination of this kind result in $\|v(\varepsilon)\| \leq \sqrt{n}$. It turns out in this case that most are around the value of $\sqrt{n}$, and you might really be trying to get a value further away from it. We will return to this in chapter ref:HERE

This shows the power of probability. Probability is an old study that pre-dates some common set-theory vocabulary Therefore, there is an equivalence in vocabulary:

notation	set theory	probability
X	set, space	Universe, Probability space
$x \in X$	element	outcome
$A \subseteq X$	subset	event
$f : X \to \mathbb{R}$	function	random variable
	average	expected value

1.2 Basic Definitions

In the previous example, the set $\Omega = \{-1, 1\}^n$ was the probability space. Notice that every element in the probability space was associated with some "chance" or "probability" of choosing that element. More generally, if we choose some subset (i.e. event) from Ω, we can give a probability of that event happening. This idea of capturing the chances of an outcome (or more precisely an event[1]) is

[1] we can think of an outcome as an event with a single outcome

captured by a special function $\mathbb{P} : P(\Omega) \to \mathbb{R}$ called a *Probability*, where $P(\Omega)$ is the powerset of Ω. If you have taken Real Analysis, this is called the *probability measure*, since it is an example of a particular measure that will follow a couple of properties we will list bellow that will make it align with our intuition of probability.

Every probability space *must* have an associated probability. In the previous example, we sneakily used the probability $\mathbb{P}(\{\varepsilon\}) \overset{!}{=} \mathbb{P}(\varepsilon) = \frac{1}{2^n}$ where $\overset{!}{=}$ is a common abuse of notation for event's with only 1 outcome. For a general event $A \subseteq \Omega$ in the previous example, we have:

$$\mathbb{P}(A) = \frac{\text{Card}(A)}{2^n}$$

this particular probability is called the *uniform measure* since every outcome was weighed equally.

What properties should $\mathbb{P}$ have? Let's take for example the roll of a dice where each side has equal probability. Should we include the chance that the dice gets stuck in a corner, so that the probability of any of the faces showing up is less than 1? Should we have a special variable called "fluke" which discounts that possibility? If one of the outcomes affect another (let' say you rolled the dice onto a sticky surface, so now one of the faces will stick to the floor more when you role the dice again). Should the probability take this into account, or should it be the responsibility of another function that takes into account dependent events? Throughout the centuries, the expected properties of what this probability function $\mathbb{P}$ must follow were greatly philosophised about. Eventually, a Russian mathematician named Kolmogorov formalized the study of statistics by making a probability need to follow the following properties:

Definition 1.2.1: Probability [measure]

Let $\mathbb{P}$ be a function from a probability space $P(\Omega) \to \mathbb{R}$. Then if $\mathbb{P}$ satisfies:

1. For all $A \subseteq \Omega$, $\mathbb{P}(A) \in [0, 1]$
2. $\mathbb{P}(\Omega) = 1$
3. If $A_1, A_2 \subseteq \Omega$ are disjoint then

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2)$$

then $\mathbb{P}$ is called a *probability [measure]*

There is in fact a couple more caveats when it comes to defining a probability that appear if Ω is infinite (in particular at least uncountably infinite). It will turn out that these rules will contradict each other[2], and so we would also restrict from $P(\Omega)$ to $\mathcal{M}(\Omega)$, which is a special subset of $P(\Omega)$ called the set of *measurable sets*. The details of this restriction is covered in a course on measure theory, and will never be a problem for us. Simply know that the only *unmeasurable sets* that we [currently] know of were always purposefully esoteric to find a counter-example, and never came up as an actual problem when doing probability.

[2] see Vitali sets

Example 1.2: probabilites

Show the following

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
2. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$
3. $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
4. $\mathbb{P}(\bigcup_{i=1}^{n} A_i) \leq \sum_{i=1}^{l} \mathbb{P}(A_i)$

If you have done some measure theory, these are are essentially the same proofs for finite measures (since $\mathbb{P}(X) = \mu(X) < \infty$)

This now gives us the language to talk about the chances of something happening.

The next piece of the puzzle in the picture of probability is a way to capture the idea of asking "probability questions" on a space. Each "probability question" will depend on a numerical values to the outcomes or events in a space. As an example, imagine you have 10 coins labeled x_1 to x_{10}, where coin x_k has value k. We might ask ourselves the following question: if we choose n coins, what's the sum of all the values? If we take $\Omega = P(\{x_1, ..., x_{10}\})$ then define

$$f : \Omega \to \mathbb{R} \qquad f(A) = \text{sum of the value}$$

Using this, we can ask some questions like:

1. What is the average value you get by picking any possible random outcome?
2. How many outcomes are "close" the average? Is it spread out? Can we bound "most" close to the average?

These two questions will be central to the beginning of our exploration of statistics

For example, we might want to know the average number of a dice roll. The probability will capture the fact that each face has a $\frac{1}{6}$ chance of happening, but not that the faces have values ranges from $1, ..., 6$. To capture this idea, we use a different notion called a *random variable.*

Definition 1.2.2: Random Variable

A random variable is a function $X : \Omega \to \mathbb{R}$ that gives a numerical value to events in Ω[a].

[a] Just like for $\mathbb{P}$, we are technically restricting to $\mathcal{M}(\Omega)$

Since the codomain is the real numbers, we take on the usual definition of

$$(XY)(a) = X(a)Y(a) \qquad (X + Y)(a) = X(a) + Y(a)$$

At the beginning, it might be confusing that random variable's are denoted with X or Y. This will be more intuitive later one when we treat these functions as variables and sum over or take the product of random variables. The word "random" here is that we are finding the numerical of a "random event". The image of a random variable is called the *weight* or the *bias*, depending on what data the random variable is representing.

Example 1.3: Random Variable

Let's say $\Omega = \{a, b, c, d\}$ with the probability of choosing any event of Ω being $1/4$ ($\mathbb{P}(x) = 1/4$, $x \in \Omega$). Define $\varepsilon : \Omega \to \mathbb{R}$, $\varepsilon(a) = \varepsilon(b) = 1$ and $\varepsilon(c) = \varepsilon(d) = -1$. Then we have:

$$\mathbb{P}(\{\omega \in \Omega \mid \varepsilon(\omega) = 1\} = 1/2$$

and similarly if $\varepsilon(\omega) = -1$.

We can define a different random variable on Ω: let $Z : \Omega \to \mathbb{R}$, $Z(a) = 2,\ \mathbb{Z}(b) = -2,\ Z(c) = 3,\ \mathbb{Z}(d) = -3$. Then

$$\mathbb{P}(\{\omega \in \Omega \mid Z(\omega) = 2\} = 1/4$$

and similarly for all other possible values. The sets in both cases are very commonly abbreviated as:

$$\{\varepsilon = 1\} = \{\omega \in \Omega \mid \varepsilon(\omega) = 1\}$$

so that we would write

$$\mathbb{P}(\{\varepsilon = 1\}) = 1/2$$

or even be even shorter with our notation and write:

$$\mathbb{P}(\varepsilon = 1) = 1/2$$

This final notation will be very common when using random variables, as it really holds the key information about how the random variable interacts with the probability.

As mentioned earlier, one of the most sought-after information we try to get from probability is the average. Average is dependent on how we numerically represent the values of the probability space, which should motivate the following definition:

Definition 1.2.3: [Finite] Expected Value

Let X be a random variable and $(\Omega, \mathbb{P})$ be a probability space. Then the expected value for the random variable X is:

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$$

Note that this is well-defined since Ω is finite. In following chapter's, we will find conditions that need to be met so that the expected value is well-defined. In particular, in chapter ref:HERE, we will work with continuum's, at which point we will be more precise with our definition of expected value. Sometimes, the notation for the discrete and continuous case is used interchangibally:

$$\mathbb{E}X = \int_{\Omega} X(\omega) d\mathbb{P}(\omega)$$

The way I like to think of the expected value is that it associated a number to each function, almost like the integral! In fact, in the continuous case, the integral *will* replace the summation in the expected value formula!

Example 1.4: Expected Value: Indicator Function

let Ω be any probability with $\mathbb{P}$ as it's probability function. Define the following random variable:

$$I_A \overset{!}{=} I(\omega \in A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

this random variable is called the *indicator function*. The $\overset{!}{=}$ equality is an abuse of notation to mean the function I is "indicating" over the set A, since the indicator function is very common.

The expected value is:

$$\mathbb{E}(I_A) = \sum_{\omega \in \Omega} I_A(\omega)\mathbb{P}(\omega) = \sum_{\omega \in A} \mathbb{P}(\omega) = \mathbb{P}(A)$$

showing that the expected value of I_A is the probability of the event of A! Take a moment to think about this case to have an idea of what a random variable is capturing

Proposition 1.2.1: Properties of Expected Value

Let f and g be random variables on $(\Omega, \mathbb{P})$ and $|\Omega| < \infty$. Then

1. $\mathbb{E}(f + g) = \mathbb{E}f + \mathbb{E}g$
2. If $f \geq 0$, then $\mathbb{E}f \geq 0$
3. More generally, if $f \leq g$ then $\mathbb{E}f \leq \mathbb{E}g$

***Proof* :**

exercise!

The next concept we shall point out is that of *independence*. When defining probability, I mentioned in the paragraph above the definition that we have to somehow deal with events affecting one another. Similarly, we have to know how to deal with events that do not affect one another, that are independent of another. The former we shall return to. For the latter, notice that we were already sneakily using the notion of independent events in the motivating example. In particular, the choice of adding or subtracting one vector did not affect the choice of another. The way we will capture this idea of independence is through the use of random variables. Notice that we can define random variables

$$\pi_i : \Omega \to \mathbb{R}, \ \pi_i(\varepsilon) = \varepsilon_i$$

every random variable π_i captures the idea of the sign of the ith vector. Often, we abuse notation and write ε_i as the random variable itself. Given these random variables, we can define an outcome from them (or more generally, an event). For some choice of signs $a_1, a_2, ..., a_n$, we can define:

$$A = \{\varepsilon \in \Omega \mid \pi_1(\varepsilon) = a_1, ..., \pi_n(\varepsilon) = a_n\}$$

or, to abuse notation a bit, we write

$$A = \{\pi_1 = a_1, ..., \pi_n = a_n\}$$

where we omit a reference to Ω or it's elements. It is a very common convention in statistics to omit Ω as much as possible when working with random variables (as we've already demonstrated in example 1.3) , either relying on the context to give us the probability space, or more generally not really caring about the particular probability space; it doesn't matter if it's 6 face of a die or or 6 cosmic events, if each have a random variable defined identically with the same probability (ex. each face of the die and each cosmic even have the same probability, and the random variable gives either number of the face or the intensity of the cosmic event from 1-6), then what we really care about is the result from the random variable and it's interaction with the probability, and not the probability space. In fact, the probability space can be different in size and we still can define the same random variables. Let's say there were 12 cosmic events, each with the same probability, and two of those events have intensity 1, two of those events have intensity 2, and so on. Then a random variable that captures the intensity will be indistinguishable to the one with the dice roll.

Example 1.5: Event with new Notation

Taking the same Ω and $\mathbb{P}$ as from the motivating example, define

$$B = \{\pi_1 = 1\}$$

Then $\mathbb{P}(B) = \frac{2^{n-1}}{2} = \frac{1}{2}$

The set A was a bit contrived to only have one outcome, however it is an excellent example of how to define in independence

Definition 1.2.4: Independence

Let $(\Omega, \mathbb{P})$ be a probability space with random variables $X_1, X_2, ..., X_n$. Then the random variables are said to be *independent* if

$$\mathbb{P}(\{X_1 = x_1, ..., X_n = x_n\}) \overset{!}{=} \mathbb{P}(X_1 = x_1, ..., X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i)$$

where $\overset{!}{=}$ is an abuse of notation where we drop the curly braces. If the above equation does not hold, then these variables are said to be *dependent*.

Example 1.6: independence

Let $\Omega = [0,1] \times [0,1]$ and let the probability be the integral over the area[a]. We can let f and g be continuous[b] functions from $[0,1] \to \mathbb{R}$. You can imagine them being random variables on Ω by extending them to $[0,1] \times [0,1]$ by mapping those points to zero. Then we know from calculus that

$$\iint\limits_{[0,1]^2} f(x)g(y)dxdy = \int_{[0,1]} f(x)dx \int_{[0,1]} g(y)dy$$

in other words, the given the "length" of one side, it is independent of the "length" of the other; each contribute independently to the area.

[a] note that this is an example where we would limit to measurable sets instead of the powerset of $[0,1]$

[b] more generally, measurable

Exercise 1.2.1

1. Let Ω be any probability space with $|\Omega| \geq 2$. Let $A, B \subseteq \Omega$ where $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$. If $A \cap B = 0$, does that imply they are independent?

Using new Vocabulary and Change of Variables

Let us re-write the motivating example with the new notation and definitions we just presented. We will re-write step ref:HERE in example 1.1 as:

$$\mathbb{E}\pi_i\pi_j \overset{\text{indep.}}{=} \mathbb{E}\pi_j\mathbb{E}\pi_j = 0$$

To show this, we simply plug into the equation:

$$\mathbb{E}\pi_i\pi_j = \sum_{\varepsilon\in\Omega} \pi_i(\varepsilon)\pi_j(\varepsilon)\mathbb{P}(\varepsilon)$$

This can be hard to compute, so we will re-write this in a different way. This is a very common strategy in statistics; many times, writing the same sum in a different way will produce much simpler summation, sometimes drastically simpler (as we'll see with distributions in the next section).

The process by which we change the summation is called *change of variables*. To do this, we will partition Ω through the random variable in the following way:

$$A_{a_i,a_j} = \{\varepsilon \in \Omega \mid \pi_i(\varepsilon) = a_i,\ \pi_j(\varepsilon) = a_j\}$$

Notice that

$$\Omega = \bigsqcup_{a_i,a_j} A_{a_i,a_j}$$

Thus, we can split up our sum which was over every element $\varepsilon \in \Omega$ in the following way:

$$\begin{aligned}
\mathbb{E}(\pi_i\pi_j) &= \sum_{\varepsilon\in\Omega} \pi_i(\varepsilon)\mathbb{P}(\varepsilon) \\
&= \sum_{a_i,a_j=\pm1} a_ia_j \sum_{\varepsilon\in A_{a_i,a_j}} \mathbb{P}(\varepsilon) \\
&= \sum_{a_i,a_j=\pm1} a_ia_j\mathbb{P}(A_{a_i,a_j}) \\
&= \sum_{a_i,a_j=\pm1} a_ia_j\mathbb{P}(\pi_i = a_i, \pi_j = a_j)
\end{aligned}$$

We have now re-written this sum to look like another expected value. In fact, this *is* an expected value over a different probability space! We will get back to this shortly, let us first finish the computation. By independence, we get

$$\sum_{a_i,a_j=\pm1} a_ia_j\mathbb{P}(\pi_i = a_i, \pi_j = a_j) = \sum_{a_i=\pm1} a_i\mathbb{P}(\pi_i = a_i) \sum_{a_j=\pm1} a_j\mathbb{P}(\pi_j = a_j)$$

Furthermore, by similar calculations, we can show that

$$\begin{aligned}
\mathbb{E}(\pi_i) &= \sum_{\varepsilon \in \Omega} \pi_i(\varepsilon)\mathbb{P}(\varepsilon) \\
&= \sum_{a_i = \pm 1} a_i \sum_{\varepsilon \in A_{a_i}} \mathbb{P}(\varepsilon) \\
&= \sum_{a_i = \pm 1} a_i \mathbb{P}(A_{a_i, a_j}) \\
&= \sum_{a_i = \pm 1} a_i \mathbb{P}(\pi_i = a_i)
\end{aligned}$$

showing that “independence” equality does indeed hold! For the last equality, we can proceed with the same reasoning we gave in the example. We can split the sum into the ones that equal -1 and those that equal to 1. There are an equal number in both cases, giving us:

$$(+1)\frac{1}{2} + (-1)\frac{1}{2} = 0$$

finishing the proof, but perhaps more clearly with this change of variables.

As we mentioned, we can think of this new equation as changing the probability space we working over. This new probability space is given by a random variable we are using to calculate our results. To best illustrate this, take π_i from the previous example. Then the image of π_i is $\{-1, 1\}$. Let $\Omega = \{-1, 1\}$ be our new probability space, and let our new probability measure on Ω' be:

$$\mathbb{P}'(\omega') = \mathbb{P}(\{\omega \in \Omega \mid \pi_i(\omega) = \omega'\})$$

This new probability measure is often called the *distribution* (sometimes called the *law*) of the random variable π_i (more generally the random variable X), or the *image measure of* $\mathbb{P}$ *by the map* π_i (more generally by the map X). For example, Given π_i, we have the new values of:

$$\mathbb{P}'(1) = \frac{1}{2} \qquad \mathbb{P}'(-1) = \frac{1}{2}$$

This new probability space is called the *sample space*. I like to think of it as the sample of *data* that we can get given an outcome, event. A bit confusingly, some statisticians called the original probability space Ω a sample space as well. It might be useful to think of the “default” random variable as $X : \Omega \to \Omega$, $X(\omega) = \omega$ (i.e. the identity), even though it doesn't necessarily map to $\mathbb{R}$ and the domain is not the powerset.

Notice that when we did a change of variables, and calculated the estimate over what now can be thought of as Ω', we have a “new” random variable which is the identity function! In the previous example, we would have $\Omega' = \{-1, 1\}$, and our new probability space would be $(\Omega', \mathbb{P}')$! We give a name to this new probability:

Definition 1.2.5: Image Measure

et $(\Omega, \mathbb{P})$ be a probability space an let X be a random variable. Define $\Omega' = \{X(\Omega)\}$ and let $\mathbb{P}'(x) = \mathbb{P}(X = x)$. Then Ω' is called the *sample space* and $\mathbb{P}'$ is called the *image measure* of Ω' with respect to X.

Many notations exist to represent the new probability on a sample space, including:

$$\mathbb{P}'(a_i) = \mathbb{P}(\pi_i = a_i) = \mathbb{P}(\pi_i^{-1}(a_i)) = \mathbb{P} \circ \pi_i^{-1}(a_i) = X\#\mathbb{P}(a_i)$$

that last one is read as "X push-forward $\mathbb{P}$".

We sum up our discoveries on expected values with change of variables the following proposition:

Proposition 1.2.2: [Finite] Expected Value Change of Variables

Let f be a random variable and $\mathbb{P}(x)$ be a probability. Let Ω' and $\mathbb{P}'$ be the sample space and image measure of the map f. Then the expected value for the random variable f can be represented as :

$$\mathbb{E}f = \sum_{\omega' \in \Omega'} \omega' \mathbb{P}'(\omega')$$

which is called the *change of variables formula for the expectation*

A word of warning to avoid confusion. We will often apply functions to random variables. In general, for a random variable X, if we apply a function to it, say $f(X)$, then $\mathbb{P}(X = x)$ doesn't change *if the bounds of the sum are the same* (i.e. if the sample space remains the same)

$$\mathbb{E}f(X) = \sum_{x \in X(\Omega)} f(x)\mathbb{P}(X = x)$$

If we change the sample space to be $f(X(\Omega))$, then we would write:

$$\mathbb{E}f(X) = \sum_{y \in f(X(\Omega))} y\mathbb{P}(X = y) = \sum_{f(x) \in f(X(\Omega))} f(x)\mathbb{P}(X = f(x))$$

Each of these can be useful in their own way.

Notice too that we can now simply define a random variable and it's interaction with the probability without referencing the probability space since the random variable will define a sample space:

Example 1.7: Practice

Let N be a non-negative random variable such that:

$$\mathbb{P}(N > 0) = e^{-1} \qquad \mathbb{E}N = e^{10}$$

Find a random variable with this property (in particular, notice that you can define one that "works", and then the same space will become the probability space)

1.3 Discrete Distributions

In this section, we will start by focusing on the discrete case, that is, where Ω is either finite or countable (i.e. they can be indexed by the natural numbers). When working in the discrete case, we will define the probability of any given outcome:

$$\mathbb{P}(\omega_n) = p_n \qquad 0 \leq p_n \leq 1$$

with the restriction that

$$\sum_{i=1}^{\infty} \mathbb{P}(\omega_i) = 1$$

to insure that $\mathbb{P}(\Omega) = 1$. Given that the probability of every term is defined, by linearity of the probability measure, it is easy to see that

$$A \subseteq \Omega, \ \mathbb{P}(A) = \sum_{a \in A} \mathbb{P}(a)$$

Another way of writing this equation would be to use the indicator function we saw before:

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(\omega_i) I(\omega_i \in A)$$

which as we saw before is $\mathbb{E}(I_A)$. Using this fact, all the properties of probabilities in example 1.2 will be satisfied.

1.3.1 Countable Expected Value

As mentioned earlier, the expected value might need a couple of conditions to be well-defined; since $X : \Omega \to \mathbb{R}$ can have both positive and negative values, we must be careful on the convergence condition. In particular, since we're now taking infinite sums, we will have to start being careful with the order of the elements over which we sum (ex. take the series $1 - 1/2 + 1/3 - \cdots \pm 1/n \mp \cdots$, and re-arrange it to sum to any number $r \in \mathbb{R}$). Because of this, we give the following restriction:

Definition 1.3.1: [Countable] expected value

Let Ω be a discrete probability space with probability measure $\mathbb{P}$ and let X be a random variable. Then the expected value of X is the same way as the finite case:

$$\mathbb{E}X = \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)$$

but with the additional condition that

$$\mathbb{E}|X| = \sum_{\omega \in \Omega} |X(\omega)|\mathbb{P}(\omega) < \infty$$

i.e. the series as absolutely convergent. Otherwise, we say that $\mathbb{E}X$ is undefined.
An expected value with this property is called *integrable*

In these notes, we will consider $\mathbb{E}X = \infty$ as an acceptable value. This means that the first thing you have to check when working with an expected value of a random variable over a discrete probability space is whether or not it's defined. In the following section, we will prove some technical lemmas which help us in computing or proving the existence of an expected value, and we'll show a couple of different ways we can view the expected value (something very useful to get a better grasp of the nature of the expected value)

Since we're working with countable probability spaces, these lemmas will be very handy to keep in mind:

Lemma 1.3.1: Equivalent Convergence Condition

Let $(a_i)_{i\geq 1}$ be a sequence of non-negative values. Then

$$\lim_{N\to\infty}\sum_{i=1}^{n} a_i = \sup_{\substack{F\subseteq\mathbb{N}\\ |F| \text{ finite}}} \sum_{i\in F} a_i$$

Proof **:**
Prove $\leq$ and $\geq$. Should be able to replicate.

A consequence of this lemma, if a series is absolutely convergent, then any permutation of the terms in the series produces converges to the same value (by the previous lemma, any permutation is upper-bounded by the limit on the right hand side, but that is equal to the original sum). Importantly choice of "order" doesn't matter, so we don't have to think about defining some order when taking the expected value of a random variable.

An equivalent to lemma 1.3.1 is the following

Corollary 1.3.1

For any bijection $\pi : \mathbb{N}\times\mathbb{N}\to\mathbb{N}$,

$$\sum_{n=1}^{\infty} c_n = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} c_{\pi(n,m)}$$

if either $c_n \geq 0$ for all $n\in\mathbb{N}$ or either side is absolutely convergent

Proof **:**
Using lemma 1.3.1, this corollary can quickly be proven, and so an alternative proof is given that was presented in Lachenko's book

Let's first consider the non-negative case. We'll show they're equal by double-inequality ($\leq$ and $\geq$). Let's first show $\leq$. Pick some K and consider

$$\sum_{k=1}^{K} c_n$$

Since π is bijective, there must exist some N where the terms on the left appear in the right in the following summation:

$$\sum_{i=1}^{k} c_N = \sum_{n=1}^{N}\sum_{m=1}^{N} c_{\pi(n,m)}$$

Since this is true for all K as $K\to\infty$, this shows the $\leq$ part of the proof.

For $\geq$, pick some numbers N and M, then consider:

$$\sum_{n=1}^{N}\sum_{m=1}^{M} c_{\pi(n,m)}$$

Then, let $K = \max\{\pi(n,m) \mid 1 \leq n, m \leq N, M\}$. Then we have:

$$\sum_{n=1}^{N}\sum_{m=1}^{M} c_{\pi(n,m)} \leq \sum_{k=1}^{K} c_k \leq \sum_{k=1}^{\infty} c_k$$

Now, let $N \to \infty$ and then $M \to \infty$ (or the other way around), and we can see that this inequality always holds. This shows the $\geq$ part of the proof, and so completing the proof in for the non-negative case.

For the absolutely convergent case, we will take advantage of what we just proved by writing the absolutely convergent case in a similar form to the first case. Suppose $\sum_{k=1}^{\infty} c_k$ is absolutely convergent. Then every term can be re-written as

$$c_n = a_n - b_n \qquad a_n = c_n \cdot I(c_n \geq 0),\ b_n = c_n \cdot I(c_n \leq 0)$$

i.e. we use an indicator function to add the positive term and subtract all the negative terms (i.e., change the sign of the term). Note that

$$\sum_{k=1}^{\infty} a_k \leq \sum_{k=1}^{\infty} |c_k| < \infty \qquad \sum_{k=1}^{\infty} b_k \leq \sum_{k=1}^{\infty} |c_k| < \infty$$

meaning both terms converge. Furthermore: Then

$$\sum_{k=1}^{\infty} a_k = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_{\pi(n,m)} \text{ and } \sum_{k=1}^{\infty} b_k = \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} b_{\pi(n,m)}$$

(finish here)

As an immediate consequence, you can check that countable additivity works.

Proposition 1.3.1: Linearity of Expected Value

Let X and Y be random variables, and let $\mathbb{E}X$, $\mathbb{E}Y$ be defined (i.e. $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$). Then

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

This proposition should be an easy application of the previous lemma.

Proposition 1.3.2: Countable Change of Variables

Let X be a random variable and $\mathbb{E}X$ be well-defined. Then the change of variables formula applies to $\mathbb{E}X$, that is

$$\mathbb{E}X = \sum_{\omega\in\Omega} X(\omega)\mathbb{P}(\omega) \overset{!}{=} \sum_{n\geq 1} a_n \mathbb{P}(X = a_n) = \sum_{n\geq 1} a_n \mathbb{P}'(a_n)$$

where the $\overset{!}{=}$ is the equality we are proving in this proposition.

***Proof* :**
This proof is essentially checking that the condition on $\mathbb{E}X$ allows us to move around and combine sums without any difficulty. To start with, enumerate the image of X so that $X(\Omega) = \{a_1, a_2, ..., a_n, ...\}$. This image can be finite or countable. Next, let's give an enumeration to the values of $X^{-1}(a_n)$:

$$X^{-1}(a) = \{\omega \in X \mid X(\omega) = a\} = \{\omega_{nm} \mid 1 \leq m \leq M_n\}$$

Note that some (or all) of the M_n's can be infinite. With this enumeration, we will start with the right hand side and work our way towards the definition of $\mathbb{E}X$:

$$\begin{aligned}\sum_{n\geq 1} a_n \mathbb{P}(X = a_n) &= \sum_{n\geq 1} a_n \sum_{m=1}^{M_n} \mathbb{P}(\omega_{nm}) \\ &= \sum_{n\geq 1} \sum_{m=1}^{M_n} a_n \mathbb{P}(\omega_{nm}) && \text{by proposition 1.3.1} \\ &= \sum_{n\geq 1} \sum_{m=1}^{M_n} X(\omega_{nm}) \mathbb{P}(\omega_{nm})\end{aligned}$$

At this point, we can apply lemma 1.3.1, and we get that

$$= \sum_{n\geq 1} a_n \mathbb{P}(X = a_n)$$

as we sought to show

Yet another way of calculating the expected value is by calculating what happened at possible "tail end" without weighing the value of the probability:

Definition 1.3.2: Tail Probability

Let $\mathbb{P}$ be a probability on a probability space and X be a random variable. Then

$$\mathbb{P}(X \geq t) = \sum_{k\geq t} \mathbb{P}(X = k)$$

Since the probability is countably additive, it is easy to see that (exercise!)

$$\lim_{t\to\infty} \mathbb{P}(X \geq t) = 0$$

Using tail probabilities, we have yet another way or writing the expected value:

Proposition 1.3.3: Expected Value as Tail Probability

Let X be a random variable and $\mathbb{E}X$ be well-defined. If $X(\Omega) \subseteq \mathbb{N}$, then for any $m \in \mathbb{N}$

$$\mathbb{E}X = \sum_{m=1}^{\infty} \mathbb{P}(X \geq m)$$

If the image of X is not contained in $\mathbb{N}$, then we can write

$$\mathbb{E}X = \int_0^{\infty} \mathbb{P}(X \geq t)dt$$

In Analysis, this is called the “layer-cake representaiton”.

***Proof* :**
By proposition 1.3.2, we can state with the change of variables representation of the expected value:

$$\mathbb{E}X = \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \sum_{k=1}^{\infty}\sum_{m=1}^{k} \mathbb{P}(X = k) \overset{\text{cor 1.3.1}}{=} \sum_{m=1}^{\infty}\sum_{k=m}^{\infty} \mathbb{P}(X = k) = \sum_{m=1}^{\infty} \mathbb{P}(X \geq k)$$

Equality $\overset{\text{cor 1.3.1}}{=}$ can be visualized as summing horizontally instead of vertically in the following diagram:

$$\begin{array}{cccc} & & & \iddots \\ & & \mathbb{P}(X = 3) & \cdots \\ & \mathbb{P}(X = 2) & \mathbb{P}(X = 3) & \cdots \\ \mathbb{P}(X = 1) & \mathbb{P}(X = 2) & \mathbb{P}(X = 3) & \cdots \end{array}$$

For the case where X doesn’t just map to the integers, then we need to use the integral. Like last time, we can start with the change of variable representation of the expected value

$$\mathbb{E}X = \sum_{n \geq 1} a_n \mathbb{P}(X = a_n)dt$$

We can now re-write a_n as $\int_0^{\infty} I(t \leq a_n)$, giving us

$$\begin{aligned} &= \sum_{n \geq 1} \int_0^{\infty} I(t \leq a_n)\mathbb{P}(X = a_n)dt \\ &\overset{\text{M.C.T}}{=} \int_0^{\infty} \sum_{n \geq 1} I(t \leq a_n)\mathbb{P}(X = a_n)dt \\ &= \int_0^{\infty} \sum_{a_n \geq t} \mathbb{P}(X = a_n)dt = \int_0^{\infty} \mathbb{P}(X \geq t)dt \end{aligned}$$

Where we use the Monotone Convergence Theorem in the second equality. If you do not know the M.C.T., please refer to Pachenko’s book on p. 17 to see a proof using the Riemann integral.

Example 1.8

1. Show that for any nonnegative random variable $X \geq 0$ and $n \geq 0$ that

$$\mathbb{E}X^n = \int_0^\infty nt^{n-1}\mathbb{P}(X \geq t)dt$$

and if X is any random variable, then:

$$\mathbb{E}X^n = \int_0^\infty nt^{n-1}\mathbb{P}(|X| \geq t)dt$$

[hint: expand the definition, you'll get $\mathbb{P}(X^n \geq t)$, do an [integral] change of varibles]

2. Show that if $c, \varepsilon > 0$ $t_0 \geq 0$,

$$\mathbb{P}(|X| \geq t) \leq \frac{c}{t^{2+\varepsilon}} \quad \text{for all } t \geq t_0$$

1.3.2 Example of Discrete Distributions

In this section, we will go over some common distributions and their representations. In particular, we will review

1. Bernoulli Distribution
2. Product Bernoulli Distribution
3. Binomial Distribution Distribution
4. Poisson Distribution

Each of these distributions will be a quantification of a "probability question", and essentially all of them will can be formulated as a question about coin flips!

We will also explore some examples of where these distribution are useful (like the Erdós-Rényi random graph).

Bernoulli Distribution

Perhaps the most common probabilistic scenario is that of a coin flip: whether an outcome will or will not happen:

Definition 1.3.3: Bernoulli Distribution

Let $\Omega = \{0, 1\}$. Then for some $p \in [0, 1]$

$$\mathbb{P}(0) = p, \ \mathbb{P}(1) = 1 - p$$

Then $\mathbb{P}$ is called the *Bernoulli Distribution* and is denoted $B(p)$.

Notice that since the probability space is in $\mathbb{R}$, we are justified in calling the probability measure a distribution. The Bernoulli Distribution is essentially representing a coin toss. In fact, essentially all the following distributions can be thought of as asking different types of questions about outcome of repeated coin tosses (as we will show)! In a sense, this is the foundational scenario for statistics. If $p = 1/2$, then we say it's a fair coin, and an unfair coin if $p \neq 1/2$.

As we will soon be encountering, we are often more interested in random variables that take on the role of such a distribution. In a sense, our random variable will now represent a "question" we are asking about our probability space:

Definition 1.3.4: Bernoulli Random Variable

Let $(\Omega, \mathbb{P})$ be a probability space, and let $p \in [0,1]$. Then if X is a random variable on Ω such that

$$\mathbb{P}(X = 1) = p \qquad \mathbb{P}(X = 0) = 1 - p$$

Then X is called a *Bernoulli Random Variable* and is denoted $X \sim B(p)$.

For every distribution we will introduce a random variable equivalent will be given too (i.e. the sample space and image measure will be the desired distribution). If $\mathbb{P}$ is an X distribution (where X is the name of a distribution, like Bernoulli) then the identity function $X(\omega) = \omega$ is always a X random variable.

Example 1.9: expected value of $X \sim B(p)$

Since $p + 1 - p = 1$, we only need to sum over 1 and 0. Since this is a finite random variable, there is no special convergence condition that needs to be checked, and so:

$$\mathbb{E}X = 1 \cdot p = 0 \cdot (1 - p) = p$$

A single Bernoulli Distribution gives us the value for a single coin toss, but this might be limiting us in what we can measure, in particular, in the fact that we are often measuring many events at once and want to know the probability of a particular result given the, say, n coins. For that, we introduce the product of Bernoulli Distributions and Variables:

Definition 1.3.5: Product of Bernoulli

Let $p_1, ..., p_n \in [0,1]$. And let $\Omega = \{0,1\}^n$. Then if the probability measure is defined as

$$\mathbb{P}(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} p_i^{x_i}(1-p)^{1-x_i}$$

Then $\mathbb{P}$ is called the *product Bernoulli Distribution* and is denoted $\bigotimes_i^n B(p_i)$.

This is indeed a well-defined probability measure. Notice that

$$\sum_{x \in \{0,1\}^n} \prod_{i=1}^{n} p_i^{x_i}(1-p)^{1-x_i} = \prod_{i=1}^{n} \sum_{x \in \{0,1\}^n} p_i^{x_i}(1-p)^{1-x_i} = \prod_{i=1}^{n} 1 = 1$$

and the rest of the conditions become easy to check.

Since we are working over a "multi-dimensional" universe, we will sometimes want to have a random vriable for "each dimension". We can define a random variable for each component, thoughnote that we must be careful on how to make it only work on 1 component, since a random variable on Ω^n must have as domain Ω^n. This might get notationally cumbersome, so to help with notation, we introduce a new concept:

Definition 1.3.6: Random Vector

Let $(\Omega^n, \mathbb{P})$ be a probability space. Then a *random vector* $X : \Omega^n \to \mathbb{R}^n$, $X = (X_1, ..., X_n)$ is an n-tuple of random variables $X_i : \Omega^n \to \mathbb{R}$, and if $A \subseteq \Omega^n$ is an event, and $A_i = \pi_i(A)$ where π_i is the ith coordinate map, then:

$$\mathbb{P}(X \in A) = \mathbb{P}(\{x \mid X_1 \in A_1, ..., X_n \in A_n\})$$

Note that it need not be that it is the same Ω every time, however it is convinient at the beginning of probability theory to stick to this special case to get a grasp of notions like indepndence or defning multiple random variables on the same space. We will only much later cover examples where we might have different spaces (in particular, we will have $\mathbb{R} \times \{0, 1\}$ in section ref:HERE).

Example 1.10: Product Of Bernoulli Random Vector

If $X = (X_1, X_2, ..., X_n)$ is a random variable where each component is the random variable X_i, then if

$$\mathbb{P}(X = x) = \prod_{i=1}^{n} p_i^{x_i} (1 - p_i)^{x_i}$$

then $X \sim \bigotimes_i^n B(p_i)$. We will call X the *random vector*. By the way we have defined the product of Bernoulli variables, it is tempting to say that by definition:

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = \prod_i^n \mathbb{P}(X_i = x_i)$$

and so the variables $X_1, X_2, ..., X_n$ are independent. However, we have to be more careful than that. In particular, notice that $B(p_i)$ not defined on $\{0, 1\}^n$ but on $\{0, 1\}$. Therefore, we must be more careful with our formulation. To simplify the proof for the moment, let's say $(\{0, 1\}^n, \mathbb{P})$ is a probability space with $\mathbb{P}$ being the product Bernoulli Distribution. The natural definition of the Bernoulli random variable on the ith coordinate X_i as a random variable on $\{0, 1\}^n$ would be

$$X_i : \Omega \to \mathbb{R} \qquad X_i(x_1, x_2, ..., x_i, ..., x_n) = x_i$$

i.e. the identity function on the ith coordinate. Then the sample space of X_i is equivalent to $\{0, 1\}$ and the $X_i \sim B(p)$! To see this, simply go through the computations remembering that the countable probability measure is additive. For notational simplicity of the change of variables, let

$A_i = \{\omega \in \Omega \mid X_i(\omega) = x_i\}$. Then:

$$\begin{aligned}
\mathbb{P}(X_i = x_i) &= \sum_{a \in A_i} \mathbb{P}(X_i = a) \\
&= \sum_{a \in A_i} \prod_{i=1}^{n} p_i^{a_i}(1-p_i)^{1-a_i} \\
&\overset{!}{=} p_i^{x_i}(1-p_i)^{1-x_i} \sum_{a \in A} \prod_{i=1}^{n} p_j^{a_i}(1-p_j)^{1-a_i} \\
&= p_i^{x_i}(1-p_i)^{1-x_i} \prod_{j \neq i} \underbrace{\sum_{a_i \in \{0,1\}} p_j^{a_i}(1-p_j)^{1-a_i}}_{1} \\
&= p_i^{x_i}(1-p_i)^{1-x_i}
\end{aligned}$$

showing us that our earlier naïve formula for independence was in fact accurate.

Before moving on to our next distribution, a quick word must be set if $p_1 = p_2 = \cdots = p_n$ are all equal. Then each random variable $X_1, X_2, ..., X_n$ will be identically distributed. Furthermore, since they are all independent, they are all *independently identically distributed* random variables (or *i.i.d* variables for short). When this is the case, the equation in definition 1.3.5 simplifies to:

$$\mathbb{P}(X = x) p^{\sum_{i=1}^{n} x_i}(1-p)^{n - \sum_{i=1}^{n} x_i}$$

To emphasize the difference, we usually denote such a distribution with $B(p)^{\otimes n}$. This random variable will be much more common in these notes thanks to it's application to the following important example:

Example 1.11: Erdós-Rényi Random Graph

Let $p \in [0, 1]$ represent the probability of a coin landing on heads, and let $V = \{v_1, v_2, ..., v_n\}$ be a set of elements called vertices. Then for each possible pair (v_i, v_j), we will flip a coin, and if it lands on heads, we add that edge to our graph. If it lands on tails, we do not add the edge. The resulting randomly generated graph is denoted $G(n, p)$.

Since each coin toss was independent and identical, then the chances of getting any one graph is the product of i.i.d. Bernoulli variables

$$(X_1, X_2, ..., X_n) \sim B(p)^{\otimes m} \qquad m = \binom{n}{2}$$

where m represents the possible number of vertices (these vertices are undirected).

These graphs have multiple application in sociology in modeling networks of people given some appropriate conditions or simplifications. We will reference back to this graph multiple times when developping new tools to see what properties we can find from it!

Binomial Distribution

Let's say we have a coin with probability of landing on heads is p. If we flip the coin n times, what is the probability that we get k heads? To represent this problem, take $\Omega = \{0,1\}^n$ to be the n coin flips. We want our probability measure $\mathbb{P}$ to measure the amount of heads we land on. As an exercise, think though why the following formula represents this scenario

Definition 1.3.7: Binomial Random Variable

Let $(\Omega, \mathbb{P})$ be a probability space, and let X be a random variable on Ω. Then if

$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ 0 \text{ if } k > n?$$

then we call X a *Binomial Random Variable*, and is denoted $X \sim B(n,p)$.

As an exercise, can you define the Binomial distribution; you have to define an explicit Ω and $\mathbb{P}$. The name of the formula is inspired by the binomial formula:

$$(a+b)^n = \sum_{k=1}^{n} \binom{n}{k} a^k b^{n-k}$$

We can in fact use this formula to show that this is a well-defined distribution, since

$$\sum_{k=1}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + 1 - p)^n = 1^n = 1$$

Given the definition, this random distribution should very naturally connect to the Bernoulli random variable, and indeed it does. If $X_1, X_2, ..., X_n$ are i.i.d. Bernoulli Random Variables, let

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

The sample space of S_n is $\{0,1,...,n\}$. Next, we'll show $S_n \sim B(n,p)$. First, notice that the set $\{S_n = k\}$ can be broken down into $\bigsqcup \{X = (X_1, X_2, ..., X_n) = x\}$ for all $x \in \{0,1\}^n$ that give the value of k. Then each $X = x$ is just a Product Bernoulli random variable, so

$$\mathbb{P}(X = x) = p^k (1-p)^{n-k}$$

Since there are $\binom{n}{k}$ possibilities, by countable additivity, we get:

$$\mathbb{P}(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

showing that $S_n \sim B(n,p)$.

Example 1.12: expected value of Binomial

Let $X \sim B(n,p)$. What is $\mathbb{E}(X)$?

Solution There are multiple ways of approaching this question. As usual in probability, it is important to have a good intuition to be able to spot an easy resolution. In this case, by what we've just proved, $X \sim S_n$ for appropriate i.i.d. Bernoulli Random variables $X_1, X_2, ..., X_n$. Then by proposition 1.3.1, the expected value is linear, so

$$\mathbb{E}X = \mathbb{E}S_n = \mathbb{E}(X_1 + X_2 + \cdots + X_n) = \mathbb{E}X_1 + \mathbb{E}X_2 + \cdots + \mathbb{E}X_n = np$$

▼

(Instead of two posibilites, n possibilities, each with prob. p_i and $\sum_i p_i = 1$. Then the multinomial formula will be used to define the multinomial distribution, which is defined since $\left(\sum_i p_i\right)^n = 1^n = 1$

Geometric Distribution

What if we ask when the first head appears? Don't forget some words on adding the "never" case, which can be thought as adding a 0 at the end, but the probability of hitting it is 0, and so we can omit it

Definition 1.3.8: Geometric Distribution

Let $(\Omega, \mathbb{P})$ be a probability space, let $p \in [0, 1]$, and let N be a random variable. Then if

$$\mathbb{P}(N = n) = (1 - p)^{n-1}p$$

we call N a *Geometric Random Variable*

The geometric random variables is not seen as often and so it usually doesn't have a special symbol like $B(p)$ or $B(n, p)$ associated to it. Some authors sometimes use $X \sim T(n)$ to represent a Geometric Random Variable. Like the binomial random variable, this random variable can also be represented in terms of the Bernoulli random variable. In particular, If $\{N = k\}$ is our set, then we can define k i.i.d. Bernoulli random variables,

$$\{N = k\} = \{X_1 = 0, X_2 = 0, ..., X_k = 1\}$$

Then using the fact that the Bernoulli are independent, we see how this is identical to the geometric random variable.

Example 1.13: Geometric Expected Value

1. Prove that the geometric distribution is indeed a probability measure
2. What is the expected value of the geometric distribution? The answer is:[a]
3. What is $\mathbb{E}X^2$? The answer is [b]

[a] 1 over the probability of hitting heads
[b] the probability of hitting heads

Poisson Distribution

The Poisson distribution is one of the harder distributions to motivate compared to the other ones. It however comes up all over the place

Look at this link to get some examples (commented):

Here are some of my favourite uses:

1. The number of call per hour a call center recives: $\mathbb{P}(X = k)$ is Poisson
2. The number of expected restaurant clinets per day
3. Number of website visitors per hour
4. Number of Banrupcies filed per month
5. Number of Network failures per week
6. The pattern of Bombing of London during WWII (more on this in section 1.4.1)

Definition 1.3.9: Poisson Distribution

Let $(\Omega, \mathbb{P})$ be a probability space and $\lambda > 0$. Then if $\mathbb{P}$ satisfies:

$$\mathbb{P}(k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

for $k \in \mathbb{Z}_{\geq 0}$, $\mathbb{P}$ is called a Poisson Distribution. If X is a random variable on some probability space such that

$$\mathbb{P}(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

then X is called a *Poisson Random Variable*, and $X \sim \text{Poiss}(\lambda)$

As an exercise, prove that it is indeed a probability measure. It essentially comes down to the fact that $e^{-\lambda}$ is a constant that can be pulled out, and everything within that's left is the Taylor series of e^{λ}.

Example 1.14: Expected Value of Poisson

If $X \sim \text{Poiss}(\lambda)$, what is its expected value?

Solution Given:

$$\mathbb{E}X = \sum_{k=1}^{\infty} k\mathbb{P}(X = k) = \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda}$$

and manipulating dividing by k and manipulating the sum so that we get back to the poisson, we can sum it out and get:

$$\mathbb{E}X = \lambda$$

▼

For this reason, λ is usually called the *mean*. It also means that in some exercises you'll do, you will be given some information about the average occurence of an event. More often than not, that implies the Poisson distribution is the appropriate distribution you use to model it.

1.4 Poisson Properties

The Poisson distribution has many interesting properties, and so it's worth taking a moment to understand them more thoroughly.

Stability Property

The first property is a sort of "closure of distribution type" known as a *stability condition.* Let $(\Omega, \mathbb{P})$ be a probability space, and let X_1, X_2 be two i.i.d. Poisson Random Variables on $(\Omega,\ \mathbb{P})$: $X_1 \sim \text{Poiss}(\lambda_1)$, $X_2 \sim \text{Poiss}(\lambda_2)$. Since they're independent:

$$\mathbb{P}(X_1 = n,\ X_2 = m) = \mathbb{P}(X_1 = n)\mathbb{P}(X_2 = m) = \frac{\lambda^n}{n!}e^{-\lambda_1}\frac{\lambda^m}{m!}e^{-\lambda_2}$$

To see that we can construct such a pair of independent random variables, simply take $\Omega = \{0, 1, ...\} \times \{0, 1, ...\}$ and re-define X_1 and X_2 in the same way we have done for the product of Bernoulli random variablesMore generally, this tells us we can always take random variables $X_1, X_2, ..., X_n$ and say they are independent without any problem about whether such random variables exist.

Lemma 1.4.1: Stability of Poisson

let $X_1 \sim \text{Poiss}(\lambda_1)$, $X_2 \sim \text{Poiss}(\lambda_2)$ be two i.i.d. Poisson random variables . Then the sum $X = X_1 + X_2$ has $\text{Poiss}(\lambda_1 + \lambda_2)$ distribution:

$$X \sim \text{Poiss}(\lambda_1 + \lambda_2)$$

***Proof* :**
The following is a common trick for the sum of two random variables[a].

First, note that both X_1 and X_2 only have in $\mathbb{Z}_{\geq 0}$. Thus $X_1 + X_2$ only has values in $\mathbb{Z}_{\geq 0}$ too. If we have $X_1 + X_2 = n$, then $X_1 = m$, $X_2 = n - m$, for any $0 \leq m \leq n$. Therefore

$$\begin{aligned}
\mathbb{P}(X = x) &= \sum_{m=0}^{n} \mathbb{P}(X_1 = m,\ X_2 = n - m) \\
&= \sum_{m=0}^{n} \mathbb{P}(X_1 = m)\mathbb{P}(X_2 = n - m) && \text{independence} \\
&= \sum_{m=1}^{n} \frac{\lambda^m}{m!}e^{-\lambda_1}\frac{\lambda^{n-m}}{(n-m)!}e^{-\lambda_2} \\
&= \frac{1}{n!}\left(\sum_{m=0}^{n} \frac{n!}{m!(n-m)!}\lambda_1^m\lambda_2^{n-m}\right)e^{-(\lambda_1+\lambda_2)} \\
&= \frac{(\lambda_1+\lambda_2)^n}{n!}e^{-(\lambda_1+\lambda_2)}
\end{aligned}$$

where the last equality comes from the binomial formula. Thus $X \sim \text{Poiss}(\lambda_1 + \lambda_2)$, as we sought to show

[a] This trick is the first steps of *convolution*; we will talk about more about this in lemma 4.4.1

As an exercise, show that this generalizes to n i.i.d. Poisson variables (use induction)

1.4.1 Poisson Approximation of the Binomial

We know resume the goal of establishing how every distribution we present is similar in some way to the Bernoulli distribution. As we saw, the binomial distribution is the sum of i.i.d. Bernoulli distributions. The Poisson distribution is related to the binomial distribution in that the limit of a binomial distribution will approach a Poisson given the probability shrinks accordingly:

$$\text{Poiss}\left(\lambda\right) = \lim_{n\to\infty} B\left(n, \frac{\lambda}{n}\right)$$

i.e. $B\left(n, \frac{\lambda}{n}\right)$ converges "point-wise" to Poiss(λ). To see this, fix any k, and notice that

$$\begin{aligned}
&= \binom{n}{k} p^k (1-p)^{n-k} \\
&= \frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!}\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}\cdots\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&\to \frac{\lambda^k}{k!} 1\cdot 1\cdot\cdots\cdot 1\cdot e^{-\lambda} \\
&= \frac{\lambda^k}{k!}e^{-\lambda}
\end{aligned}$$

where it converges to that value since k is fixed, so each term in the middle converges to 1. If we look back at the list of examples given to motivate the Poisson distribution, this limit makes sense: the further "away" you go, the less likely it is that you hit the event. The bombing of London might be particularly insightful: the mathematicians were trying to determine whether or not the Germain where bombing randomly, or targetting specific areas. If there is a central location that is usually bombed, then as you move out radially from that location, and so the amount of area that *can* be bombed there goes up (which parrallels our $\frac{\lambda}{n}$), and so the probability will srhink accordingly. If the bombing were random, then we might expect no such concentration, but such concentration was found!

This stability result can be made more precise with a more exact error bound on the difference between the Poisson and the Binomial, and generalized to the product of Bernoulli distributions (and hence, with different p_i's). The way will be measuring this difference between two random variables is by a new process called *coupling*

Theorem 1.4.1: Approximation of Binomial using Poisson

Let $X_i \sim B(p_i)$ be n Bernoulli random variables on the same probability space with $p_i \in (0,1)$ for $1 \leq i \leq n$. Define

$$S_n = X_1 + X_2 + \cdots + X_n \qquad \lambda = p_1 + p_2 + \cdots + p_n$$

Then for any subset $A \subseteq \mathbb{Z}_{\geq 0}$

$$|\mathbb{P}(S_n \in A) - \text{Poiss}_\lambda(A)| \leq \sum_{i=1}^{n} p_i^2$$

If $p_1 = p_2 = \cdots = p_n$ and $S_n \sim B(n,p)$, then $\lambda = np$ and the aforementioned equation becomes

$$|\mathbb{P}(S_n \in A) - \text{Poiss}_\lambda(A)| \leq np^2 = \frac{\lambda^2}{n}$$

which if λ is fixed and we take $n \to \infty$, $B(n,p)$ is approximated by a Poisson.

***Proof* :**

We'll start off with a single Bernoulli variable $X_1 \sim B(p)$ and $Y_1 \sim \text{Poiss}(p)$ both defined on the same probability space. Note that the probability space in question does not matter as the random variables that take on these values are not related to the probability space. We will first construct a new random variable that will give us a sense of how X_1 and Y_1 will be "close" when p is small. First, consider the usual sample space for Y_1:

$$\Omega = \{0, 1, 2, ...\} \quad \mathbb{P}'(k) = \frac{p^k}{k!}e^{-p}$$

Notice that $\mathbb{P}(0) = e^{-p} \geq 1 - p$ and is only equal if p were 0 (something we're not allowing). What we'll do is split e^{-p} into two numbers so that one of them can become the new probability of $\mathbb{P}(0)$ which will be the probability for both X_1 and Y_1, and the other will be the probability for 1 X_1, with the second number being the probability of $\mathbb{P}(-1)$ for a new -1 we're adding to Ω to take it into account (say Ω_+). To split e^{-p}, notice that the function $1 - x$ is tangent to e^{-x} at zero, and so we can slit e^{-p} into two numbers:

$$e^{-p} = (1 - p) + (e^{-p} - 1 + p)$$

Thus, let $\mathbb{P}_p(-1) = 1 - p$, $\mathbb{P}_p(0) = e^{-p} - 1 + p$. All other values of $\mathbb{P}_p$ are defined as for the distribution of Y:

$$\mathbb{P}_p(k) = \frac{p^k}{k!}e^{-p},\ k \geq 1$$

Notice now that on our new probability space, we can still define the Bernoulli distribution X and Poisson distribution Y:

$$X(\omega) = \begin{cases} 0 & \omega = -1 \\ 1 & \text{otherwise} \end{cases} \qquad Y(\omega) = \begin{cases} 0 & \omega \in \{-1, 0\} \\ \omega & \text{otherwise} \end{cases} \tag{1.1}$$

This construction now let's us compare these two random variables! In particular, we can ask what is $\mathbb{P}_p(X = Y) = \mathbb{P}_p(\{\omega \in \Omega \mid X(\omega) = Y(\omega)\})$ and $\mathbb{P}_p(X \neq Y)$. For the first case, notice the condition is satisfied if and only if $\omega \in \{-1, 1\}$ and so

$$\mathbb{P}(X = Y) = 1 - p + pe^{-p} \geq 1 - p + p(1-p) = 1 - p^2$$

using the fact that $e^{-p} \geq 1 - p$. Thus, by the properties of a probability measure, the compliment of the set will have value:

$$\mathbb{P}(X \neq Y) \leq p^2 \tag{1.2}$$

Notice that when p is small, then the value is very small, showing that the two probabilities are very similar for small p's.

In the more general case, Let $X_i \sim B(p_i)$ for $1 \leq i \leq n$ and $Y_i \sim \text{Poiss}(p_i)$. We will couple X_i to Y_i on the same distribution in the same way. To achieve this, we will use the Product Bernoulli distribution for the X_i's and the stability property of the Poisson distribution. For the new sample space, take the same Ω_+ as before, and define

$$\Omega = (\Omega_+)^n$$

If $\omega = (\omega_1, \omega_2, ..., \omega_n)$, define $\mathbb{P}_+$ on Ω to be

$$\mathbb{P}_+ = \prod_{i=1}^{n} \mathbb{P}_{p_i}(\omega_i)$$

On this, notice that we can define the X_i in the following way:

$$X_i(\omega) = X(\omega_i) \qquad Y_i(\omega) = Y(\omega_i)$$

where X, Y where defined in equation(1.1). Clearly

- $X_i \sim B(p_i)$
- These random variables are independent (For Bernoulli, use the same trick as for the product of Bernoulli's shown earlier. Similar trick for the Poisson random variables)

Let S_n be the sum of Bernoulli's and S_n' be the sum of Poisson's. If $S_n \neq S_n'$, then they are not equal on at least on pair X_i, Y_i, which means that

$$\{S_n \neq S_n'\} \subseteq \bigcup_{i=1}^{n} \{X_i \neq Y_i\}$$

Thus, by the properties of measures:

$$\mathbb{P}_+(S_n \neq S_n') \leq \sum_{i=1}^{n} \mathbb{P}_+(X_i \neq Y_i)$$

By equation (1.2), and independence of the variables, we have that:

$$\mathbb{P}_+(S_n \neq S_n') \leq \sum_{i=1}^{n} (p_i)^2$$

To get our desired equality from the definition, notice that $\mathbb{P}(S_n \in A) = \mathbb{P}(S_n \in A,\ S'_n \in A) + \mathbb{P}(S_n \in A,\ S'_n \notin A)$, the two sets being disjoint. Then:

$$\begin{aligned}\mathbb{P}_+(S_n \in A) - \mathbb{P}_+(S'_n \in A) &= \mathbb{P}(S_n \in A,\ S'_n \in A) + \mathbb{P}(S_n \in A,\ S'_n \notin A)\\ &\quad - \mathbb{P}(S'_n \in A,\ S_n \in A) - \mathbb{P}(S'_n \in A,\ S_n \notin A)\\ &= \mathbb{P}(S_n \in A,\ S'_n \notin A) - \mathbb{P}(S'_n \in A,\ S_n \notin A))\\ &\leq \mathbb{P}(S_n \in A,\ S'_n \notin A) + \mathbb{P}(S'_n \in A,\ S_n \notin A))\\ &= \mathbb{P}(S_n \neq S'_n)\end{aligned}$$

with the same proof working if we take $\mathbb{P}_+(S'_n \in A) - \mathbb{P}_+(S_n \in A)$, hence justifying the absolute values. Combining with our earlier result, this gives us:

$$|\mathbb{P}_+(S_n \in A) - \mathbb{P}_+(S'_n \in A)| \leq \mathbb{P}_+(S_n \neq S'_n) \leq \sum_{i=1}^{n}(p_i)^2$$

completing the proof

1.5 $\mathbb{E}f(X_1, .., X_n)$ and Independence

So far, we showed that the sum of i.i.d. Bernoulli random variables forms a binomial random variable, and the limit of a particular type of Binomial (i.e. $B\left(n, \frac{\lambda}{n}\right)$) is the Poisson random variable, and that summing two independent Poisson is also a Poisson.

A goal that will pop up once in a while in this book is to understand what happens when we combine random variables in certain ways. In future sections of this book, we will be more general with what functions we are taking between random variables (essentially treating random variables as elements in a function space, and hence the term “variable”). So we will see notation like $f(X_k)$ or $f((X_k)_{k\in I})$ to represent a function on a random variable or variables. I like to think of this as finding relations between statistical questions. This idea will be explored in more depth in the next section when we will try to find the expected value of the not just a random variable $\mathbb{E}X$, but of $\mathbb{E}f(X_1, X_2, ..., X_n)$. There is a very neat result that will allow us to compute this value easily, and if the random variables are all independent, then this becomes even easier!

1.5.1 Independence Revisited

We now take more time to discuss independence. Recall that if we flip a twice, and assume that there is no relation between the chance of the first and second coin, then the probabilities of the sequence $(1, 1)$ (i.e. heads, heads) is equal to the probability of 1 times the probability of 1. This motivates the following definition:

Definition 1.5.1: Independence of Events

Let $(\Omega, \mathbb{P})$ be a probability space, and let $A, B \subseteq \Omega$ be events. Then the events A and B are said to be *independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$$

that, the probability of event A and B happening is equal to the probability of A times the probability of B.

Another way to think of this in terms of conditional

Definition 1.5.2: Conditional Probability

Let $(\Omega, \mathbb{P})$ be a [discrete] probability space, and let $A, B \subseteq \Omega$ be events with $\mathbb{P}(B) > 0$. Then

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

My intuition behind this is that we are changing the probability space to the smaller one. To see this, consider:

$$\mathbb{P}(A|\Omega) = \frac{\mathbb{P}(A \cap \Omega)}{\mathbb{P}(\Omega)} = \frac{\mathbb{P}(A)}{1} = \mathbb{P}(A)$$

This relates to independence by noticing that A and B are independent if $\mathbb{P}(A|B) = \mathbb{P}(A)$, so changing the probability space to looking at B does not change the value of A (i.e. A is not affected).

Independence can be generalized to multiple variables. It is important to remember that to be independent, non of the events must effect any other event. That means we actually need to be stronger than pairwise independent.

Example 1.15: Independence of multiple events

Take a tetraherdron coloured red, green, and blue on 3 of the sides, and all 3 colours on the 4th. Let the chances of any side be equal. Calculate the chance of landing on one of the colours. This is pair-wise independent, but not independent.

Thus, we generalize independence in the following way:

Definition 1.5.3: Independence over Multiple Events

Let $(\Omega, \mathbb{P})$ be a probability space and $A_i \subseteq \Omega$ be an events with indexes between $1 \leq i \neq n$. Then the events are called independent if and only if for any finite subset of indices $I \subseteq \{1, 2, ..., n\}$

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$$

Just like with the case with two events, (p.39 with conditional event similarity)

Example 1.16: exercise

1. with compliments of sets
2. example on p.39 with pair-wise independence not implying independence (tetrahedron)

As usual, if we can define something for a probability measure, we can do the same for random variables:

Definition 1.5.4: Independent Random Variables

Let $(\Omega, \mathbb{P})$ be a probability space, and let $X_1, ..., X_n$ be random variables on Ω. Then these random variables are said to be independnet if

$$\mathbb{P}(X_1 = a_1,\ X_2 = a_2,\ \cdots X_n = a_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = a_i)$$

for all possible $a_i \in \Omega$.

Notice that instead of saying it must be true for all posible $a_i \in \Omega$, we can replace $X_1 = a_1, ...X_n = a_n$ with $X_1 \in A_1, ..., X_n \in A_n$ for arbitrary subsets A_i. This also means that $A_i = \mathbb{R}$, and so the value can essentially be "eliminated" from our calculations (since $\mathbb{P}(X \in \mathbb{R}) = 1$), meaning if we have a family of independent random variables, any subfamily is also independent.

Similarly to when we conditional probability, we can define conditional random variables:

Definition 1.5.5: Conditional Random Variables

et $(\Omega, \mathbb{P})$ be a probability space and let X, Y be two random variables on Ω. THen

$$\mathbb{P}(X = a|Y = b) = \frac{\mathbb{P}(X = a, Y = b)}{\mathbb{P}(Y = b)}$$

This also implies that we have another way of writing $\mathbb{P}(X = a, Y = b) = \mathbb{P}(X = a|Y = b)\mathbb{P}(Y = b)$.

To get an intuition independent random variables, let's explore a generalization of how we defined the Binomial distribution to show how to internalize the idea. We can generalize our previous notion of product of Bernoulli random variables to the following:

Definition 1.5.6: Product Space

Let $(\Omega_i, \mathbb{P}_i)$ be probability spaces for $1 \leq i \leq n$. Define

$$\Omega = \Omega_1 \times \Omega_2 \times \cdots \times \Omega_n$$

and

$$\mathbb{P}(\omega) = \mathbb{P}(\omega_1, \omega_2, ..., \omega_n) = \prod_{i=1}^{n} \mathbb{P}_i(\omega_i)$$

Then $(\Omega, \mathbb{P})$ is called the *product measure* of the $(\Omega_i, \mathbb{P}_i)$ spaces.

The fact that the product space is defined so that every component (or "coordinate") is independent has some nice results. For example, if we take some hyper-rectangle

$$A = A_1 \times A_2 \times \cdots \times A_n$$

Then

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\omega) = \sum_{\omega \in A} \prod_{i=1}^{n} \mathbb{P}_i(\omega_i) \overset{!}{=} \prod_{i=1}^{n} \sum_{\omega_i \in A_i} \mathbb{P}_i(\omega_i) = \prod_{i=1}^{n} \mathbb{P}_i(A_i)$$

showing that product and probability spaces "commute" when the random variables are independent. Due to this, we will define a *random vector* $(X_1, X_2, ..., X_n)$ to be a n-tuple of random variables, each acting on their particular coordinates.

An important reason to care more about the independence of random variables instead of events is because they behave much more nicely. The following theorem is an excellent illustration of this: If $X_1, X_2, ..., X_n$ are random variables, and we partition them into different groups forming new random variables $Y_1, ..., Y_m$, then these random variables are also independent!

In fact, we can be even stronger. If I_k represents the partition of our n random variables for $1 \leq k \leq m$, then define the new random variables Y_k to be

$$Y_k = f_k((X_i)_{i \in I_k})$$

for some arbitrary functions f_k. Then these Y_k's are still independent! In other words, if we have random variables, and we partition them and define new functions based off of random variable, these are still independent. A simple example would be each $I_k = \{X_k\}$ and $Y_k = X_k^2$. Another would be $I_1 = \{X_1, ..., X_{n-1}\}$ and $I_2 = \{X_n\}$, and $Y_1 = X_1(X_2)^2(X_3)^3 \cdots (X_{n-1})^{n-1}\}$, and $Y_2 = \log(X_k)$. Then Y_1 and Y_2 are still independent!

Lemma 1.5.1: Grouping Lemma

Let $X_1, X_2, ..., X_n$ be arbitrary random variables and let $Y_1, Y_2, ..., Y_m$ be defined as

$$Y_k = f_k((X_i)_{i \in I_k})$$

for appropriate I_k that partition the X_i random variables. Then the $Y_1, Y_2, ..., Y_m$ are independent

***Proof* :**
We need to show that

$$\mathbb{P}(Y_1 = a_1,\ Y_2 = a_2, ..., Y_m = a_m) = \prod_{i=1}^{m} \mathbb{P}(Y_i = a_i)$$

for all a_i, or more concisely

$$\mathbb{P}(Y_1 \in A_1,\ Y_2 \in A_2, ..., Y_m \in A_m) = \prod_{i=1}^{m} \mathbb{P}(Y_i \in A_i)$$

for all $A_i \subseteq \mathbb{R}$. To that end, take $\{Y_1 \in A_1, Y_2 \in A_2, ..., Y_m \in A_m\}$ where $A_i \subseteq \mathbb{R}$. Next, we want to find what a set $\{Y_i \in A_i\}$ contains. Consider

$$B_k = f^{-1}(A_k) = \{(x_i)_{i \in I_k} \mid f_k((x_i)_{i \in I_k}) \in A_k\}$$

that is, B_k is the set of *vectors* that map into A_k. Then if we let $x^k = (x_i)_{i\in I_k}$ and $X^k = (X_i)_{i\in I_k}$ represent an element of X^k and X^k is the product random variables with indices in I_k, then

$$\{Y_k \in A_k\} = \{f_k(X^k) \in A_k\} = \{X^k \in B_k\}$$

From here, we have now established enough of our notation to proceed with the proof:

$$\mathbb{P}(Y_1 \in A_1,\ Y_2 \in A_2, ..., Y_m \in A_m) = \sum_{x\in B} \mathbb{P}(X^1 = x^1,\ X^2 = x^2, ..., X^m = x^m) \tag{1.3}$$

Essentially, we reduced the problem the sum of product spaces, which we know to be independent:

$$\mathbb{P}(X^1 = x^1, ..., X^m = x^m) = \mathbb{P}(X^1 = x^1)\cdots\mathbb{P}(X^m = x^m)$$

thus, we continue with our calculations with this new result:

$$\begin{aligned}
&= \sum_{x\in B} \mathbb{P}(X^1 = x^1)\cdots\mathbb{P}(X^m = x^m)\\
&= \sum_{x^1\in B_1} \mathbb{P}(X^1 = x^1)\cdots \sum_{x^m\in B_m} \mathbb{P}(X^m = x^m)\\
&= \mathbb{P}(X^1 \in B_1)\cdots\mathbb{P}(X^m \in B_m)\\
&= \mathbb{P}(Y_1 \in A_1)\cdots\mathbb{P}(Y_m \in B_m)
\end{aligned}$$

Combining this with equation (1.3), we get:

$$\mathbb{P}(Y_1 \in A_1,\ Y_2 \in A_2, ..., Y_m \in A_m) = \mathbb{P}(Y_1 \in A_1)\cdots\mathbb{P}(Y_m \in B_m)$$

completing the proof.

Example 1.17: Poisson Revisited

Recall that if X_1 and X_2 are independent Poisson Random distributions with λ_1 and λ_2, then $(X_1 + X_2) \sim \text{Poiss}(\lambda_1 + \lambda_2)$. We also saw in section 1.4.1 that $\lim_{n\to\infty} B(n, n/\lambda) = \text{Poiss}(\lambda)$. We will show another way of seeing that the sum of two Poisson distributions ought to be Poisson.

Let' say that λ_1 and λ_2 had a rational ration:

$$\frac{\lambda_1}{\lambda_2} = \frac{r_1}{r_2} \qquad r_1, r_2 \in \mathbb{Z}$$

If not, then any rational number is arbitrarily close to λ_1/λ_2, and so modifying one of the lambda's should be no problem. Then re-arranging so the denominator are rational we get:

$$\frac{\lambda_1}{r_1} = \frac{\lambda_2}{r_2}$$

Now, divide the denominator by an n large enough so that

$$\frac{\lambda_1}{nr_1} = \frac{\lambda_2}{nr_2} = p \in (0, 1)$$

Now set $m_1 = nr_1$, $m_2 = nr_2$, and $m = m_1 + m_2$. Since a binomial $B(m, p)$ is the sum of p i.i.d.

Bernoulli with probability p, let X_i $B(p)$ and

$$Y_1 = X_1 + X_2 + \cdots + X_{m_1} \sim B(m_1, p) \approx \text{Poiss}(m_1 p) = \text{Poiss}\left(m_1 \frac{\lambda_1}{m_1}\right) = \text{Poiss}(\lambda_1)$$

$$Y_2 = X_{m_1+1} + X_{m_1+2} + \cdots + X_{m_1+m_2} \sim B(m_2, p) \approx \text{Poiss}(m_2 p) = \text{Poiss}\left(m_2 \frac{\lambda_2}{m_2}\right) = \text{Poiss}(\lambda_2)$$

And since Y_1 and Y_2 are sums of independents random variables,

$$Y_1 + Y_2 \sim B(m, p) \approx \text{Poiss}(mp) = \text{Poiss}(\lambda_1 + \lambda_2)$$

And so we see how connecting the Poisson distribution to The Bernoulli distribution through many many flips of a coin.

1.5.2 Expected Value of Functions of RV's

We now return to the question of finding $\mathbb{E}f(X_1, X_2, ..., X_n)$ for random variables X_i. We will show that this is simple to compute if the random variables are independent, and the formula for the dependent case is not much more complex! We start off with the case of $f(X, Y) = XY$:

Proposition 1.5.1: Product of Independent Expected Values

Let $(\Omega, \mathbb{P})$ be a probability space, Let X, Y be indepenent random variables on Ω. If $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$, then $\mathbb{E}|XY| < \infty$ and

$$\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$$

***Proof* :**

This is an application of Change of Variables and of lemma 1.3.1 (or corrollary 1.3.1). First, By definition

$$\mathbb{E}XY = \sum_{\omega \in \Omega} X(\omega)Y(\omega)\mathbb{P}(\omega)$$

our goal is

$$\mathbb{E}X\mathbb{E}Y = \left(\sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega)\right)\left(\sum_{\omega \in \Omega} Y(\omega)\mathbb{P}(\omega)\right)$$

to accomplish this, we will use change of variables. Since we're working in the countable case, give some order to the image of X and Y: $a_1, a_2, ..., a_n, ...$ and $b_1, b_2, ..., b_n, ...$ respectively. Define the events

$$\Omega_{nm} = \{\omega \in \Omega \mid X(\omega) = a_n,\ Y(\omega) = b_n\}$$

Since $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$, by change of variables and lemma 1.3.1

$$\begin{aligned}
\mathbb{E}XY &= \sum_{\omega\in\Omega} X(\omega)Y(\omega)\mathbb{P}(\omega) \\
&= \sum_{n,m} a_n b_m \mathbb{P}(X = a_n,\ Y = b_m) && \text{Change of Variables} \\
&= \sum_{n,m} a_n b_m \mathbb{P}(X = a_n)\mathbb{P}(Y = b_m) && \text{independence} \\
&= \left(\sum_{i=1}^{\infty} a_i \mathbb{P}(X = a_i)\right)\left(\sum_{i=1}^{\infty} b_i \mathbb{P}(Y = b_i)\right) && \text{lemma 1.3.1} \\
&= \mathbb{E}X\mathbb{E}Y
\end{aligned}$$

Notice we can apply lemma 1.3.1 since $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$.

Finally, the fact that $\mathbb{E}|XY|$ comes essentially immediately. Notice the proof is identical for $\mathbb{E}|X||Y|$, so:

$$\mathbb{E}|XY| = \mathbb{E}|X||Y| = \mathbb{E}|X|\mathbb{E}|Y| < \infty$$

Completing the proof.

(I don't get this comment) In our calculations, when we did $\sum_{\omega\in\Omega} X(\omega)Y(\omega)\mathbb{P}(\omega) = \sum_{\omega\in\Omega_{nm}} a_n b_m \mathbb{P}(X = a_n,\ Y = b_m)$, we could just as well have used random vectors. We will use this fact later.

Example 1.18: Properties

1. Use the Grouping lemma and induction to show that

$$\mathbb{E}\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} \mathbb{E}(X_i)$$

2. Let $X_1, X_2, ..., X_n$ be i.i.d. Bernoulli Random variables. Computer

$$\mathbb{E}(X_1 + X_2 + \cdots + X_n)^2$$

Notice that in the proof, the fact that we used a "particular" function XY on the random variables X and Y is something that can be generalized. Given $f(X, Y)$ for some function f and we have reached this step in the computation of $\mathbb{E}f(X, Y)$:

$$\sum_{\omega\in\Omega_{nm}} a_n b_m \mathbb{P}(X = a_n)\mathbb{P}(Y = b_m) = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty} a_i b_j \mathbb{P}(X = a_i)\mathbb{P}(Y = b_j)$$

where we replace ∞ with the cardinality of the domain of X and Y if it is finite, then by since $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$, by lemma 1.3.1

$$\sum_{i=1}^{\infty}\left(\sum_{j=1}^{\infty} a_i b_j \mathbb{P}(X = a_i)\right)\mathbb{P}(Y = b_j)$$

This might actually look quite familiar from a result in calculus:

Theorem 1.5.1: Fubini's Theorem

Let $(\Omega, \mathbb{P})$ be a probability space, and let X, Y be independent random variables with $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$. Then if $f(X,Y)$ is some function of X and Y such that $\mathbb{E}|f(X,Y)| < \infty$, then

$$\mathbb{E}f(X,Y) = \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} f(a_i, b_j)\mathbb{P}(X = a_i) \right) \mathbb{P}(Y = b_j)$$

Essentially, we can take the average over all fixed b_j, and then average over all b_j's! This is Analogous to the famous Calculus Theorem:

$$\iint\limits_{[0,1]^2} f(x,y)dxdy = \int_0^1 \left(\int_0^1 f(x,y)dy \right) dx$$

Example 1.19: Different Perspective

Fubini's Theorem can be "brought back down" to the probability space to get a second perspective. In particular, if X and Y are independent and $\mathbb{E}|f(X,Y)| < \infty$ then:

$$\sum_{\omega \in \Omega} \left[\sum_{\omega' \in \Omega} f(X(\omega), Y(\omega'))\mathbb{P}(\omega') \right] \mathbb{P}(\omega)$$

where we sum over Ω twice.

What if X and Y are dependent? Then our calculations will break down at the following step:

$$\begin{aligned} \mathbb{E}f(X,Y) &= \sum_{\omega \in \Omega} f(X(\omega), Y(\omega))\mathbb{P}(\omega) \\ &= \sum_{\omega \in \Omega_{nm}} f(a_n, b_m)\mathbb{P}(X = a_n,\ Y = b_m) && \text{Change of Variables} \end{aligned}$$

At this point, we were able to use independence to split the result. In the dependent case, we can introduce a sneaky 1 in the form of $\frac{\mathbb{P}(X=a_n)}{\mathbb{P}(X=a_n)}$ and get:

$$\begin{aligned} \sum_{n,m} f(a_n, b_m)\mathbb{P}(X = a_n,\ Y = b_m) &= \sum_{n,m} f(a_n, b_m)\frac{\mathbb{P}(X = a_n,\ Y = b_m)}{\mathbb{P}(X = a_n)}\mathbb{P}(X = a_n) \\ &= \sum_{m} \left[\sum_{n} f(a_n, b_m)\frac{\mathbb{P}(X = a_n,\ Y = b_m)}{\mathbb{P}(X = a_n)} \right] \mathbb{P}(X = a_n) \end{aligned}$$

where the last line is by lemma 1.3.1. Notice that the inside term is the conditional distribution! This result is important enough to get called out as a theorem:

Theorem 1.5.2: Fubini's Theorem II

Let $(\Omega, \mathbb{P})$ be a probability space, and let X, Y be random variables (not necessarily independent) with $\mathbb{E}|X| < \infty$ and $\mathbb{E}|Y| < \infty$. Then if $f(X, Y)$ is some function of X and Y such that $\mathbb{E}|f(X, Y)| < \infty$, then

$$\sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} f(a_i, b_j) \mathbb{P}(Y = b_i \mid X = a_i) \right) \mathbb{P}(X = a_i)$$

The $\mathbb{P}(X = a|Y = b)$ term is similar to conditional probability, but given random variables. It is called the *conditional distribution* given fixed $Y = b$. It an easily be verified that this in fact a probability measure (for a fixed $Y = b$ since it's probabilities sum to 1. A special notation is used for the expected value of this

Definition 1.5.7: Conditional Distribution

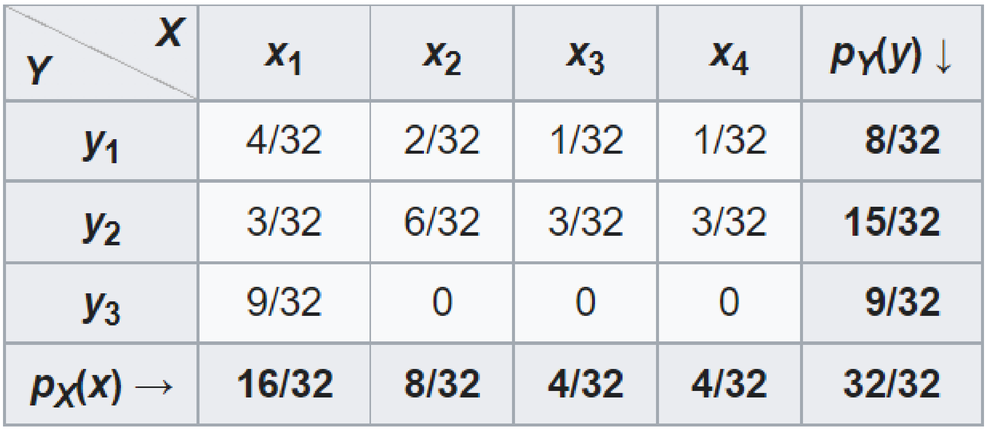
Let X, Y be random variables. Then $\mathbb{P}(Y = b_m|X = a_n)$ is called the *conditional distribution with respect to* a_n.

The expected value of this distribution is:

$$\mathbb{E}(f(a_n, Y)|X = a) := \sum_m f(a_n, b_m)\mathbb{P}(Y = b_m|X = a)$$

Example 1.20: Conditional Distribution

Let's say we had random variable X and Y on some space Ω for which the following table outlines the probability:



X Y	x1	x2	x3	x4	pY(y) ↓
y1	4/32	2/32	1/32	1/32	8/32
y2	3/32	6/32	3/32	3/32	15/32
y3	9/32	0	0	0	9/32
pX(x) →	16/32	8/32	4/32	4/32	32/32

Then any row or column in this table is a conditional distribution if we switch the denominator to the sum of the numerators of the row's or columns (given what we are taking it "with respect to").

Notice that in the image, the table actually represents the probability of $\mathbb{P}((X, Y) = (x_i, y_j))$. and there is a bottom row and right column that holds the values of what is summed of the rows and columns: this is given a name:

Definition 1.5.8: Marginal Distribution

Let X and Y be random variables on Ω. We an define $\mathbb{P}_Y(X = x) = \sum_{y_i} \mathbb{P}((X, y_i) = (x, y_i))$ to be the *marginal distribution with respect to* Y. Similarly for $\mathbb{P}_X$.
More generally, if $\{X_i\}_I$ is a collection of random variables, then (can we define this more generally?)

Note that the expected value of the marginal distribution is calculated precisely in the way Fubini's Theorem is calculated.

Definition 1.5.9: Joint Distribution

Let $(\Omega, \mathbb{P})$ be a probability space and let X, Y be two random variables on Ω. Define the *joint distribution* (X, Y) to be

$$\mathbb{P}((X,Y) = (a,b)) := \mathbb{P}(X = a,\ Y = b) = \mathbb{P}(Y = b|X = a)\mathbb{P}(X = a)$$

(More on interpreting Fubini II as joint probability, then do shark attack and Colouring Poisson example to illustrate how to work with Fubinni II, mention how the probability space is not really mentioned, and will be getting mentioned less often in future theorems. Then elaborate how we can extend it to not just two events, but multiple dependent events, which leads to Markov Chains, which will be studied in chapter ref:HERE)

Lemma 1.5.2: Poisson Colouring

Let $N \sim \text{Poiss}(\lambda)$ and:

$$\mathbb{P}(N_1 = k|N = n) = \binom{n}{k} p^k (1-p)^{n-k}$$

that is, for every $N = n$, get a binomial distribution $B(n, p)$. Let $N_2 = N - N_1$. Then N_1 and N_2 are independent and have distribution's $\text{Poiss}(p\lambda)$ and $\text{Poiss}((1-p)\lambda)$ respectively

Here is a way to make this theorem intuitive: If a Chicken lays a Poisson number of eggs, and then we colour them red or blue depending on a coin toss, then knowing the number of red eggs gives us nothing about the number of blue eggs, and the number of red and blue eggs are also Poisson

> ***Proof*** **:**
> as an exercise for midterm

1.6 Inequality With Expected Value

In this section, we start exploring one of the basic types of result we try to find in Probability: *approximation.* Approximating is probability the most important statistical concept as it lets you bound difficult problems with simpler variables, and gives important heuristics in solving more complicated problems. Usually, this means we bound using inequalities the values of a problem from above and/or bellow. In this section, we introduce two famous inequalities that let's us do exactly that. In particular, we will be focusing on bounding from above.

Theorem 1.6.1: Chebyshev's Inequality

Let X be a random variable where $X \geq 0$.

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}X}{x}$$

and if X is any random variable, then for $x > 0$

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(XI(|X| \geq x))}{x}$$

***Proof* :**
Try deriving it for yourself! For the $X \geq 0$ case,

$$\mathbb{E}X = \sum_{x=1}^{\infty} \mathbb{P}(X \geq x)$$

$$\mathbb{E}X = \sum_{k=1}^{\infty} k\mathbb{P}(X = k) \geq \sum_{k=x}^{\infty} k\mathbb{P}(X = k) \geq \sum_{k=x}^{\infty} x\mathbb{P}(X = k) = x\sum_{k=x}^{\infty} \mathbb{P}(X = k) = x\mathbb{P}(X \geq x)$$

so

$$\mathbb{P}(X \geq x) \leq \frac{\mathbb{E}X}{x}$$

If X is not restricted to a codomain of $\mathbb{N}$, then we "limit" it by taking an appropriate indicator function. First, notice that:

$$X \cdot I(X \geq x) \geq x \cdot I(X \geq x)$$

since if $X < x$, the indicator makes both values 0. Taking the expected value on both sides, and recalling that $\mathbb{E}(I(X \geq x)) = \mathbb{P}(X \geq x)$ and the expected value is linear, we get:

$$x\mathbb{P}(X \geq x) \leq \mathbb{E}(X \cdot I(X \geq x)) \qquad \Leftrightarrow \qquad \mathbb{P}(X \geq x) \leq \frac{\mathbb{E}(X \cdot I(X \geq x))}{x}$$

completing the proof.

Though simple, this inequality gives us one of our first real insights on approximating problems, in particular that the "tail" of a probability given a random variable will never really exceed the expected value! For example, let's say X random variable such that $\{X > 0\} = \{X \geq 1\}$[3]. Then using Chebyshev's inequality at $x = 1$ give us

$$\mathbb{P}(X > 0) = \mathbb{P}(X \geq 1) \leq \mathbb{E}X$$

If $\mathbb{E}X \ll 1$, then that means that $\mathbb{P}(X = 0) = 1 - \mathbb{P}(X \geq 1)$ is very large, meaning most of our values

[3] Don't forget this is a shorthand for all values such that $X > 0$ and $X \geq 1$, see example 1.5

will be 0!

Example 1.21: Chebyshev's Inequality
here

The next theorem let's get bounds when applying differentiable functions to $\mathbb{E}X$. Recall a function is convex if a line segment between any two points in the graph does not lie bellow the graph. The function is concave otherwise:

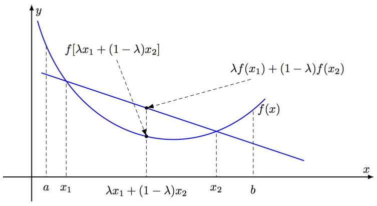

y
$f[\lambda x_1 + (1-\lambda)x_2]$
$\lambda f(x_1) + (1-\lambda)f(x_2)$
$f(x)$
x
a
x_1
$\lambda x_1 + (1-\lambda)x_2$
x_2
b

Figure 1.1: Convex Function

Lemma 1.6.1: Jensen's Inequality

Let $f : \mathbb{R} \to \mathbb{R}$ be a convex function (assumed to be differentiable) with respect to the x-axis, and $\mathbb{E}X$ is well-defined. Then:

$$f(\mathbb{E}X) \leq \mathbb{E}f(X)$$

If f is concave, the inequality is reversed:

$$f(\mathbb{E}X) \geq \mathbb{E}f(X)$$

***Proof* :**
Let $\mu = \mathbb{E}X$. Since f is convex, the tangent line of $f(\mu)$ is bellow the graph, so

$$f(x) \geq f(\mu) + f'(\mu)(x - \mu)$$

Thus, for all $x \in X$, we have:

$$f(X) - f(\mu) - f'(\mu)(x - \mu) \geq 0$$

Thus, taking the expected value, we get:

$$\mathbb{E}f(x) - f(\mu) - f'(\mu)(\mathbb{E}x - \mu) = \mathbb{E}f(X) - f(\mu) \geq 0$$

As we sought to show. The proof in the concave case is very similar.

Note that this theorem works on any interval as long as the random variable X takes on those values

Example 1.22: Jensens Inequality

Suppose $\mathbb{E}X^2 < \infty$. Notice that $f(x) = x^2$ is a convex function, and so

$$(\mathbb{E}X)^2 \leq \mathbb{E}X^2$$

On The other hand, $f(x) = \sqrt{x}$ is concave, and so

$$\sqrt{\mathbb{E}X^2} \geq \mathbb{E}\sqrt{X^2} = \mathbb{E}|X|$$

Furthermore, $X \leq |X|$, and so by monotonicity of the expected value (proposition 1.2.1) $\mathbb{E}|X| \geq \mathbb{E}X$. Notice that since $\mathbb{E}X^2 < \infty$, then $\sqrt{\mathbb{E}X^2} < \infty$ and so this also shows that if $\mathbb{E}X^2$ is defined, so is $\mathbb{E}X$!

1.7 Computation trick: Linearity of Expectation

In the preceding sections, the importance of the expected value of a random variable should now be evident in how often it is a value we want to know. In this section, we show how it is sometimes very easy to find the expected value of a random variable if the random variable is the sum of indicators:

$$N = \sum_{i=1}^{n} I_{A_i} \qquad A_i \subseteq \Omega$$

The key idea is that this will circumnavigate the change of variable trick we've been doing so far, which can be sometime difficult to come up with. As an example of the power of random variables which are sums of indicators, recall that when we computed the expected value of the Binomial distribution, we took advantage of the fact that we can represent it as the sum of i.i.d. Bernoulli variables. This trick allowed us to incredibly quickly calculate the expected value. Similarly, we can take the expected value of N by taking advantage of the linearity of the expected value and the known value of the expected value of an indictaor function (example 1.4):

$$\mathbb{E}N = \mathbb{E}\left(\sum_{i=1}^{n} I_{A_i}\right) = \sum_{i=1}^{n} \mathbb{E}(I_{A_i}) = \sum_{i=1}^{n} \mathbb{P}(A_i)$$

the key insight here is that we avoided the possible ugliness of taking $\mathbb{E}X = \sum_{k=1}^{\infty} k\mathbb{P}(X = k)$ by "evaluating" the expected value only once we reached the indicator functions, which have a very simple expected value!

The purpose of this is that it is often easier to take advantage of trick to to compute $\mathbb{P}(A_i)$ instead of finding $\mathbb{P}(N = n)$ and then sum over all n for $\mathbb{E}N = \sum_{i=1}^{n} \mathbb{P}(N = n)$. We will show how seomtimes $\mathbb{P}(A_i)$ is easier to compute.

We will apply this in the study of finite random variables (which will be interesting since we use such functions all the time) and cliques in the Erdos-Reyni graph

1.7.1 Functions on Finite sets

Let $X = \{1, 2, ..., n\}$ and Ω be the set of all function from X to X. We can define a probability function representing the change of choosing one of these functions by

$$\mathbb{P}(\omega) = \frac{1}{n^n}$$

Let Given any $\omega \in \Omega$, let $N = \text{card}(\omega(X))$ be the cardinality of the range:

$$N = |\{(\omega(1), \omega(2), ..., \omega(n)\}|$$

Then N is a random variable, and since $|\Omega| < \infty$, $\mathbb{E}N$ is clearly well-defined. If we try to compute

$$\mathbb{E}N = \sum_{k=1}^{n} k\mathbb{P}(N = k)$$

this would be a greatly difficult task (as we will show soon). A much easier way to compute this is to notice that we can represent

$$N(\omega) = \sum_{i=1}^{n} I(i \in \omega(X))$$

that is, it is equivalent to think of N as the sum of indicator's that check whether i is in the range of ω. Since we are working with the uniform probability, we have that $\mathbb{P}(I(i \in \omega(X))) = \mathbb{P}(I(1 \in \omega(X)))$ and so

$$\mathbb{E}N = \mathbb{E}\sum_{i=1}^{n} I(i \in \omega(X)) = \sum_{i=1}^{n} \mathbb{P}(I(i \in \omega(X))) = \sum_{i=1}^{n} \mathbb{P}(I(1 \in \omega(X))) = n\mathbb{P}(I(1 \in \omega(X)))$$

Thus, we simply calculate $\mathbb{P}(1 \in \omega(X))$. It in fact easier to calculate $\mathbb{P}(1 \notin \omega(X))$:

$$\mathbb{P}(1 \in \omega(X)) = 1 - \mathbb{P}(1 \notin \omega(X)) = 1 - \frac{(n-1)^n}{n^n}$$

since if $1 \notin \omega(X)$, we are essentially counting all functions which don't map to 1, which is all functions of the form $\{1, 2, ..., n\} \to \{2, 3, ..., n\}$, giving us the value at the end. Thus, multiplying by n, we get:

$$\mathbb{E}N = n\left(1 - \left(1 - \frac{1}{n}\right)^n\right) \sim n\left(1 - \frac{1}{e}\right)$$

where $\sim$ means the ratio between the two go to 1 as $n \to \infty$. This gives a very succinct view of the expected value!

If we were to do $\mathbb{E}N = \sum_{k=1}^{n} k\mathbb{P}(N = k)$, we require a lot more work!. We start by finding $\mathbb{P}(N = k)$. First, if $N(\omega) = k$, then there are $\binom{n}{k}$ ways of choosing k values from the range of n variables. With this, we can now count all function which map to $\{1, 2, ..., k\}$. If we start by denoting Ω_k to be the set of all functions from $X \to \{1, 2, ..., k\}$, we then must ask how many surjective functions are there in Ω_k.

To calculate this, we will partition Ω_k to be sets that miss one of the values in the codomain:

$$A_i = \{\omega \in \Omega_k \mid i \notin \omega(X)\}$$

Then the set $A_1 \cup A_2 \cup \cdots A_n$ is the collection of all functions that miss at least one element in their ranges. Defining $\mathbb{P}_k$ to be the uniform probability on Ω_k, $\mathbb{P}_k(A) = \frac{\text{card}(A)}{n^k}$, we are essentially finding $\mathbb{P}(\cup_i^n A_i)$. To find this value, we can use the principle of Inclusion-Exclusion

$$here$$

(Finish typing this up later, I don't feel rn this is the most important)

The conclusion is

$$\mathbb{P}(N = k) = \frac{1}{n^n}\binom{n}{k}\sum_{\ell=0}^{k}(-1)^\ell\binom{k}{\ell}(k-\ell)^n$$

And so:

$$\mathbb{E}N = \sum_{i=1}^{n} i\left(\frac{1}{n^n}\binom{n}{k}\sum_{\ell=0}^{k}(-1)^\ell\binom{k}{\ell}(k-\ell)^n\right)$$

which it is perhaps immediately evident to see that this is much harder to compute.

1.7.2 Longest Increasing Subsequence

Let's now restrict our attention to all bijective functions from X to X. This probability space already has the name of S_n from group theory, and is usually called the *symmetric group* (since it has a group structure). We will not need any reference to it as a group, but we will use the S_n notation for mathematical consistency. Notice that $|S_n| = n!$, and so we can define the uniform probability to be $\mathbb{P}(\sigma) = \frac{1}{n!}$.

What we will focus on is the random variable L which gives the length longest increasing subsequence. In particular, if $i_1 < i_2 < \cdots < i_k$ and

$$\sigma(i_1) < \sigma(i_2) < \cdots < \sigma(i_k)$$

Then $L(\sigma)$ would be the maximum length of all subsequences satisfying this property.

Example 1.23: chain examples

If $\sigma = (1\ 5\ 2\ 4\ 3)$, then $L(\sigma) = 3$. If $\tau = (1\ 2\ 4)$ or $(1\ 2\ 3$, then $L(\tau) = 3$. If $\rho = (5\ 4\ 3\ 2\ 1)$ then $L(\rho) = 0$

To find the actual distribution of L is in fact notoriously difficult, and it has been an open problem (till recently) to show that

$$\lim_{n\to\infty}\frac{\mathbb{E}L}{\sqrt{n}} = 2$$

We will not show how to get perfect bound of 2, but we will show that we can upper bound it by 4

(finish this later)

If N_k is the number of chains of length k, then:

$$\mathbb{E}N_n = \binom{n}{k}\frac{1}{k!}$$

1.7.3 Cliques in Erdös-Rényi graph

Let $G = G(n,p)$ be a random graph. We will show that the expected number of cliques is typically (i.e. close to 1) around $\mathcal{O}(\log(n))$. Let N_k be the number of cliques of size k. We want to computer $\mathbb{E}N_k$. There are $\binom{n}{k}$ possible subsets of vertices of size k, and we need each possible combination to have had edge added to it. If W is our clique of size k, then:

$$\mathbb{P}(W \text{ is a clique}) = p^{\binom{k}{2}}$$

Notice that we can represent N_k as the sum of indicator functions indicating whether a subset W is a clique or not, giving us:

$$\mathbb{E}N_k = \binom{n}{k} p^{\binom{k}{2}}$$

If N_k is the number of cliques of size k, then:

$$\mathbb{E}N_k \binom{n}{k} p^{\binom{k}{2}}$$

We can use this formula to some more information. First, recall the Stirling's formula:

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

Now, let $f(k) = \mathbb{E}N_k$. Notice that $f(1) = n$ and $f(n) = p^{n(n-1)/2} \ll 1$, and that $\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1}p^k$, meaning this ratios grows for a bit, and then starts shrinking. So, there exists a k_0 such that

$$f(k_0) \geq 1 > f(k_0 + 1)$$

We will show that we can accurately estimate k_0:

Lemma 1.7.1: Clique Number

For large n, we have:

$$k_0 = k_0(n) \sim \frac{2}{\log(1/p)} \log(n)$$

***Proof* :**
We'll first bound $f(k)$ from above and bellow. Notice that the binomial can be bounded by

$$\left(\frac{n-k}{k}\right)^k \leq \binom{n}{k} \leq n^k$$

and so for $f(k)$:

$$\left(\frac{n}{k} - 1\right)^k p^{k(k-1)/2} \leq f(k) \leq n^k p^{k(k-1)/2}$$

(here) for $m \geq 1$

$$\mathbb{P}(N_{k_0+m+1} > 0) \leq \frac{1}{n^{m(1-\varepsilon)}}$$

and

$$k_0 = k_0(n) = \frac{2\log(n)}{\log(1/p)}$$

2

Second Moment: Variation and Correlation

So far, the big statistical question we've been asking centered around finding the expected value of a distribution. We now will study some of the statistical information around the expected value, in particular, how concentrated are values around the expected value.

Like in the previous chapter, all distributions will be discrete, however many results easily generalize to the non-discrete case.

2.1 Variance and Covariance

Let X be a random variable. For now, the following definition is introduced for technical reasons. We will call $\mathbb{E}(X)^k$ the kth moment of X. The kth moment is defined if $\mathbb{E}|X|^k < \infty$. Note that if $\mathbb{E}X^k < \infty$ then $\mathbb{E}|X|^{k-i} < \infty$ for all $0 \leq j \leq k$. To see this, Take the case of $k = 2$. Then

$$|X| = |X|I(|X| < 1) + |X|I(|X| > 1) \leq 1 + X^2$$

And if we add absolute values to X^k, the inequality becomes even stronger.

The next concept gives us some idea of how dense values are around the expected value:

Definition 2.1.1: Variance

Let X be a random variable and $\mu = \mathbb{E}X$, $\mathbb{E}|X|^2 < \infty$. Then the *variance of* X is defined as

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

i.e. the variance is the second moment of the random variable $X - \mu$. The value $\sqrt{\text{Var}(X)}$ is called the *standard deviation*

Some common terminology is associated with the aforementioned definition. If X is any random variable, then $X - \mathbb{E}X$ is called *centering*, since $\mathbb{E}(X - \mathbb{E}X) = 0$:

$$\mathbb{E}(X - \mathbb{E}X) = \mathbb{E}(X) - \mathbb{E}X = 0$$

since $\mathbb{E}(\mathbb{E}(X)) = \mathbb{E}(X)$ since $\mathbb{E}(X)$ is a constant. For that reason, we will usually denote it as μ. In general, if $\mathbb{E}X = 0$ a random variable X, then X is called centered.

Let's say the codomain of X is $\mathbb{N}$. Notice that if most values are close to μ, then examining the expanded definition of expected value:

$$\mathbb{E}(X - \mu)^2 = \sum_{x=1}^{\infty} (x - \mu)^2 \mathbb{P}(X = x)$$

(HERE)

For computational purposes, it is often easier to represent the variance like so:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}(X - \mu)^2 \\
&= \mathbb{E}(X^2 - 2\mu X + \mu^2) \\
&= \mathbb{E}X^2 - 2\mu^2 + \mu^2 \\
&= \mathbb{E}X^2 - \mu^2 \\
&= \mathbb{E}X^2 - (\mathbb{E}X)^2
\end{aligned}$$

Thus, we get

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

Using Yensen's Inequality, we can also be used to show that if $X \geq 0$ then $\mathbb{E}X^2 \geq \mathbb{E}|X|$ using $Y = X^2$ and taking advantage of the fact that the square-root is concave.

Example 2.1: Variance Examples

1. if $X \sim B(p)$ show that $\text{Var}(X) = p(1 - p)$. What if $X \sim B(n, p)$?
2. if $X \sim \text{Poiss}(\lambda)$ show that $\text{Var}(X) = \lambda$. That means that not only is the mean λ, but it varies by λ! [Hint: Compute $\mathbb{E}(X(X - 1))$
3. If $\mathbb{E}X^2 < \infty$, then $\text{Var}(aX + b) = a^2 \text{Var}(X)$ for all $a, b \in \mathbb{R}$

The next value we will look at is a way of seeing how much two random variable make an "observable" effect on each other

Definition 2.1.2: Covariance

Let X, Y be random variables where $\mathbb{E}|XY|, \mathbb{E}|X|, \mathbb{E}|Y| < \infty$ (or equivalently $\mathbb{E}X^2, \mathbb{E}Y^2 < \infty$). Then we define the *covariance* between X and Y to be

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

The 2nd inequality comes from expanding the definition. The equivalent limitation comes from the fact that

$$|XY| \leq \frac{1}{2}(X^2 + Y^2) \leq X^2 + Y^2$$

which is an application of the Cauchy-Schwarz's Inequality (a famous result from 1st year Analysis, which we will soon translate into the language of probability). By letting either X or Y be the identity, we sese that $|X|$ and $|Y|$ are also bounded.

Lemma 2.1.1: Cauchy-Schwarz Inequality

Let X, Y be random variables defined on the same space. Then

$$\mathbb{E}|XY| \leq (EX^2)^{1/2}(EY^2)^{1/2}$$

***Proof* :**

This is the same proof but translated into statistics. if $a = \mathbb{E}X^2$, $b = \mathbb{E}XY$ and $c = \mathbb{E}Y^2$, then for all $t \in \mathbb{R}$:

$$0 \leq \mathbb{E}\left((tX - Y)^2\right) = at^2 = 2bt + c$$

For this quadratic to be non-negative, it can at most have 1 root, and so the discriminant must be:

$$D = 4b^2 - 4ac \leq 0 \quad \Leftrightarrow \quad b^2 \leq ac$$

replacing our variables with their original symbols, we get:

$$\mathbb{E}|XY| \leq \mathbb{E}X^2\mathbb{E}Y^2$$

replacing X and Y with with $|X|$ and $|Y|$ gives us our desired result.

As an exercise, show that it is equivalent to write $|\operatorname{Cov}(X,Y)| \leq \sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}$.

Example 2.2: Uncorrelated Random Variables

Let X and Y be independent. Then

$$\operatorname{Cov}(X,Y) = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y = \mathbb{E}X\mathbb{E}Y - \mathbb{E}X\mathbb{E}Y = 0$$

showing that independent random variables are uncorrelated! Thus being independent implies uncorrelated. However, being uncorrelated does not imply independent. Take for example independent random variables ε and Z such that

$$\mathbb{P}(\varepsilon = 1) = \mathbb{P}(\varepsilon = -1) = \frac{1}{2}$$

$$\mathbb{P}(Z = \pm 2) = \mathbb{P}(Z = \pm 3) = \frac{1}{4}$$

Let

$$X = Z \qquad Y = \varepsilon Z$$

Then:

$$\mathbb{E}X = 0 \qquad \mathbb{E}Y \overset{\text{indep.}}{=} \mathbb{E}\varepsilon\mathbb{E}Z = 0$$

and

$$\mathbb{E}XY = \mathbb{E}\varepsilon Z^2 \overset{\text{indep.}}{=} \mathbb{E}\varepsilon\mathbb{E}Z^2 = 0$$

Thus

$$\operatorname{Cov}(X, Y) = 0$$

However:

$$0 = \mathbb{P}(Z = 2,\ \varepsilon Z = 3) \neq \mathbb{P}(Z = 2)\mathbb{P}(\varepsilon Z = 3) = 1/4 \cdot 1/4 = 1/8$$

Thus, the two random variable are *not* independent! In other words, knowing something about one random variable gives us information about the outcome of the other!

If we have random variables $X_1, X_2, ..., X_n$ on the same probability space, and f is a linear combination of these functions, then the variance of f is very simple to compute. For simplicity, let's say $f = S_n = X_1 + X_2 + \cdots + X_n$ and $\mathbb{E}(X_i)^2 < \infty$. Then to compute $\operatorname{Var}(S_n)$, we first find:

$$\begin{aligned}(S_n - \mathbb{E}S_n)^2 &= \left(\sum_{i=1}^{n}(X_i - \mathbb{E}X_i)\right)^2 \\ &= \sum_{i,j=1}^{n}(X_i - \mathbb{E}X_i)(X_j - \mathbb{E}X_j)\end{aligned}$$

Taking the expected value of both sides gets us:

$$\operatorname{Var}(S_n) = \sum_{i,j=1}^{n}\operatorname{Cov}(X_i, X_j)$$

Since $\operatorname{Cov}(X_i, X_i) = \operatorname{Var}(X_i)$ and $\operatorname{Cov}(X_i, X_j) = \operatorname{Cov}(X_j, X_i)$, we get that:

$$\operatorname{Var}(S_n) = \sum_{i=1}^{n}\operatorname{Var}(X_i) + 2\sum_{i<j}\operatorname{Cov}(X_i, X_j)$$

If the random variables are uncorrelated, we get that:

$$\operatorname{Var}(S_n) = \sum_{i=1}^{n}\operatorname{Var}(X_i)$$

Furthermore, if they are i.i.d., then:

$$\operatorname{Var}(S_n) = n\operatorname{Var}(X_i)$$

Example 2.3: Variance of Binomial
Using what we've just proved, find $\mathrm{Var}(B(n,p))$

2.2 Law of Large Numbers

You might have heard that if you flip a coin 100 times, you will get ∼50 coins to be heads, i.e., half the coins will be heads. The idea here is that as the amount of times you repeat an experiment becomes larger, you expected the changes of the event to happen to regress to the mean. This is known as the law of Large Numbers and will be the focus of this section. In a sense, what the law of large number's is trying to gain from the world is whether there particular scenario we have was quantified in such a way that we can see patterns happen if it is repeated, in particular whether there is a particular "favoured direction" in which the scenario will lean towards (the expected value). Using the law of large numbers will be how we actually determine from the noise of repeated experiments whehter there are actuall something going on in the randomness! Naturally in the real world, it is very hard (one can even say impossible) to have an identical experiment repeated mulitple times. However, with enough control, we can get close enough to get hopefully meaningfull results.

We'll first generalize Chebyshev's ineqality to use Variances. In some textoboks, this is considered *the* Chebyshev's inequality:

Lemma 2.2.1: Chebyshev's Inequality II

Let X be a random variable and $\mathrm{Var}\,|X| < \infty$. Then for any $x > 0$

$$\mathbb{P}(|X-\mu| \geq x) \leq \frac{\mathrm{Var}(X)}{x^2}$$

The key is that the variance is bounded, so the variance *exists*, so to speak, and so we can bound our values:

Proof **:**
We can convert this to Chebyshev's inequality by making $Y = (X-\mu)^2$ and considering

$$\mathbb{P}(Y \geq y) \leq \frac{\mathbb{E}Y}{y}$$

Then substituting back in and doing some change of variables, we get:

$$\mathbb{P}((X-\mu)^2 \geq (x-\mu)^2) \leq \frac{\mathbb{E}(X-\mu^2)}{(x-\mu)^2} = \frac{\mathrm{Var}(X)}{(x-\mu)^2}$$

$$\mathbb{P}((X-\mu)^2 \geq x^2) \leq \frac{\mathrm{Var}(X)}{x^2}$$

Furthermore, notice that $\{(X-\mu)^2 \geq x^2\} = \{|X-\mu| \geq x\}$ and so:

$$\mathbb{P}(|X-\mu| \geq x) \leq \frac{\mathrm{Var}(X)}{x^2}$$

completing the proof

Essentially, the probability that the random variable X is far away from the mean gets smaller at a rate of $\mathcal{O}(x^2)$.

One more thing we need before proving the law of large numbers: if $X_1, ..., X_n$ are random variables, then let

$$\overline{X_n} := \frac{1}{n}\sum_{i=1}^{n} X_i$$

which will be our random variable that takes into consideration the “average outcome” of many events

Theorem 2.2.1: Law of Large Numbers

Let $X_1, X_2, ..., X_n$ be i.i.d. random variables and define $\overline{X}_n$ to be their average. Let $\mu = \mathbb{E}X_1$ and $\sigma^2 = \text{Var}(X_1) < \infty$. Then for any $\varepsilon > 0$

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Proof **:**

This is simply a matter of using the previous lemma, and so the details are omitted:

$$\mathbb{E}\overline{X}_n = \frac{n\mathbb{E}X_1}{n} = \mathbb{E}X_1 = \mu$$

$$\text{Var}(\overline{X}_n) = \frac{n\,\text{Var}(X_1)}{n^2} = \frac{Var(X_1)}{n} = \frac{\sigma^2}{n}$$

and so

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

Notice that no matter what ε is chosen, however small, as $n \to \infty$, $\frac{\sigma^2}{n\varepsilon} \to 0$, meaning the average of the i.i.d. random variables (i.e. the same experiment repeated several times perfectly and independently) will approach the average.

2.3 Bernstein Polynomials

It is a famous result in analysis that all continuous functions can be arbitrarily well approximated by polynomials[1]. There is a way to prove this fact using statistics!

[1] i.e. the polynomials are dense in the metric space of continuous functions

Definition 2.3.1: Bernstein Polynomials

Let $f : [0,1] \to \mathbb{R}$ be a [uniformally] continuous function. Then the *Bernstein Polynomial of order* n associated to the function f is:

$$B_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right) \binom{n}{k} x^k (1-x)^{n-k}$$

Note that f is automatically uniformally continuous since it's domain is compact.

Theorem 2.3.1: Weiestrass Approximation Theorem

If $f[0,1] \to \mathbb{R}$ is continuous function, then:

$$\lim_{n\to\infty} \max_{x\in[0,1]} |f(x) - B_n(x)| = 0$$

***Proof* :**

Let $p \in [0,1]$, $X_1, ..., X_n$ be i.i.d. Bernoulli random variables. $B(p)$, and

$$S_n = X_1 + \cdots + X_n \qquad \overline{X}_n = \frac{S_n}{n}$$

Since S_n has a binomial distribution, we have:

$$\mathbb{E}(f(\overline{X}_n)) = \mathbb{E}\left(f(\frac{S_n}{n})\right) = \sum_{k=1}^{n} f\left(\frac{k}{n}\right)\binom{n}{k} p^k (1-p)^{n-k}$$

which is equal to the Bernstein polynomial evaluated at $x = p$: $B_n(x)$. In other words, the Bernstein polynomials are the expected value of taking the "average" of random selections which are put through f. Now, consider

$$\begin{aligned} |B_n(p) - f(p)| &= |\mathbb{E}f(\overline{X}_n) - f(p)| \\ &= |\mathbb{E}(f(\overline{X}_n) - f(p))| \\ &\leq \mathbb{E}|(f(\overline{X}_n) - f(p))| && \text{monotonicity} \\ &= \mathbb{E}\left[|(f(\overline{X}_n) - f(p))| I(|\overline{X}_n - p| \leq \varepsilon) + |(f(\overline{X}_n) - f(p))| I(|\overline{X}_n - p| > \varepsilon)\right] \end{aligned}$$

the last line is a common trick in many cases where a function acts differently on two different parts (ex. For x^2, if $x \geq 1$ then $x^2 \geq x$, but if $0 < x < 1$ then $x^2 < x$). For the first term, since f is differentiable, we can bound it by the modulus of continuity:

$$\Delta(\varepsilon) = \max_{|x-y|\leq\varepsilon} |f(x) - f(y)|$$

For the second term, we use uniform continuity again to use the fact that there exists a C such that $|f| \leq C$, so we get

$$\mathbb{E}|f(\overline{X}_n) - f(p)| \leq \mathbb{E}\left[\Delta(\varepsilon) + 2C \cdot I(|\overline{X}_n - p| > \varepsilon)\right] = \Delta(\varepsilon) + 2C\mathbb{P}(|\overline{X}_n - p| > \varepsilon)$$

Finally, we're at a position were we can use Chebyshev's inequality. First, $\mathbb{E}(B_n(p)) = n\frac{\mathbb{E}(B(p))}{n} = p$. Next, since the variance of the Bernoulli random variable is $\operatorname{Var}(B(p)) = p(1-p) \leq 1/4$ (i.e., is bounded by a fair coin), we have $\operatorname{Var}(B_n(p)) = \frac{n}{n^2}\operatorname{Var}(B(p)) = \frac{1}{4n} \leq \frac{1}{4}$. Combining these, we get:

$$|B_n(p) - f(p)| \leq \Delta(\varepsilon) + \frac{C}{2n\varepsilon^2}$$

Notice that p is not in the upper bound. So this is true for all p, including the one that maximizes the equation:

$$\max_{p\in[0,1]} |B_n(p) - f(p)| \leq \Delta(\varepsilon) + \frac{C}{2n\varepsilon^2}$$

And taking $n \to \infty$:

$$\lim_{n\to\infty} \max_{p\in[0,1]} |B_n(p) - f(p)| \leq \Delta(\varepsilon)$$

and since f is uniformly continuous, as $\varepsilon \to 0$, $\Delta(\varepsilon) \to 0$, and since this true for all $\varepsilon > 0$, it must be that:

$$\lim_{n\to\infty} \max_{p\in[0,1]} |B_n(p) - f(p)| = 0$$

as we sought to show

2.4 Cliques on Erdös Reyni Graph

In section 1.7.3 we found that the expected number of cliques of size k is:

$$\mathbb{E}N_k = \binom{n}{k} p^{\binom{k}{2}}$$

and that the number of cliques grow up to the tipping point of $f(k_0) \leq 1 > f(k_0+1)$ where the number of expected clicks of size k becomes less than 1 is approximately $\frac{2}{\log(1/p)}\log(n)$. At this point, we had the probability of finding a click to be:

$$\mathbb{P}(N_{k_0+m+1} > 0) \leq \frac{1}{n^{m(1-\varepsilon)}}$$

showing that it's is very unlikely of having a click of size $k_0 + 2$ or bigger. Overall, we get a good idea of what happens when $k > k_0$. We had little to analyze what happens when $k < k_0$. Using the variance, we can take a peek at what happens on the other side. It will turn out that N_k will concentrate, similar to the law of large numbers, meaning that the number of such cliques is not only large, but typical (i.e. with probability close to 1).

We first consider a $k < k_0$. In particular, taking a small $\varepsilon > 0$, we consider:

$$\frac{(2-\varepsilon)\log(n)}{\log(1/p)} \leq k < k_0 \tag{2.1}$$

To simplify the equation, let's exponentiate both sides. Manipulating the result around, we get:

$$p^k \leq \frac{1}{n^{w-\varepsilon}}$$

This might seem familiar to when we considered the ratio of $f(k+1)/f(k)$:

$$\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1}p^k \leq np^k \leq \frac{1}{n^{1-\varepsilon}}$$

Since $f(k_0) \geq 1$, we get that $f(k_0 - 1) \geq n^{1-\varepsilon}$, $f(k_0 - 2) \geq N^{2(1-\varepsilon)}$ and by induction $f(k_0 - m) \geq n^{m(1-\varepsilon)}$. In particular, this means that the number of cliques of size k in the range of the inequalities in equation (2.1) is quite large. To get more detailed results, we will compute the variance of N_k and apply Chebyshev's inequality.

To get more data, we will introduce a new trick in applying Chebyshev's inequality:

Trick Let's say $\mathbb{E}X > 0$, and take $x = \delta \mathbb{E}X$. Take some $\delta > 0$ and consider:

$$\mathbb{P}(|X - \mathbb{E}X| \geq \delta \mathbb{E}X) \geq \frac{\text{Var}(X)}{\delta^2 (\mathbb{E}X)^2}$$

usually, $\delta \ll 1$. The key comes if the variance is very small as compared to $\mathbb{E}(X)^2$: $\text{Var}(X) \ll \mathbb{E}(X)^2$. Then since we can re-write $|X - \mathbb{E}X| < \delta \mathbb{E}X$ as:

$$1 - \delta \leq \frac{X}{\mathbb{E}X} \leq 1 + \delta$$

then the probability of the compliment of the Chebyshev inequality presented earlier is:

$$\mathbb{P}(1 - \delta \leq \frac{X}{\mathbb{E}X} \leq 1 + \delta) \geq 1 - \frac{\text{Var}(X)}{\delta^2 (\mathbb{E}X)^2}$$

here

3

Exponential Inequalities

So far, the best we have been able to bound probabilities is with the use of Chebyshev's inequality. We used it when analyzing the sum of i.i.d. random variables with $\mathbb{E}X^2 < \infty$ to get:

$$\mathbb{P}(|\overline{X}_n - \mu| \geq \varepsilon) \leq \frac{\operatorname{Var}(X)}{n\varepsilon^2}$$

The key assumption here is that $\mathbb{E}X^2 < \infty$. If we can assume the decay is faster (which, in fact, it often is!), then we can get a much stronger bound. In particular, if

$$\mathbb{E}e^{\lambda X} < \infty \quad \text{for all } \lambda \in \mathbb{R}$$

then we can get some much stronger bounds; the error of this bound is much better, since e^{λ} for negative lambda shrinks much faster than $1/x^2$.

3.1 Markov and Hoeffding Inequality

This bound is not arbitrary: notice that if a random variable X is bounded by a constant, then this condition holds.

> **Example 3.1: Bounded Random Variables**
> Let $X < c$. Then
>
> $$\mathbb{E}X \leq \mathbb{E}c = c$$
>
> and so
>
> $$\mathbb{E}e^{X\lambda} \leq \mathbb{E}e^{c\lambda} = e^{c\lambda}$$

Furthermore, the exponential is one of the quicker shrinking functions we have at our disposal: it shrinks faster than any polynomial time. Thus, if we find ourselves being able to study random variables that grow rather quickly, we can bound our function even further:

Theorem 3.1.1: Markov's Inequality

Let X be a random variable with $\mathbb{E}e^{\lambda X} < \infty$ for all $\lambda \geq 0$. Then:

$$\mathbb{P}(X \geq t) \leq e^{-\lambda t}\mathbb{E}e^{\lambda X}$$

or more concisely, since $\mathbb{P}(X \geq t)$ is independent of any lambda, we can write:

$$\mathbb{P}(X \geq t) \min_{\lambda \in \mathbb{R}_{\geq 0}} \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda t}}$$

***Proof* :**
If $\lambda = 0$, the right hand side is equal to 1, so the equality trivially holds. Thus, assume $\lambda > 0$. Since e^x is a monotone function, $X \geq t$ if and only if $e^{\lambda X} \geq e^{\lambda t}$. Furthermore, e^x is strictly positive, and so we can apply the non-negative version of Chebyshev's inequality (the one without the indicator function):

$$\mathbb{P}(X \geq t) = \mathbb{P}(e^{\lambda X} \geq e^{\lambda t}) \overset{\text{C.}}{\leq} \frac{\mathbb{E}e^{\lambda X}}{e^{\lambda t}}$$

completing the proof.

Keep the minimum formula in the back of your mind when working with Markov's inequality; we will often take advantage of this fact. As mentioned earlier, many random variables satisfy the exponential condition. As a key example, let ε_i be a random variable such that:

$$\mathbb{P}(\varepsilon = -1) = \mathbb{P}(\varepsilon = 1) = 1/2$$

that is, it is the Bernoulli random variable, but modified so that $\mathbb{E}\varepsilon_i = 0$. This is called the *Rademacher* random variable. Now, consider collection of i.i.d. ε_i and the new random variable:

$$X = \sum_{i=1}^{n} a_i \varepsilon_i$$

for some arbitrary $a_1, ..., a_n \in \mathbb{R}$. This random variable takes on as values all possible ± combinations of the elements a_i. As a use of such a random variable, imagine there are n events for which if you win, you get a_i amount of money, and if you loose, you give a_n amount of money. In each event, you have a 50% chance of winning. Then this random variable captures the amount of money you will win. Notice that this is a bounded random variable, and so we can apply Markov's inequality and see that the chances that all the signs are positive is in fact rather slim!

Theorem 3.1.2: Hoeffding's Inequality

Let $\varepsilon_1, \varepsilon_2, ..., \varepsilon_i$ be i.i.d Rademacher random variables and $a_1, a_2, ..., a_n \in \mathbb{R}$. Then for any $t \geq 0$

$$\mathbb{P}\left(\sum_{i=1}^{n} \varepsilon_i a_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}\right)$$

and

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} \varepsilon_i a_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^{n} a_i^2}\right)$$

Proof **:**

First, we can directly apply Markov's inequality:

$$\mathbb{P}\left(\sum_{i=1}^{n} \varepsilon_i a_i \geq t\right) \leq \frac{\mathbb{E}e^{(\lambda \sum_{i=1}^{n} \varepsilon_i a_i)}}{e^{\lambda t}} = e^{-\lambda t}\mathbb{E}\prod_{i=1}^{n} e^{(\lambda \varepsilon_i a_i)}$$

Next, applying independence of the ε_i, by induction on proposition 1.5.1 ($\mathbb{E}XY = \mathbb{E}X\mathbb{E}Y$), we have:

$$e^{-\lambda t}\mathbb{E}\prod_{i=1}^{n} e^{(\lambda \varepsilon_i a_i)} = e^{-\lambda t}\prod_{i=1}^{n} \mathbb{E}e^{(\lambda \varepsilon_i a_i)}$$

At this point, we can directly compute the expected value of each $\mathbb{E}e^{\lambda \varepsilon_i}$, since we can then use some knowledge from complex analysis to equate this value to a known function:

$$\mathbb{E}e^{\lambda \varepsilon_i} = \frac{1}{2}e^{\lambda a_i} + \frac{1}{2}e^{-\lambda a_i} = \cosh(\lambda a_i)$$

Since we have that every expected value is related to $\cosh(x)$, we can use some Taylor series to show that $\cosh(x) \leq e^{x^2/2}$:

$$\cosh(x) = \frac{e^x + e^{-x}}{2} = \sum_{k=0}^{\infty} \frac{1}{(2k)!}x^{2k} \leq \sum_{k=0}^{\infty} \frac{1}{2^k k!}x^{2k} = e^{x^2/2}$$

Thus, for every ε_i, we get that $\mathbb{E}e^{\lambda \varepsilon_i a_i} \leq e^{\frac{\lambda^2 a_i^2}{2}}$. Substituting back into our equation, we get:

$$\mathbb{P}\left(\sum_{i=1}^{n} \varepsilon_i a_i \geq t\right) \leq \exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{n} a_i^2\right)$$

This inequality is true over all λ, and so we can minimize it over our choice of λ:

$$\mathbb{P}\left(\sum_{i=1}^{n} \varepsilon_i a_i \geq t\right) \leq \min_{\lambda \in \mathbb{R}_{\geq 0}}\left\{\exp\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{n} a_i^2\right)\right\}$$

To find this, we take advantage our some calculus knowledge. In particular, we start by finding the critical points of the value in the exponent (since e^x is monotone, it is equivalent to minimize the exponent). Thus:

$$\frac{d}{d\lambda}\left(-\lambda t + \frac{\lambda^2}{2}\sum_{i=1}^{n} a_i^2\right) = -t + \lambda\sum_{i=1}^{n} a_i^2 \quad \Leftrightarrow \quad \lambda = \frac{t}{\sum_{i=1}^{n} a_i^2}$$

and so, plugging this value back in, we get:

$$\exp\left(-t\frac{t}{\sum_{i=1}^n a_i^2}+\frac{\left(\frac{t}{\sum_{i=1}^n a_i^2}\right)^2}{2}\sum_{i=1}^n a_i^2\right)=\exp\left(-\frac{t^2}{\sum_{i=1}^n a_i^2}+2\frac{t^2}{(\sum_{i=1}^n a_i^2)^2}\sum_{i=1}^n a_i^2\right)=\exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right)$$

Thus:

$$\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \geq t\right)\leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right)$$

proving the first inequality. For the second inequality, we take advantage of the symmetry of ε_i. In particular, the random variable $(-\varepsilon_i)$ has the same distribution:

$$\mathbb{P}(-\varepsilon_i=1)=\mathbb{P}(-\varepsilon_i=-1)=1/2$$

and so:

$$\mathbb{P}\left(\sum_{i=1}^n -\varepsilon_i a_i \geq t\right)=\mathbb{P}\left(\sum_{i=1}^n \varepsilon_i a_i \leq -t\right)\leq \exp\left(-\frac{t^2}{2\sum_{i=1}^n a_i^2}\right)$$

Since:

$$\left\{\left|\sum_{i=1}^n a_i\varepsilon_i\right|\geq t\right\}=\left\{\sum_{i=1}^n a_i\varepsilon_i\geq t\right\}\cup\left\{\sum_{i=1}^n a_i\varepsilon_i\leq -t\right\}$$

Then since $\mathbb{P}(A\cup B)\leq\mathbb{P}(A)+\mathbb{P}(B)$ (i.e., the union bound), we have the desired result.

As an immediate use-case of this inequality, let's improve our bound from the law of large numbers when we have i.i.d. Bernoulli random variables:

Example 3.2: Law of Large Numbers with fair coins

In theorem 3.1.2, if we let $a_i=\frac{1}{n}$ for all a_i, we get:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i\right|\geq t\right)\leq 2\exp\left(-\frac{n^2t^2}{2\sum_{i=1}^n 1}\right)=2e^{\frac{nt^2}{2}}$$

We can relate this to $\overline{X}_n=\frac{1}{n}\sum X_i$ where the X_i are i.i.d. Bernoulli random variables by taking $X_i=(\varepsilon_i+1)/2$. Since each ε_i are i.i.d, each X_i is i.i.d, and further more $X_i\sim B(1/2)$. Isolating $\varepsilon_i=2X_i-1$ and substituting back in, we get:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n 2X_i-1\right|\geq t\right)=\mathbb{P}\left(\left|\overline{X}_n-1/2\right|\geq\frac{t}{2}\right)\leq 2e^{\frac{nt^2}{2}}$$

Substituting $\varepsilon=t/2$, we get:

$$\mathbb{P}\left(\left|\overline{X}_n-1/2\right|\geq\varepsilon\right)\leq 2e^{-2n\varepsilon^2}$$

If we were to do Chebyshev's inequality, we get:

$$\mathbb{P}\left(\left|\overline{X}_n-1/2\right|\geq\varepsilon\right)\leq\frac{1}{4n\varepsilon^2}$$

Notice how much stronger our new inequality is! Let's take $n=10000$ and $\varepsilon=0.02$. Then the two bounds we get are:

$$0.00067 \qquad \text{vs.} \qquad 0.0625$$

meaning the Hoeffding bound is over 100 times smaller than what we get with Chebyshev!

We might further ask if we can estimate any better than this. We will ask this question again when we cover the Central Limit Theorem, at which point we'll show that n and ε on the right hand side is almost as good as it gets.

Another use of the inequality comes from student higher moments of the Rademacher Random variables:

Example 3.3: Higher moments of Rademacher RV

Let $a_1, a_2, ..., a_n \in \mathbb{R}$ such that:

$$a_1^2 + a_2^2 + \cdots + a_n^2 = 1$$

such a linear combination is called a *convex combination* since if $v_1, v_2, ..., v_n$ are vector's in a vector space, if we take the convex hull of these vectors, if the condition satisfies the aforementioned equality, the vector's will lie in side the convex hull (I think? that condition is usually satisfied on the sum of the coefficients, not the sum of the square...)

Take the random variable:

$$X := a_1\varepsilon_1 + \cdots + \alpha_n\varepsilon_n$$

Take a moment to see that all of it's odd moments are equal to zero:

$$\mathbb{E}X^{2k+1} = \mathbb{E}(a_1\varepsilon_1 + \cdots + \alpha_n\varepsilon_n)^{2k+1} = 0$$

It's second moment is:

$$\mathbb{E}X^2 = \mathbb{E}(a_1\varepsilon_1 + \cdots + \alpha_n\varepsilon_n)^2 = a_1^2 + \cdots + a_n^2 = 1$$

We'll use Hoeffding's inequality bound the other even moments. Using the tail definition of the expected value (proposition 1.3.3 and exercise ref:HERE (it was an exercise in the book, see p.110):

$$\mathbb{E}X^{2k} = \int_0^\infty 2kt^{2k-1}\mathbb{P}(|X| \geq t)dt \overset{!}{\leq} \int_0^\infty 4kt^{2k-1}e^{-t^2/2}dt$$

where we used Hoeffding's inequality in $\overset{!}{\leq}$ by replacing $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2}$. If we set $t = \sqrt{2u}$, we get:

$$\mathbb{E}X^{2k} \leq k2^{k+1} \int_0^\infty u^{k-1}e^{-u}du = k2^{k+1}\Gamma(k) = k2^{k+1}(k-1)! = 2^{k+1}k!$$

where we recognized the integral is the gamma function, and we can evaluate the Gamma function to be $(k-1)!$ (we will in fact take lot's of time to explore this function in the next chapter, so no need to worry about not recognizing it at this stage).

Finally, we leave off this chapter with some build-up for the next section. Given Rademacher random variables $\varepsilon_1, ..., \varepsilon_n$, let:

$$Y = X^2 - 1 = (a_1\varepsilon_1 + \cdots + a_n\varepsilon_n)^2 - 1$$

We will use the random variable to do (HERE). We'll be doing some bounding with it that requires the following lemma:

Lemma 3.1.1

Let $a_1, ..., a_n \in \mathbb{R}$ such that $a_1^2 + \cdots + a_n^2 = 1$, and let Y be defined as above. Then for any $0 \leq \lambda \leq \frac{1}{4}$,

$$\mathbb{E}e^{\lambda Y} \leq e^{16\lambda^2} \text{ and } \mathbb{E}e^{-\lambda Y} \leq e^{16\lambda^2}$$

***Proof* :**
We introduce another bounding trick. Taking the Taylor series for e^x, notice that:

$$e^x \leq 1 = x + \frac{x^2}{2} + \sum_{k=3}^{\infty} \frac{(x_+)^k}{k!}$$

where $x_+ = \max(x, 0)$. If $x \geq 0$, then equality holds in the above equation. However, if $x \leq 0$, then since e^x is monotone and strictly increasing, we get $e^x \leq 1 + x + \frac{x^2}{2}$. Thus, with this bound we can use countable linearity of the expected value to write:

$$\mathbb{E}e^{\lambda Y} \leq 1 + \lambda \mathbb{E}Y + \frac{\lambda^2}{2}\mathbb{E}Y^2 + \sum_{k=3}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(Y_+)^k$$

To get better results, we start decomposing our variables: since $\mathbb{E}X^2 = 1$ (as shown in example 3.3, then $\mathbb{E}Y = \mathbb{E}X^2 - 1 = 1 - 1 = 0$. On the other hand, $\mathbb{E}Y^2 = \mathbb{E}(X^2 - 1)$, and so

$$\mathbb{E}Y^2 = \mathbb{E}X^4 - 2\mathbb{E}X^2 + 1 = \mathbb{E}X^4 - 1 \leq \mathbb{E}X^4$$

which gives us an idea for the first two terms. For the infinite summand, notice that $Y_+ = \max(X^2 - 1, 0) \leq X^2$, so we have $\mathbb{E}(Y_+)^k \leq \mathbb{E}X^{2k}$. Putting this all together, we get:

$$\mathbb{E}e^{\lambda Y} \leq 1 + \frac{\lambda^2}{2}\mathbb{E}X^4 + \sum_{k=3}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}X^{2k} 1 + \sum_{k=2} \frac{\lambda^k}{k!}\mathbb{E}X^{2k}$$

At this point, we can again use the inequality for even moments of the random variable X from example 3.3 to get:

$$\begin{aligned}
\mathbb{E}e^{\lambda Y} &\leq 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!} 2^{k+1} k! \\
&= 1 + \sum_{k=2}^{\infty} \lambda^k 2^{k+1} \\
&= 1 + \sum_{k=2}^{\infty} (2\lambda)^k 2^k \\
&\overset{!}{=} 1 + \frac{8\lambda^2}{1 - 2\lambda} \\
&\leq 1 + 16\lambda^2
\end{aligned}$$

where $\overset{!}{=}$ is since $2\lambda \leq \frac{1}{2}$, finishing the proof. For the second inequality, we would go the other way by taking $1 + x \leq e^x$.

3.2 Johnson-Lindenstrauss Lemma

In this section, we are going to take a detour to show an application of the previous result to prove a really important result that comes up often in the analysis of big data and functional analysis.

Let $N \geq 1$, and set pick m points in $\mathbb{R}^N$:

$$V = \{v_1, ..., v_m\} \subseteq \mathbb{R}^N$$

Usually, we think of N being able to be arbitrarily large and m being large. What the Johnson-Lindenstrauss lemma tells us is that there exists a *linear map* $\mathbb{R}^N$ into $\mathbb{R}^n$ (where n is possibly of much lower dimension) that preserves the distances between all points in the set V up to a small relative error! This is called a *low-distortion embedding*. It was discovered by Johnson and Lindenstrauss while studying functional analysis, but it has been very useful in computational algorithms as a pre-processing tool to reduce the number of dimensions of high-dimensional data (which is becoming important). We formalise this idea in the following theorem:

Theorem 3.2.1: Johnson-Lindenstrauss Lemma

Let $\{v_1, ..., v_m\} \subseteq \mathbb{R}^N$ be m points in $\mathbb{R}^N$. Then for any $\varepsilon \in (0, 1)$ where

$$n > \frac{128}{\varepsilon^2} \log(m)$$

there exits a linear map $f : \mathbb{R}^N \to \mathbb{R}^n$ such that

$$\sqrt{1-\varepsilon} \leq \frac{\|f(v_k) - f(v_l)\|}{\|v_k - v_l\|} \leq \sqrt{1+\varepsilon}$$

for all $1 \leq k < l \leq m$, with $\|\cdot\|$ being the euclidean norm.

Notice that the dimension n is only dependent on m and ε, and not N. This means that this estimate can even work if N is small, however in most cases we are more interesting in having a large N and embedding into dimensions. Furthermore, the dependence on m is logarithmic, meaning that in general, increasing the number of vector only slightly changes the bound on n. Finally, the number 128 is not optimized. It can be shown that it can be diminished to 8, but such a bound is harder to prove.

For this proof, we will define a linear map with random variable as entries (i.e. a random matrix)

$$\mathcal{E} = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1N} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nN} \end{pmatrix}$$

where ε_{ij} are i.i.d. Rademacher random variables. We'll show that there exists an n satisfying the inequality such that:

$$f(x) = \frac{1}{\sqrt{n}} \mathcal{E} x$$

will satisfy the low distortion property and exist with positive probability. That is, there exists at least one matrix with ± 1 entries that satisfy the bounding equation given a properly chosen dimension n.

Before giving a proof, we will reformulate the bounds. Let $1 \leq k < l \leq m$ and define

$$a_{kl} = \frac{v_k - v_l}{\|v_k - v_l\|}$$

that is, each a_{kl} represents the difference between any two vectors v_k and v_l which is normalized to have length 1. Then if we square the main bound and replace f with the linear function we define above, we get:

$$1 - \varepsilon \leq \frac{1}{n}\|\mathcal{E}a_{kl}\|^2 \leq 1 + \varepsilon$$

Since each a_{kl} was rescaled so that $\|a_{kl}\| = 1$, then each a_{kl} belong to the unit sphere $a_{kl} \in S^{N-1} \subseteq \mathbb{R}^N$ We will further manipulate this so that the values we want will cluster (just like in the law of large number's). To get a feeling of how the Large numbers apply, consider an arbitrary vector

$$a = (a_1, ..., a_N) \in S^{N-1}$$

and denote the ith coordinate of $\mathcal{E}a$ by

$$X_i(a) := (\mathcal{E}a)_i = \varepsilon_{i1}a_1 + \cdots + \varepsilon_{iN}a_N$$

and with this denote:

$$Y_i(a) := X_i^2(a) - 1 = (\varepsilon_{i1}a_1 + \cdots + \varepsilon_{iN}a_N)^2 - 1$$

With the new X_i notation, we can re-write our bounds again as:

$$1 - \varepsilon \leq \frac{1}{n}\sum_{i=1}^{n} X_i^2(a) \leq 1 + \varepsilon$$

and if we re-write as Y_i, we get:

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a)\right| \leq \varepsilon$$

At this point, this inequality is starting to look like the law of large numbers for $Y_1(a), ..., Y_n(a)$ since $\mathbb{E}Y_i(a) = 0$ as we've shown in last section. Even better, we've managed to bound these inequalities in lemma 3.1.1 for $0 \leq \lambda \leq \frac{1}{4}$:

$$\mathbb{E}e^{\lambda Y_i(a)} \leq e^{16\lambda^2} \quad \text{and} \quad \mathbb{E}e^{-\lambda Y_i(a)} \leq e^{16\lambda^2}$$

Furthermore, notice that by the grouping lemma, the random variables $Y_1(a), ..., Y_n(a)$ are independent since the rows of the matrix $\mathcal{E}$ are all independent by construction (as a high-level way of seeing this, if you choose values for one row of the random matrix, it tells you nothing about what values you'll choose for the other row of the matrix)

Using this inequality and independence, we get to prove the following lemma:

Lemma 3.2.1: Bounding $Y_i(a)$

Let $Y_i(a), ..., Y_n(a)$ be the random variables we defined above, $\varepsilon \in (0, 1)$ and $a = (a_1, ..., a_N) \in S^{N-1}$. Then:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a)\right| \geq \varepsilon\right) \leq 2e^{\frac{-n\varepsilon^2}{64}}$$

***Proof* :**
Just like for the Hoeffding inequality, we will drop the absolute values and use the union bound to get our final value. To bound without the union bound, we can use Markov's inequality: for any $0 \leq \lambda \leq \frac{1}{4}$

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i(a) \geq \varepsilon\right) &= \mathbb{P}\left(\sum_{i=1}^{n} Y_i(a) \geq n\varepsilon\right) \\
&\leq \min_{\lambda\in[0,1/4]} e^{-n\lambda\varepsilon}\mathbb{E}(e^{\lambda \sum_{i=1}^{n} Y_i(a)}) \\
&= \min_{\lambda\in[0,1/4]} e^{-n\lambda\varepsilon}\prod_{i=1}^{n}\mathbb{E}e^{\lambda Y_i(a)} && \text{independence} \\
&\leq \min_{\lambda\in[0,1/4]} e^{-n\lambda\varepsilon+16n\lambda^2} && \text{lemma 3.1.1} \\
&\overset{!}{=} e^{-n\varepsilon^2/64} && \lambda = \frac{\varepsilon}{32}
\end{aligned}$$

where the $\overset{!}{=}$ comes finding the critical point with the further restriction that $\lambda \in (0, 1/4)$.

The exact same thing works if we do $-Y_i(a)$ as we've done with Hoeffding inequality, and so using the union bound, we get the final result:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a)\right| \geq \varepsilon\right) \leq 2e^{\frac{-n\varepsilon^2}{64}}$$

as we sought to show

Using this lemma, we can prove Johnsons-Lindenstrauss Lemma

***Proof* :**
The goal is to show that there exists a matrix $\mathcal{E}$ that works with the given restriction in Johnsons-Lindenstrauss Lemma with non-zero probability. As we've shown, we can characterize the effect of the rows of the random matrix on a random element on the unit sphere, re-write the inequality as:

$$\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{kl})\right| \leq \varepsilon$$

for all $1 < k < l \leq m$. What we'll take the compliment of the above set, use the union bound to deal with image of vectors $v_1, ..., v_m$ from of the random matrix separately (since the image is represented through the random variables Y_i, we already understand their exponential expected value), and then take the compliment again and see what restrictions we need to put on our parameters, namely m and ε, so that the probability is positive. First, we take the compliment and split up the set into individual Y_i random variables:

$$\left\{\exists k < l \text{ such that } \left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{kl})\right| \geq \varepsilon\right\} = \bigcup_{1\leq k<l\leq m}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{kl})\right| \geq \varepsilon\right\}$$

Using the union bound and the previous lemma, we get that:

$$\sum_{1\leq k<l\leq m} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{kl})\right| \geq \varepsilon\right) \leq \sum_{1\leq k<l\leq m} 2e^{-n\varepsilon^2/64} \leq m^2 e^{-n\varepsilon^2/64}$$

Thus, taking the compliment again, we have that:

$$\mathbb{P}\left(\forall 1 \leq k < l \leq m, \left|\frac{1}{n}\sum_{i=1}^{n} Y_i(a_{ik})\right| \leq \varepsilon\right) \geq 1 - m^2 e^{-n\varepsilon^2/64}$$

This probability is strictly positive if $m^2 e^{-n\varepsilon^2/64} < 1$, which is equivalent to writing:

$$n > \frac{128}{\varepsilon^2}\log(m)$$

which is exactly the assumption we have on n! Thus, given such an n, we can find a matrix with positive probability, as we sought to show

As a final observation, notice that in the end we had $m^2 e^{-n\varepsilon^2/63} < 1$, and so there is a balance between m and n to make the value less than 1. However, since the n is in the exponential decay, by increasing the dimension of n, we can make sure that the random matrix we picked will yield a low-distortion embedding with not just positive probability, but with high probability (i.e. close to 1).

3.3 Hoeffding-Chernoff Inequality

In the proof of the Hoeffding inequality, we relied on the symmetry of the Rademacher random variables. Therefore, the same calculation does not apply for, say, a biased coin flip $B(p)$ $(p \neq 1/2)$. We want to generalize our result, so we need to replace the i.i.d. restriction with a more general one. Since many random variables are bounded, we will replace the i.i.d. of random variables to the sum of bound independent random variables. Since the random variables are bounded, they can all be re-scaled to be within $[0, 1]$.

Let $X_1, ..., X_n$ be independent random variables taking values in $[0, 1]$ with common expectation $p = \mathbb{E}X_i \in [0, 1]$. Notiec that they are not identically distributed, and so they may have different distributions. In particular, there will be a bias now towards one part of the distribution. This is captured in the following concept:

Definition 3.3.1: Kullback-Leibler Divergence

Let $p \in (0, 1)$ and $q \in [0, 1]$. Then define:

$$D(q||p) := q\log\left(\frac{q}{p}\right) + (1-q)\log\left(\frac{1-q}{1-p}\right)$$

This function is called the *Kullback-Leibler divergence* or *relative entropy*, where if $q \in \{0, 1\}$, it is understood that $0 \cdot \log(0) = 0$

The last condition guarantee's that $D(q||p)$ is continuous. $D(q||p)$ is *almost* like a metric, as in:

1. $D(p||q) \geq 0$ and $D(p||q) = 0$ if and only if $p = q$

> ***Proof* :**
> Since $-\log(x)$ is convex, $-\log(x) \geq 1 - x$ (its tangent line at $x = 1$ with equality only at $x = 1$. Then:
>
> $$\begin{aligned} D(q||p) &= q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} \\ &= q\left(-\log \frac{p}{q}\right) + (1-q)\left(-\log \frac{1-p}{1-q}\right) \\ &\geq 1\left(1 - \frac{p}{q}\right) + (1-q)\left(1 - \frac{1-p}{1-q}\right) = 0 \end{aligned}$$
>
> and equality only when $p = q$

2. $D(p + q||r) \leq D(p||r) + D(q||r)$

> ***Proof* :**
> Will leave as an exercise

However, $D(p||q) \neq D(q||p)$. This therefore defines a *quasi-metric* – a metric that is not symmetric. The way I like to think about it is imagine a metric is capturing the energy required to go from point a to point b on a flat surface. However, if you surface has hills and dips, then the energy required to go from point a to point b might be different from the energy required for going from point b to point a (due to relative entropy). $D(p||q)$ captures a particular landsacpe:

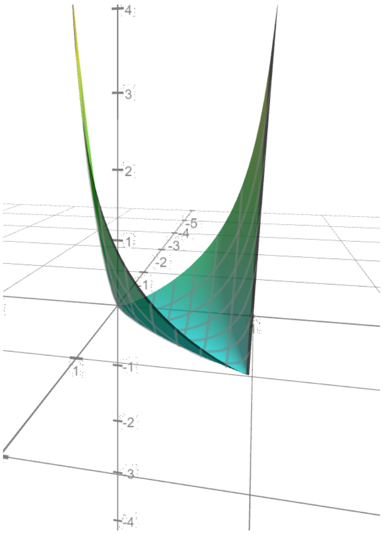


Figure 3.1: $D(p||q)$ one angle

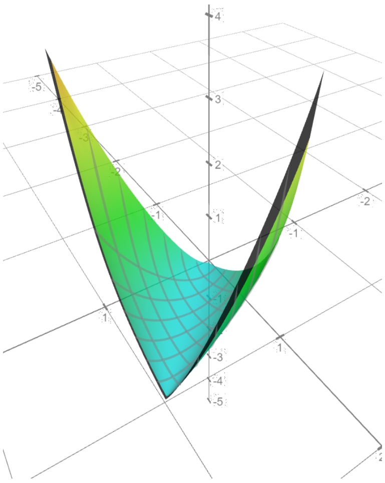


Figure 3.2: $D(p||q)$ another angle

Using this, we can generalize Hoeffding's inequality:

Theorem 3.3.1: Hoeffding-Chernoff's Inequality

Let $X_1, ..., X_n$ be bounded independent random variables where $X_i \in [0,1]$ and $p = \mathbb{E}X_i$ for each i. Then:

$$\mathbb{P}(\overline{X}_n \geq p + t) \leq e^{-nD(p+t||p)}$$

for any $0 \leq t \leq 1 - p$ and

$$\mathbb{P}(\overline{X}_n \geq p - t) \leq e^{-nD(p+t||1-p)}$$

for any $0 \leq t \leq p$

Just like Hoeffding's inequalities, the n is in the exponent, so these are again much better bounds than Chebyshev's inequalities. Furthermore, notice that since the average is always less than 1, the probability that $p + t > 1$ is 0, so the restriction that $t \leq 1 - p$ is in fact superfluous (and similarly for the $t \leq p$ constraint).

Furthermore, the facto $D(p+t||p)$ in the upper tail bound is not the same as $D(1-p+t||1-p)$ in the exponent in the lower tail bound, so these bounds are *not* symmetric. This should make sense, since heuristically, when $p < 1/2$, there is "less room" between 0 and p than above p between p and 1. We can make the proof easier by simplifying (though slightly worsening) the bounds to:

$$\mathbb{P}(\overline{X}_n \geq p + t) \leq e^{-2nt^2}$$

and

$$\mathbb{P}(\overline{X}_n \geq p - t) \leq e^{-2nt^2}$$

Proof **:**
First, apply Markov's inequality: for any $\lambda \geq 0$,

$$\begin{aligned}\mathbb{P}(\overline{X}_n \geq p + t) &= \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq n(p+t)\right)\\ &\leq \min_{\lambda} e^{-\lambda n(p+t)} \mathbb{E}e^{\lambda \sum_{i=1}^{n} X_i}\\ &= \min_{\lambda} e^{-\lambda n(p+t)} \prod_{i=1}^{n} \mathbb{E}e^{\lambda X_i}\end{aligned}$$

At this point, we take advantage of convexity of e^x to apply Jensen's inequality to get rid of the x in the exponent: for $x \in [0,1]$

$$e^{\lambda x} = e^{x\lambda + (1-x)\cdot 0} \leq xe^{\lambda} + (1-x)e^0 = 1 - x + xe^{\lambda}$$

Since each $X_i \in [0,1]$, we have that:

$$\mathbb{E}e^{\lambda X_i} \leq 1 - \mathbb{E}X_i + \mathbb{E}X_i e^{\lambda} = 1 - p + pe^{\lambda}$$

thus:

$$\min_{\lambda} e^{-\lambda n(p+t)} \prod_{i=1}^{n} \mathbb{E}e^{\lambda X_i} \leq \min_{\lambda} e^{-\lambda n(p+t)} (1 - p + pe^{\lambda})^n$$

To minimize λ, notice that since e^x and log is monotone, minimizing the log of the expression is equivalent:

$$\min_{\lambda}(-\lambda n(p+t) + n\log(1-p+pe^{\lambda}))$$

which we do by finding the critical point as we've done in similar proofs:

$$-n(p+t) + \frac{npe^{\lambda}}{1-p+pe^{\lambda}} = 0$$

and solving for λ, we get:

$$e^{\lambda} = \frac{(1-p)(p+t)}{p(1-p-t)}$$

This expression is greater than or equal to 1 since $p+t \geq p$ and $1-p \geq 1-p-t$, which implies $\lambda \geq 0$, as needed for Markov's inequality! Thus, we can plug this back into our expression and get:

$$\begin{aligned} e^{-\lambda n(p+t)}(1-p+pe^{\lambda})^n &= \left[\left(\frac{(1-p)(p+t)}{p(1-p-t)}\right)^{p+t}\left(1-p+\frac{(1-p)(p+t)}{p(1-p-t)}\right)\right]^n \\ &= \left[\left(\frac{p}{p+t}\right)^{p+t}\left(\frac{1-p}{1-p-t}\right)^{1-p-t}\right]^n \\ &= \exp\left[-n\left((p+t)\log\frac{p+t}{p} + (1-p-t)\log\frac{1-p-t}{1-p}\right)\right] \\ &= e^{-nD((p+t||p)} \end{aligned}$$

and squishing these inequalities, we get:

$$\mathbb{P}(\overline{X}_n \geq p+t) \leq e^{-nD(p+t||p)}$$

proving the first inequality.

For the lower-tail bound, consider $Z_i = 1 - X_i \in [0,1]$. Then $\mathbb{E}X_i = 1-p$. Applying what we've just proved, we get:

$$\mathbb{P}(\overline{Z}_n \geq 1-p+t) \leq e^{-nD(1-p+t||1-p)}$$

since

$$\overline{Z}_n = 1 - \overline{X}_n \geq 1-p+t \quad \Leftrightarrow \quad \overline{X}_n \leq p-t$$

completing the lower tail bound, and hence completing the proof

Example 3.4: Generalization error of Classification Algorithm

A classification algorithm takes in a lot of data, and tries to map it to the appropriate classification. For example, given pictures of cats and dogs to a classification algorithm, the algorithm must correctly output "cat" or "dot". For every "image object", we'll have a tuple (X, Y) with X representing the data (let's say, the pixels of the image) and Y representing the label (in this example, it can either cat or dog). The algorithm will then be a function f that should take in X and give Y: $f(X) = Y$. The way these algorithms are created these days is by training them with a lot of sample data (X_i, Y_i) (for which Y_i is the known value) which the algorithm uses to find patterns to be able to pick up on non-sample data.

The algorithm will try to find a function f that minimizes the propostion of incorrectly guessed output's given the training data:

$$E_n(f) := \frac{1}{n}\sum_{i=1}^{n} I(Y_i \neq f(X_i)) \to \text{minimiaze}$$

Or more generally, instead of using an indicator, a different function called a *loss function* can be used:

$$E_n(f) := \frac{1}{n}\sum_{i=1}^{n} L((Y_i, f(X_i)) \to \text{minimiaze}$$

This average is called the *empirical error* of the classifier f. Different algorithms will give different ways to minimize this error over different classes of functions $\mathcal{F}$ to find a good candidate function $f \in \mathcal{F}$. Now, let's say the algorithm found a function that has a low error estimate on the training data. How do we know it will do a good job outside the training data? If we give it new input not in the training data, will it correctly classify the object?

One way to try and answer this question is to assume that the examples are coming from an [unkonwn] probability distribution $\mathbb{P}$ over a population of possible examples Ω, and that the training data consists of i.i.d. observations (let's say the finite collction (X_i, Y_i) for $1 \leq i \leq n$) from the distribution $\mathbb{P}$. Then the ability to classify future examples (called the *generalisation ability* is measured by the *generalization error*, which we will define as the expected value of the loss function:

$$E(f) := \mathbb{E}(L(Y, f(X))$$

We can ask ourselves if a small value in $E_n(f)$ means there is a small value for $E(f)$? If f is fixed, then as n gets larger, the values will concentrate; in fact this is precisely a use-case of the law of large numbers. However, f is not fixed, it depends on the training data, and so we cannot apply the law of large numbers. To see intuitively why the Law of Large Numbers can't be implied, imagine that the algorithm brute-force memorizes the labels given an X_i (like memorizing the answers to a solution without understanding them). Then even if the amount of solution's memorized increased, it will not necessarily generalize well. In statistics, this is called *over-fitting*

3.4 Azuma's Inequality

In the Hoeffding-Chernoff inequality we dealt with bounded independent variable. We'll now try find a way take more general combination of (not necessarily bounded) random variables (i.e. not just sums of random variables), and still be able to find a bound. We will require to keep the independence and boundedness condition to easily calculate the expected value, In particular, instead of boundedness, we will define the new random variable

$$Z = f(X_1, X_2, ..., X_n)$$

with the following *stability condition*: for some constants $a_1, ..., a_n \in \mathbb{R}$, we have that for any two $x_i, x_i' \in \mathbb{R}$:

$$|f(x_1, x_2, ..., x_i, ..., x_n) - f(x_1, x_2, ..., x_i', ..., x_n)| \leq a_i$$

for all $1 \leq i \leq n$. That is, if we change one coordinate and fix all other coordinates, the value of the function cannot deviate more than a_i (for appropriate i). In the next section, we'll show how this function shows up in some very interesting situation. Notice that by this stability condition, f must be bounded, and so $\mathbb{E}e^{\lambda f} < \infty$

Before we proceed, we first prove a tehcnical lemma to make a bounding argument smoother in the proof of Azuma:

Mention that we will often treat the random X_i as random vectors, as long as:

$$\mathbb{P}(X_1 = x_1, ..., X_n = X_n) = \prod_{i}^{n} \mathbb{P}(X_i = x_i)$$

Lemma 3.4.1: Special Bounding Condition

Let X be a random variable that satisfies $|X| \leq 1$ and $\mathbb{E}X = 0$. Then for any $\lambda \geq 0$

$$\mathbb{E}e^{\lambda X} \leq e^{\lambda^2/2}$$

***Proof* :**
We take advantage of convexity. Notice that we can write λX as:

$$\lambda X = \frac{2X}{2}\lambda = \frac{X+1+X-1}{2}\lambda = \frac{1+X}{2}\lambda + \frac{1-X}{2}(-\lambda)$$

so that $\frac{1-X}{2} + \frac{1+X}{2} = 1$. By the convexity of e^x:

$$e^{\lambda X} \leq \frac{1+X}{2}e^{\lambda} + \frac{1-X}{2}e^{-\lambda}$$

Taking the expected value and using the fact that $\mathbb{E}X = 0$, we get:

$$\mathbb{E}e^{\lambda X} \leq \frac{1}{2}e^{\lambda} = \frac{1}{2}e^{-\lambda} = \cosh(\lambda)$$

and as we've proven in the proof of Hoeffding's inequality, we have that $\cosh(\lambda) \leq e^{\lambda^2/2}$

Theorem 3.4.1: Azuma's Inequality

Let $X_1, ..., X_n$ be independent random variables and let f have the aforementioned stability condition. Then for any $t \geq 0$:

$$\mathbb{P}(|f - \mathbb{E}f| \geq t) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^{n} a_i^2}\right)$$

***Proof* :**
the main idea behind this proof will be a clever representation of $f - \mathbb{E}f$ by breaking it down into a representation of expected values given of "$f(X_1)$", "$f(X_1, X_2)$", and so forth. We start this proof by showing how to define the expected value of these functions:

First, recall Fubini's Thoerem:

$$\mathbb{E}f(X,Y) = \sum_k \left[\sum_\ell f(a_k, b_\ell)\mathbb{P}(Y = b_\ell) \right] \mathbb{P}(X = a_k)$$

In the case we are workign with, we have $f(X_1, ..., X_n)$. We will split this into two random variables in the following way: let's say $A_k = (X_1, X_2, ..., X_k)$ and $a = (x_1, ..., x_k)$ so that $\{A = a\} = \{(X_1 = x_1, ..., X_k = x_k\}$. For notation simplicity, we will write the right hand side as a vector, so $\{(X_1, X_2, ..., X_n) = (x_1, x_2, ..., x_n)\}$ which represents the same thing (think of it as the product space that is has the the product probability from definition 1.5.6). Similarly, we'll split the other random variable's with $B_k = (X_{k+1}, ..., X_n)$ and $b = (x_{k+1}, ..., x_n)$. The random variables (really vectors) A_k and B_k are independent by the grouping lemma, and so

$$\mathbb{P}((X_1, ..., X_n) = (a, b)) = \mathbb{P}((X_1, ..., X_k) = a)\mathbb{P}((X_{k+1}, ..., X_n) = b)$$

and so using Fubini, we get:

$$\mathbb{E}f(X_1, ..., X_n) = \sum_{a_k} \left[\sum_{b_\ell} f(a_k, b_\ell)\mathbb{P}(B_\ell = b_\ell) \right] \mathbb{P}(A = a_k)$$

summing over all possible $a_k = (x_1, ..., x_k)$ and $b_k = (x_{k+1}, ..., x_n)$. To simplify notation a bit, notice that the inside of the square parenthesis is the marginal distribution $\mathbb{P}_{B_\ell}$ with respect to A_k. Notice that what's in the square brakets is the expected value of this marginal distribution: $\mathbb{E}_k f(X_1, ..., X_n) = \sum_{b_\ell} f((X_1, ..., X_k), b_\ell)\mathbb{P}(B_l = b_\ell)$ so that

$$\mathbb{E}f(X_1, ..., X_n) = \sum_k \left[\mathbb{E}_k f(X_1, ..., X_n)]\right] \mathbb{P}(A = a_k)$$

notice that this is simply equal to

$$\mathbb{E}f(X_1, ..., X_n) = \mathbb{E}\left[\mathbb{E}_k f(X_1, ..., X_n)]\right]$$

where the expected value on the outside is with respect to $(X_1, ..., X_k)$ since we have already calculated the (marginal) distribution of the other variables. Notice further that by Fubini's theorem, we can also write:

$$\mathbb{E}_{i-1} f(X_1, ..., X_n) = \mathbb{E}_{i-1}(\mathbb{E}_i f(X_1, ..., X_n))$$

since averaging of $(x_i, ..., x_n)$ is the same as first averaging over $(x_{i+1}, ..., x_n)$, and then averaging over x_i.

With this definition under our belt, we can move onto the next step of this proof. Define:

$$Y_i = \mathbb{E}_i f(X_1, ..., X_n) - \mathbb{E}_{i-1} f(X_1, ..., X_n) \qquad 1 \le i \le n$$

where $\mathbb{E}_0 f(X_1, ..., X_n) = \mathbb{E}f(X_1, ..., X_n)$ and $\mathbb{E}_n = f(X_1, ..., X_n)$. Then using the stability condition of f, we can show that $|Y_i| \le a_i$. To see this, notice that $\mathbb{E}_{i-1}$ differs from $\mathbb{E}_i$ by also averaging over X_i, and those coordinates are fixed in $\mathbb{E}_i$. Since the difference can't be more than a_i, and we are averaging over them, then:

$$|Y_i| \le a_i$$

More particularly, using the fact that the marginal distributions are absolutely convegent to re-arrange the sums:

$$\begin{aligned}
&= |\mathbb{E}_i f(X_1, ..., X_n) - \mathbb{E}_{i-1} f(X_1, ..., X_n)| \\
&= \left| \sum_{b_\ell} f((X_1, ..., X_i), b_\ell)\mathbb{P}((X_{i+1}, ..., X_n) = b_\ell) \right. \\
&\qquad \left. - \sum_{(x_i', b_\ell)} f((X_1, ..., X_{i-1}), x_i', b_\ell)\mathbb{P}((X_i, ..., X_n) = (x_i', b_\ell)) \right| \\
&\overset{!}{=} \left| \sum_{b_\ell} f((X_1, ..., X_i), b_\ell)\mathbb{P}((X_{i+1}, ..., X_n) = b_\ell) \right. \\
&\qquad \left. - \sum_{b_\ell} \left[\sum_{x_i'} f((X_1, ..., X_{i-1}), x_i', b_\ell)\mathbb{P}(X_i = x_i') \right] \mathbb{P}((X_{i+1}, ..., X_n) = b_\ell) \right| \\
&= \left| \sum_{b_\ell} \left[f((X_1, ..., X_i), b_\ell) - \sum_{x_i'} f((X_1, ..., X_{i-1}), x_i', b_\ell)\mathbb{P}(X_i = x_i') \right] \mathbb{P}((X_{i+1}, ..., X_n) = b_\ell) \right| \\
&\leq \sum_{b_\ell} \left| f((X_1, ..., X_i), b_\ell) - \sum_{x_i'} f((X_1, ..., X_{i-1}), x_i', b_\ell)\mathbb{P}(X_i = x_i') \right| \mathbb{P}((X_{i+1}, ..., X_n) = b_\ell) \\
&\overset{!}{\leq} \sum_{b_\ell} |f((X_1, ..., X_i), b_\ell) - f((X_1, ..., X_{i-1}), x_i, b_\ell)| \, \mathbb{P}((X_{i+1}, ..., X_n) = b_\ell) \\
&\overset{!!}{=} \sum_{b_\ell} a_i \mathbb{P}((X_{i+1}, ..., X_n) = b_\ell) \\
&= a_i \underbrace{\sum_{b_\ell} \mathbb{P}((X_{i+1}, ..., X_n) = b_\ell)}_{1} \\
&= a_i
\end{aligned}$$

where $\overset{!}{=}$ comes from Fubini, $\overset{!}{\leq}$ comes from fixing some x_i that increases the sum, and $\overset{!!}{=}$ comes from the stability condition of f. Now, notice that:

$$Y_1 + Y_2 + \cdots + Y_n = f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n)$$

In other words, we just decompseod $f - \mathbb{E}f$ into a sum of marginal distributions! This is called the *martingale-difference representation.* Notice that:

$$\mathbb{E}_{i-1}Y_i = \mathbb{E}_{i-1}(\mathbb{E}_i f(X_1, ..., X_n) - E_{i-1} f(X_1, ..., X_n)) = \mathbb{E}_{i-1} f(X_1, ..., X_n) - E_{i-1} f(X_1, ..., X_n) = 0$$

so the average of Y_i with respect to the last coordinate X_i. We are now setup to use Markov's inequality:

$$\mathbb{P}(f - \mathbb{E}f \geq t) = \mathbb{P}(\sum_i^n Y_i \geq t) \leq \frac{\mathbb{E}e^{\lambda Y_1 + \cdots + \lambda Y_n}}{e^{\lambda t}}$$

To compute it further, using the $\mathbb{E}_i f = \mathbb{E}_i(\mathbb{E}_{i-1} f)$ equation with $i = n-1$, we get

$$\begin{aligned}\mathbb{E}e^{\lambda Y_1 + \cdots + \lambda Y_n} &= \mathbb{E}[\mathbb{E}_{n-1} e^{\lambda Y_1 + \cdots + \lambda Y_n}] \\ &= \mathbb{E}[e^{\lambda Y_1 + \cdots + \lambda Y_{n-1}} \mathbb{E}_{n-1} e^{\lambda Y_n}]\end{aligned}$$

since $\mathbb{E}_{n-1}$ is the average only in terms of X_n, and the terms $Y_1, ..., Y_{n-1}$ do not depend on X_n. Take $X = Y_n/a_n$. This function is only dependend on X_n, and has $|X| \leq 1$ and $\mathbb{E}_{n-1}X = 0$, and so by lemma 3.4.1 we get:

$$\mathbb{E}_{n-1} e^{\lambda Y_n} = \mathbb{E}_{n-1} e^{\lambda a_n X} \leq e^{()\lambda a_n)^2/2}$$

substituing back into our previous equation we get:

$$\mathbb{E}e^{\lambda Y_1 + \cdots + \lambda Y_n} \leq e^{(\lambda a_n)^2/2} \mathbb{E}e^{\lambda Y_1 + \cdots + \lambda Y_{n-1}}$$

from here, we can proceed with the same techniqu for X_{n-1}, ..., X_1 to get that:

$$\mathbb{E}e^{\lambda Y_1 + \cdots + \lambda Y_n} \leq e^{\sum_i^n (\lambda a_i)^2/2}$$

and so:

$$\mathbb{P}(\sum_i^n Y_i \geq t) \leq \exp\left(-\lambda t + \frac{\lambda^2}{2} \sum_i^n \lambda a_i^2\right)$$

and optimizing over λ get's us:

$$\mathbb{P}(|f - \mathbb{E}f| \geq t) \leq \exp\left(\frac{-t^2}{2\sum_{i=1}^n a_i^2}\right)$$

as we sought to show

3.5 Application's of Azuma's Inequality

In this section, we will go over 4 examples of Azuma's inequality.

3.5.1 Chromatic Number of Erdós Rényi graph

Let $G(n,p)$ be the Erdós Rényi graph on n vertices. Now, define $\chi(G(n,p))$ to be the smallest number of colours needed to colour the vertices so that no two adjacent vertices share the same colour. The function $\chi(G(n,p))$ is called the *chromatic number* of $G(n,p)$. To represent this problem, we must keep track of the edges. Let

$$e_{i,j} = I(\text{there is an edge between vertex } v_i \text{ and } v_j)$$

Then by definition, each $e_{i,j} \sim B(p)$ and are all independent of each other. Next, let $X_i = (e_{1,i}, e_{2,i}, ..., e_{i-1,i})$ for $2 \leq i \leq n$ be random vectors consisting of all possible edges of the random graph. Therefore, taking as input all possible edges and giving a chromatic number is equivalent as taking as input the random variables $X_2, ..., X_n$ and giving an output, that is for appropriate f:

$$\chi(G(n,p)) = f(X_2, ..., X_n)$$

We want to apply Azuma, and so we must verify that f satisfie's the stability condition. Let's say that we have changed some X_i to X_i' with different values (i.e. we modify some edges between v_i all vertices with a lower index). Since we can always assign new colours, the chromatic number cannot increase by more than 1, and so

$$f(X_2, ..., X_i, ..., X_n) - f(X_2, ..., X_i', ..., X_n) \leq 1$$

and by the same reasoning:

$$f(X_2, ..., X_i', ..., X_n) - f(X_2, ..., X_i, ..., X_n) \leq 1$$

and so

$$|f(X_2, ..., X_i', ..., X_n) - f(X_2, ..., X_i, ..., X_n)| \leq 1$$

Giving us the condition of Azuma's inequality! Thus, aplying it both to f and $-f$, we get:

$$\mathbb{P}(|\chi(G(n,p)) - \mathbb{E}\chi(G(n,p))| \geq t) \leq 2e^{\frac{-t^2}{2(n-1)}}$$

To make the inequality more enlightening, if we take $t = \sqrt{2n \log n}$ and consider the compliment probability, we get:

$$\mathbb{P}(|\chi(G(n,p)) - \mathbb{E}\chi(G(n,p))| \leq \sqrt{2n \log n}) \geq 1 - \frac{2}{n}$$

Since n can be really large, with high probability the chromatic number will be close to the expected value. Part of the beauty of Azuma's inequality is that we don't need to know what the expected value is to get our result! It is a known (but difficult to prove) fact that:

$$\mathbb{E}\chi(G(n,p)) \sim \frac{n}{2 \log n} \log \frac{1}{1-p}$$

meaning the deviation $\sqrt{2n \log n}$ is of a smaller order than that of the expected value, meaning the chromatic numbers are usually close to the expected value (as n gets large). This shows an example of the law of large numbers for a very non-trivial random variable!

3.5.2 Balls and boxes

Let's say you have m boxes and n balls. You through the balls at random, with each box having equal chance of having a ball land in it (i.e. $1/m$). Each ball through is independent of each other. Let N be the random variable representing the number of non-empty boxes. To use Azuma's, we wan to characterize N in terms of a function of random variables. To that end, let X_i $(1 \leq i \leq n)$ represent the number of the box for which the ith ball lands (i.e. $X_i \in \{1, ..., m\}$). Then

$$N = |\{X_1, ..., X_n\}| = \text{card}\{X_1, .., X_n\}$$

is the number of distinct boxes hit, meaning we have represented N as a function of $X_1, ..., X_n$. To check the stability condition, let's say we replace X_i with X_i'. Then at most the size of the set will change is 1, and so the stability condition holds, letting us use Azuma's inequality to find:

$$\mathbb{P}(|N - \mathbb{E}N| \geq t) \leq 2e^{-\frac{t^2}{2n}}$$

To make the right hand side nicer, repace $t = \sqrt{2n \log n}$ and take the comliment to get:

$$\mathbb{P}(|N - \mathbb{E}N| \leq \sqrt{2n \log n}) \geq 1 - \frac{2}{n}$$

Recall from section 1.7.1 that

$$\mathbb{E}N = m\left(1 - \left(1 - \frac{1}{m}\right)^n\right)$$

If we let n get very large with respect to m, then we can write $m = \alpha n$, with $\alpha > 0$, then we get that this value is asymptotically equal to:

$$\mathbb{E}N \sim n\alpha(1 - e^{-1/\alpha}) \tag{3.1}$$

Thus, we see that in general that the typical number of non-empty boxes is close to the expected value since the $|\cdot|$ term grows much faster than the term on the other side, and since $1 - 2/n \to 1$ as $n \to \infty$, we see it should be very closed to the expected value.

3.5.3 Max-cut Sparse Graph

This example will be like the last one except we will have a more complicated function and re-interpret the boxes and valls in terms of a random graph.

Let $G(n,p)$ be a random graph where $V = \{v_1, ..., v_n\}$ and E are the vertices and edges respectively. We want to find a way to cut the graph in two in such a way that a maximal number of edges connect the two components: this is called the *max-cut problem*. This comes in often in comptuer science when trying to figure out the layout of electronic circutry, or for a way to reformulate certain combinatorial problems. For example, consider the set of vertices to be people and edges represent if the two people dislike each other. Then we would want to as much as possible seperate the group of people so as many "enemies" are on opposite sides.

To find a random variable that captures this idea, we first construct the random graph in a somewhat different way than we have done before. Start off by choosing some natural number $d > 0$ (the variable d is supposed to hint that it will represent in some way the "degree" of the vertices), and let $m = dn$ (where as usuall we will think of n as being very large). The number m represents the number of edges we want to put in our graph. We now distribute them randomly and independently into the $\binom{n}{2}$ possible places they can go into our random graph (if an edge goes to the same place twice, we only "keep" on of them). Notice how this is similar to the balls and boxes scenario from last section: a possible location for an edges is a box, and the edge itself is the ball. So we can define the same random variables $X_1, ..., X_m$ from the last section that can take on values $1 \leq X_i \leq \binom{n}{2}$. Since $m < \binom{n}{2}$, this model produces a *sparse graph* since there are relatively few edges compared to all possible edges, and there is a fixed number of edges per vertices (namely d). Just like before, we define

$$e_{i,j} = I(\text{there is an edge between } v_i \text{ and } v_j)$$

sinces edges are not repeated between vertices, this random variable captures all edges between two vertices.

We will now define the function *Max-Cut* as follows: let each vertex be assigned the a value from $\{-1, 1\}$ randomly. We consider all vertice swith the same value to be part of the same group. Let

$$\sigma = (\sigma_1, ..., \sigma_n) \in \{-1, 1\}^n$$

Then a σ represents a possible cut of the graph into 2 groups. Let $E(\sigma)$ represent the number of edges connecting the two groups:

$$E(\sigma) = \text{card}\,\{i < j \mid e_{i.j} = 1,\ \sigma_i \neq \sigma_j\}$$

This random variable is a bit hard to work with, and so here is another way to represent the same value: notice that $\sigma_i \neq \sigma_j$ if and only if $\sigma_i\sigma_j = -1$ (otherwise $\sigma_i\sigma_j = 1$). Thus, we can write:

$$I(\sigma_i \neq \sigma_j) = \frac{1 - \sigma_i\sigma_j}{2}$$

and so:

$$E(\sigma) = \frac{1}{2}\sum_{i<j} e_{i,j}(1 - \sigma_i\sigma_j)$$

Finally, we define M to represent the maximal number of $E(\sigma)$:

$$M = \max_{\sigma} E(\sigma)$$

We'll show that M respects the stability condition. Remember that M is ultimately a function of $X_1, ..., X_n$, and so we can check the stability conditions of those variables. let's say we change on X_i to X_i', that is we move one edge to a different possition. Then M can change by at most 1 since it will effect $E(\sigma)$ by at most 1 value. Thus, by Azuma's inequality:

$$\mathbb{P}(|M - \mathbb{E}M| \geq t) \leq e^{\frac{-t^2}{2m}} = e^{\frac{-t^2}{2dn}}$$

and as usual we can replace $t = \sqrt{2dn \log dn}$ and take the compliment to get:

$$\mathbb{P}(|M - \mathbb{E}M| \leq \sqrt{2dn \log dn}) \geq 1 - \frac{2}{n}$$

We would like to now know if the deviation $\sqrt{2dn \log dn}$ has order greater than or less than $\mathbb{E}M$, since that will determine whether the the typical max-cut is close to the expected value. In fact, this turns out to be the case: there exists a cut such that at least half of all edges are between the vertices that belong to opposite groups. Notice that $\sigma_1, ...\sigma_n$ are Rademacher random variables. Then:

$$\begin{aligned} M &= \max_{\sigma} E(\sigma) \\ &\overset{!}{\geq} \mathbb{E}E(\sigma) \\ &= \frac{1}{2}\sum_{i<j} e_{i,j}(1 - \sigma_i\sigma_j) \\ &= \frac{1}{2}\sum_{i<j} e_{i,j} \end{aligned}$$

where $\overset{!}{=}$ comes from how the expected value would be less than the maximum by the nature of the expected value. Notice that the last term rerpesents half of all edges! Thus, taking the expected value of both sides, we get:

$$\mathbb{E}M \geq \frac{1}{2}\sum_{i<j} \mathbb{P}(\text{edge between } v_i,\, v_j) = \frac{1}{2}\binom{n}{2}\left(1 - \left(1 - \frac{1}{\binom{n}{2}}\right)^{dn}\right)$$

where we get the last equality for the same reason we got as in equation (3.1). To simplify this lower bound of $\mathbb{E}M$, notice that by convexity that $1 - x \leq e^{-x}$ and so:

(here I don't follow, it feels like the inequalities are going the wrong way...)

3.5.4 Hamming Cube

A 3d-cube has vertices that can be represented by taking a path $\{0,1\}^3$ where 0 means don't move, and the 1 means go in direction x, y, z, dependent (i.e. each vertex is represented by a vector). More generally, $\{0,1\}^n$ represents an n-cube. For any $x, y \in \{0,1\}^n$, the number of coordinates that they differ is called the *hamming distance* and is represented by:

$$\rho(x,y) = \sum_{i=1}^{n} I(x_i \neq y_i)$$

If you know about the taxi-cab metric or path metric, this is precisely what this is. We'll use Azuma's inequality to show that on average, for any $A \subseteq \{0,1\}^n$, there is a path from one point A to any point of $\{0,1\}^n$ that requires at most $\sqrt{n}$ points. In particular, for some $\varepsilon > 0$, and $A \subseteq \{0,1\}^n$ such that A $> \varepsilon 2^{\varepsilon}$, the haming distance will be in the order of $\sqrt{n}$.

Let's first define balls in this metric (i.e. the set of all points whic htake at most t steps away from A)

$$B_t(A) = \{x \in \{0,1\}^n \mid \rho(x,y) < t \text{ for some } y \in A\}$$

With this, we establish the result in the following lemma (a lemma since I think we'l link this to something later, in that case it is at ref:HERE). For notational simlicity, for any $\varepsilon, \delta \in (0,1)$, define:

$$t_{\varepsilon,\delta} = \sqrt{2\log\frac{1}{3}} + \sqrt{2\log\frac{1}{\delta}}$$

Lemma 3.5.1: Hamming Lemma

For any $\varepsilon, \delta \in (0,1)$, if $\operatorname{card}(A) > \varepsilon 2^n$, then

$$\operatorname{card} B(A, t_{\varepsilon,\delta}\sqrt{n}) \geq (1-\delta)2^n$$

That is, the $t_{\delta,\varepsilon}\sqrt{n}$-neigbhorhood contains at least $(1-\delta)$ propostion of all points $\{0,1\}^n$

***Proof* :**
Let $X = (X_1, X_2, ..., X_n)$ consist of i.i.d. Bernoulli random variables, so that for any subset $S \subseteq \{0,1\}^n$,

$$\mathbb{P}(X \in S) = \frac{\text{S}}{2^n}$$

Now define a new random variable Z which is equal to the distance from X to the set A:

$Z = \min_{y \in A} \rho(X, y)$. The advantage of representing the point in the 1st component of the previous equation as X is because X is a vector of i.i.d. random variables, and so notice that if we change one coordiante X_i to X_i', then Z changes by at most 1, and so Azuma's inequality (on Z and $-Z$) aplies, and we get (with some change of variables)

$$\mathbb{P}(Z \leq \mathbb{E}Z - t\sqrt{n}) \leq e^{\frac{-t^2}{2}}$$

$$\mathbb{P}(Z \geq \mathbb{E}Z + t\sqrt{n}) \leq e^{\frac{-t^2}{2}}$$

We take both these cases seperately to find different information: If we take $t_\varepsilon = \sqrt{2 \log \frac{1}{\varepsilon}}$ then

$$\mathbb{P}(Z \leq \mathbb{E}Z - t_\varepsilon\sqrt{n}) \leq \varepsilon$$

At this point, we will do a new trick that is specific to this problem: we can conlude that $\mathbb{E}Z \leq t_\varepsilon\sqrt{n}$: If $\mathbb{E}Z - t_\varepsilon\sqrt{n} > 0$, then

$$\varepsilon = \mathbb{P}(Z \leq \mathbb{E}Z - t_\varepsilon\sqrt{n}) = \mathbb{P}(Z \leq s) \geq \mathbb{P}(Z = 0)$$

where $s > 0$ is some positive number. However, $Z = 0$ if and only if $X \in A$, and so:

$$\varepsilon \geq \mathbb{P}(Z = 0) = \mathbb{P}(X \in A) = \frac{\text{card}(A)}{2^n} \overset{!}{>} \varepsilon$$

where $\overset{!}{>}$ comes by assumption, which implies $\varepsilon > \varepsilon$, a contardiction. Thus:

$$\mathbb{E}Z \leq t_\varepsilon\sqrt{n}$$

Now using the second inequlaity, we see that:

$$\mathbb{P}(Z \geq t_\varepsilon\sqrt{n} + t\sqrt{n}) \leq \mathbb{P}(Z \geq \mathbb{E}Z + t\sqrt{n}) \leq e^{-t^2/2}$$

If we take $t_\delta = \sqrt{2 \log \frac{1}{\delta}}$ then:

$$\mathbb{P}(Z \geq t_\varepsilon Z + t_\delta\sqrt{n}) \leq \delta$$

and taking the compliment gives us:

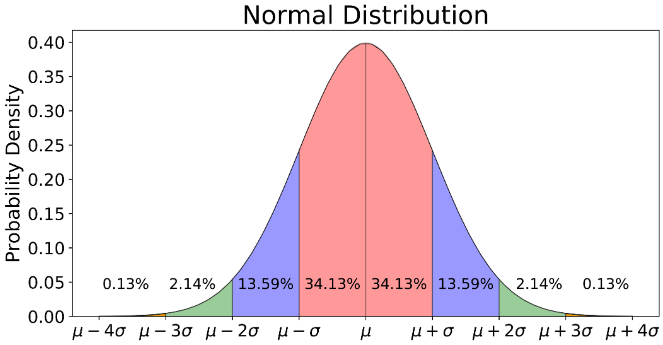
$$\mathbb{P}(Z \leq t_\varepsilon Z + t_\delta\sqrt{n}) \leq 1 - \delta$$

Since this set is exactly the set $B(A, t_{\varepsilon,\delta}\sqrt{n})$ of all points with hamming distance $t_{\varepsilon,\delta}\sqrt{n}$ from A, and the probability is greater than 0, we have the desired result.

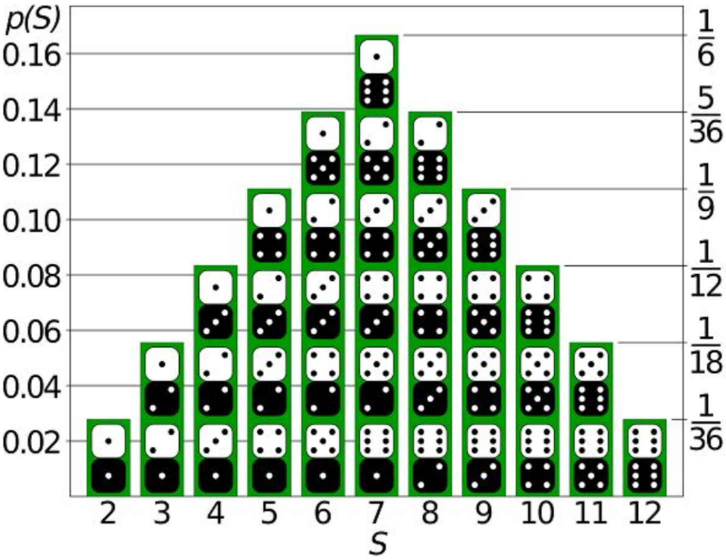
4

Gaussian Distribution

Many might have heard of the normal distribution: many metrics (such as I.Q., weight of people, height, SAT scores, blood pressure, error's in measurements produced by a machine) turn out to fall into a normal distribution, with the chances of being within 1 standard deviation away to be 68%, being 2 standard deviations away to be 95%, and 3 standard deviation to be 99%:



All of the given examples were required a large pool of data to show, but we can do it with a pair of dice: the values are normally distributed:



Since we can relate it to dice, we can relate it to a fair coin (take 12 i.i.d. $B(1/2)$ random variables). In fact, we almost encountered this distribution in example 3.2 (we will repalce t with x because t will be used for a different purpose in a moment): recall that

$$\overline{X}_n = \frac{S_n}{n} = \frac{X_1, X_2, ..., X_n}{n}$$

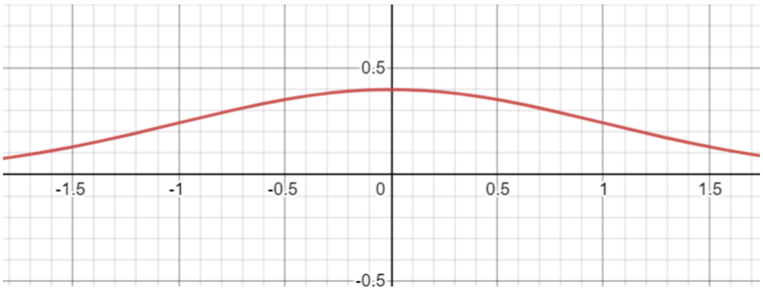
with $X_1, X_2, ..., X_n$ being i.i.d. Bernoulli random variable $B(1/2)$. Given $x \geq 0$, we proved in example 3.2 that:

$$\mathbb{P}(|\overline{X}_n - 1/2| \geq x) \leq 2e^{-2nx^2}$$

doing a change of variables $x = \frac{t}{2\sqrt{n}}$ and defining $Z_n := 2\sqrt{n}(\overline{X}_n - 1/2)$, we get:

$$\mathbb{P}(|Z_n| \geq t) \leq 2e^{\frac{-t^2}{2}} \tag{4.1}$$

If we graph this funciton, it will be really close to the normal distribution we discussed earlier. In fact, if we replace the constant 2 with $\sqrt{1}\sqrt{2\pi}$, then the function $\frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}$ will be:



Taking the integral of this function on $[-a, a]$ for $a \in \{1, 2, 3\}$, and letting $f(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-t^2}{2}}$, we get:

$$\int_{-1}^{1} f(x) \approx 0.6826 \qquad \int_{-2}^{2} f(x)dx \approx 0.9545 \qquad \int_{-3}^{3} f(x)dx \approx 0.9973$$

showing that we got back the normal distribution! Even better, notice that

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

(we will prove this soon), and if $A, B \subseteq \mathbb{R}$ are disjonit, then:

$$\int_{A \sqcup B} f(x)dx = \int_A f(x)dx + \int_B f(x)dx$$

meaning the function $\mathbb{P}(E) = \int_E f(x)dx$ satisfies the definition of a probability! How then do we get $f(x)$, in particular the constant $\frac{1}{\sqrt{2\pi}}$ from equation 4.1? That will be the contents of the *Central Limit Theorem*, a key result of this section. What the CLT will tell us is given the sum of independent random variable actually converge towards this distribution!

In particular, we will get that:

$$\lim_{t\to\infty} \mathbb{P}(Z_n \leq t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{-x^2}{2}} dx$$

The name central limit theorem might not seem clear. Under a different perspective, it might be more clear: recall that $\mu = \mathbb{E}X_1 = \frac{1}{2}$, $\sigma^2 = \text{Var}(X_1) = \frac{1}{4}$, so we can re-write Z_n as :

$$Z_n = \frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu) = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$$

From this perspective, the aforementioned limit becomes:

$$\lim_{t\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \leq t\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{t} e^{\frac{-x^2}{2}} dx$$

which looks a lot more like the law of large numbers modified in some way! This hopefully gives some motivation for the coming study of the Gaussian distribution

4.1 Gaussian Distribution on $\mathbb{R}$

We'll first put a box around the gaussian distribution for ease of reference:

Definition 4.1.1: Gaussian Distribution and Measure

Let $p(x) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$. Then this function is called the *standard Gaussian distribution.* Define $\gamma : P(\mathbb{R}) \to \mathbb{R}$ to be:

$$\gamma(A) = \int_A p(x)dx$$

Then γ is called the *standard gaussian measure on* $\mathbb{R}$. The continuous probability space with gaussian probability meausre is usually denoted $(\mathbb{R}, \gamma)$.

A quick word before we continue: the domain of γ was said to be $P(\mathbb{R})$, however this is not strictly speaking true. As would be covered in a course on measure theory, this set is too large (there exists sets that will contardict γ being a probability[1]). Therefore, we would usually limit ourselves to "measurable set" and give a new symbol for the domain (for example $\mathcal{M}(\mathbb{R})$) to show the restriction from $P(\mathbb{R})$ to the set of measurable sets[2]. In this book, we will not worry about that; instead we will only work with countable subsets (since the integral over a countabler subset is always 0), and sets of the form:

$$(a, b), (a, b], [a, b), [a, b]$$

where $a, b \in \overline{\mathbb{R}}$ (i.e. $\mathbb{R} \cup \{\pm\infty\}$), and all finite unions of such intervals[3]. Notice that since the integral does not care if we drop a finite number of points, then it doesn't really matter if we deal with (a, b), $(a, b]$, $[a, b)$, or $[a, b]$ (if $-\infty < a, b < \infty$, since we don't define the integral on infinity by considering infinity to be a point, but as the asymptotic behavior of $\lim_{b\to\infty} \int_a^b f(x)dx$)

Notice that we have just defined our first *continuous distribution*! Unlike discrete distributions, we define the probability of events over the events of the from described above, and not for each individual outcome. That means we will have to go over how to define change of variables, expected value, random variable, and so forth. We will do all of this in a moment, but let us first verify that γ does ineed satisfy the conditions of a probability. In particular, we should show that $\gamma(\mathbb{R}) = 1$ (since $\gamma(A \sqcup B) = \gamma(A) + \gamma(B)$ comes for free from the linearity of the domain of the integral). To compute this, we use a familiar trick from multivariable calculus:

$$\begin{aligned}
\gamma(\mathbb{R})^2 &= \left(\frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-x^2/2}\right) \\
&= \frac{1}{\sqrt{2\pi}} \iint\limits_{\mathbb{R}^2} e^{-(x^2+y^2)/2} dxdy \\
&= \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} dr d\theta \\
&= \frac{1}{2\pi} \int_0^{2\pi} d\theta \int_0^{\infty} e^{-r^2/2} dr \\
&= 1
\end{aligned}$$

showing that γ is indeed a probability space! This means that we can use the $\mathbb{P}$ to represent it as well, however we will often stick to γ to make it more explicit that we are dealing with a continuous probability space. As an exercise, all the properties covered in section 1.2. One important comment to make is proving countable additivity of disjonit sets. This is not always possible with if we use the Riemann integral definition (prove that χ_q for $q \in \mathbb{Q}$ is Riemann integrable, but that $\chi_{\mathbb{Q}}$ is not Riemann integrable, in particular it is continuous nowhere).

Let us now ponder over how random variables would look like over this distribution. The simplset such random variable would be the identity function $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x$. It is easy to see that for any event $A \subseteq \mathbb{R}$ that:

$$\mathbb{P}(g \in A) = \gamma(\{x \mid g(x) \in A\}) = \gamma(A)$$

[1] Google Vitalis sets to see the construction of such a set

[2] In fact, there is not real "natural" restriction to some subsets $\mathcal{M}(\mathbb{R})$, meaning we have to make some choice of what restriction we make. This leads to measure-theorists needing to keep track of what choice of $\mathcal{M}(\mathbb{R})$ was made, and so instead of writing $(\mathbb{R}, \gamma)$ we would write $(\mathbb{R}, \mathcal{M}, \gamma)$ to keep track of the choice

[3] If we wanted to use more machinery from measure theory, we would employ the Lebesgue intergral and allow countable unions, however this is not necessary for this book

A little more generaly, for any event A, we can define χ_A to be a random variable (i.e. the indicator function on the event, i.e. finite union and intersections of intervals, A). Another example would be $g(x) = -x$ (you should prove this as an exercise). However, to define random variable's more generally is not as easy as it was in the discrete case. In particular, if we let $X : \mathbb{R} \to \mathbb{R}$ a meaningful random variable, then for any event $A \subseteq \mathbb{R}$, we should be able to calculate $\mathbb{P}(X \in A)$, so we must do:

$$\begin{aligned}\mathbb{P}(X \in A) &= \gamma(\{x \mid X(x) \in A)\} \\ &= \gamma(X^{-1}(A)) \\ &= \int_{X^{-1}(A)} p(x)dx\end{aligned}$$

which means we must be able to integrate over $X^{-1}(A)$. However, just like not every set is "measurable", not every function will produce sets $X^{-1}(A)$ that are "measurable" (in our case, that will be finite unions and intersection of intervals). Function's that will always produce such sets are called "measurable functions". The natural environment to define such a function is, again, measure theory, but we want to avoid the technicalities of measure theory to make the study of probability in the continuous case accessible, so instead we will limit ourslves to a smaller class of measurable functions that will suffice for this book. It must be noted that this isn't too big of a restriction when it comes to the types of random variables that are used in probability, so if you are not aware of the technicalities of measurable functions you can just think of this next definition as trying to avoid the "pathologies" that would come up when dealing with measure theory that are not interesting from a probability point of view:

Definition 4.1.2: Nice Function

We will say a function $f : \mathbb{R} \to \mathbb{R}$ is *nice* if:

1. The function is piece-wise continuous
2. The funtion is bounded on compact sets
3. The pre-image of any interval A (i.e. $f^{-1}(A)$) is a Rieman-integrable set

In most cases, we will usually deal with functions where $f^{-1}(A)$ is the finite union of intervals.

Example 4.1: Nice Functions

1. Let:
$$X(x) = I(x \geq 0) - I(x < 0)$$
Then the image of X is $\{1, -1\}$. Notice that:
$$\mathbb{P}(X = 1) = \gamma(\{x \mid x \geq 0\}) = \int_0^\infty p(x)dx = 1/2$$
$$\mathbb{P}(X = -1) = \gamma(\{x \mid x < 0\}) = \int_{-\infty}^0 p(x)dx = 1/2$$
which shows that X is in fact a *Rademacher random variable* within a continuous probability space $(\mathbb{R}, \gamma)$.
2. Let
$$X(x) = |x|$$

The image of X is $[0,\infty)$, and so as a start let's consider events of the form $[a,b]$:

$$\mathbb{P}(X \in [a,b]) = \gamma(\{x \mid X(x) \in [a,b]\}) = [-b,-a] \cup [a,b]$$

Since the Gaussian density $p(x)$ is symmetric, we get that:

$$\mathbb{P}(X \in [a,b]) = 2\gamma([a,b]) = 2\int_a^b p(x)dx$$

given this new insight on how this random variable gets "converted" into some function of γ (namely 2γ) and that it's never negative, define:

$$p_X(x) := 2p(x)I(x \geq 0)$$

Then for any inteval $[a,b]$ we can write:

$$\mathbb{P}(a \leq x \leq b) = \int_a^b p_X(x)dx$$

and in general, if A is the finite disjoint union of intervals, then:

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

Notice that $\mathbb{P}$ in the previous example is now defined in terms of a new density (the function inside the integral)! Additivity clearly works, and for $\mathbb{P}(\mathbb{R}) = 1$, we still get:

$$\int_{\mathbb{R}} p_X(x)dx = 1$$

meaning $p_X(x)$ *also* defines a probability measure! In fact, we can construct many new probability measures through the use of such random variables:

Definition 4.1.3: Continuous Distributions From Densities

Let $(\mathbb{R},\gamma)$ be the standard gaussian probability space and let X be a random variable. Then if there exists a function p_X such that

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

where $\mathbb{P}(X) = 1$, then p is called the *density* with respect to X

Note that not all random variables and and distributions can be expressed in this way: there does not always exist a function $p_X(x)$ such that $\mathbb{P}(X \in A) = \int_A p_X(x)dx$.In this class, all continuous random variables will have this property. For those taking measure theory, you will encounter some continuous random variables that don't admit distributions. More genreally, distributions that admit a density are called *absolutely continuous*.

Example 4.2: Continuous Distributions From Densities

1. Define the density
$$p_X(x) = \lambda e^{-\lambda} I(x \geq 0)$$
then the resutling probability is called the *exponential distribution with paramater* λ

2. Define the density
$$p_X(x) = \frac{1}{b-a} I(a \leq x \leq b)$$
then the resutling probability is called the *uniform distribution on the interval* $[a, b]$

An equivalent to the tailsum formula for the continuous case is the following:

Definition 4.1.4: Cumulative Ditribution Function (c.d.f.)

Let $(\mathbb{R}, \mathbb{P})$ be a probability space. Then $F(x) = \mathbb{P}(X \leq x)$ is called the *cumulative distribution function* (or c.d.f for short).

The nice thing about the c.d.f. is that it gives you an easy way to calculate the probability of an event over an interval, since:
$$\mathbb{P}(a \leq x \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F(b) - F(a)$$

We get some more properties if F has nicer properties. For examples, if F is continuous, then:
$$\mathbb{P}(X = a) \leq \mathbb{P}(a - \varepsilon \leq x \leq a + \varepsilon) = F(a + \varepsilon) - F(a - \varepsilon)$$

Letting $\varepsilon \to 0$, we see that $\mathbb{P}(X = a) = 0$! It might seem like an obvious statement that this is the case since we are integrating over a single point, however this consideration becomes more important in more general environments, but it is interesting to see it here as a property of continuity of F.

Furthermore, if F is differentiable, then the deritive $F'(x) = P_X(x)$ is piece-wise continuous (by assumption of nice function), so by the Fundamental Theorem of Calculus:
$$\mathbb{P}(a \leq X \leq b) = F(b) - F(a) = \int_a^b p_X(x)dx$$

Hence, $p_X(x)$ is the density of the distribution of X. Using all this information, we can deduce what a density ought to be given a random variable:

Example 4.3: Deducing Density From X

1. Let $X = g^2$ where $g(x) = x$ is the standard Gaussian Random variable on $(\mathbb{R}, \gamma)$. Then for $x > 0$ (since $x < 0$ is not in the range, and we will cover the $x = 0$ case soon):
$$F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(g^2 \leq x)$$
re-arranging so that we may apply the fundamental theorem of calculus and take advantage of the fact that g is the standard gaussian:
$$\mathbb{P}(-\sqrt{x} \leq g \leq \sqrt{x}) = \int_{-\sqrt{x}}^{\sqrt{x}} p(t)\, dx$$

Taking the derivative with respect to x and taking advantage of the symmetry of the Guassian, we get:

$$F'(x) = \frac{1}{\sqrt{x}}p(\sqrt{x}) - p(-\sqrt{x})\frac{-1}{\sqrt{x}} = p(\sqrt{x})\frac{1}{\sqrt{x}} = \frac{1}{\sqrt{2\pi x}}e^{-x/2}$$

Since X takes on non-negative values, we define the density to be:

$$p_X(x) = \begin{cases} 0 & x \leq 0 \\ \frac{1}{\sqrt{2\pi x}}e^{-x/2} & x > 0 \end{cases}$$

Thus, $p_X(x)$ is the density of distribution for X. This distribution in fact has a name: $\Gamma\left(\frac{1}{2}, \frac{1}{2}\right)$; what this notation means will be addressed in section ref:HERE

2. Let $\mu \in \mathbb{R}$, $\sigma \in \mathbb{R}_{>0}$, and $X = \mu + \sigma g$ where $g(x) = x$ is the standard gaussian distribution. Then, doing the same calculations as in the last example, we get:

$$p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

This distribution is usually denoted $N(\mu, \sigma)$ and is called the *Gaussian distribution with mean μ and variance σ^2* (the name will become clear shortly). For convienence, when $\mu = 0$, we usually denote $p_\sigma(x) := p_{0,\sigma}(x)$

The next concept to update is that of the expected value:

Definition 4.1.5: Expected Value Over Gaussian

Let $(\mathbb{R}, \gamma)$ be the probability space. The expected value of $X : \mathbb{R} \to \mathbb{R}$ on $(\mathbb{R}, \gamma)$ is:

$$\mathbb{E}X = \int_{\mathbb{R}} X(x)p(x)dx$$

Assuming $\mathbb{E}|X| < \infty$.

Example 4.4: Expected Values Over Gaussians

1. Take $g(x) = x$ to be the random variable. Then:

$$\mathbb{E}g = \int_{\mathbb{R}} xp(x) = 0$$

Since $p(x)$ is symmetic, and combined with x makes $xp(x)$ an odd function.

2. Let $g(x) = x^2$. Then:

$$\begin{aligned}\mathbb{E}g^2 &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} x^2 e^{-x^2/2}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} xd(e^{-x^2/2}) \\ &= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-x/2} \\ &= 1\end{aligned}$$

by doing integratin by parts. Using this piece of information, we see that $\text{Var}(g) = 1$

3. We can combine these two facts to find the expected value and variance of $X = \mu + \sigma g \cong N(\mu, \sigma^2)$. In particular, we that:

$$\mathbb{E}X = \mu \qquad \text{Var}(X) = \mathbb{E}(\sigma g^2) = \sigma^2$$

justifying the names of the parameters of $N(\mu, \sigma^2)$. Note that the usual Gaussian would thus be $N(0, 1)$ sine by the result of the last section: $\text{Var}(g) = \mathbb{E}(g^2) - (\mathbb{E}g)^2 = 1 - 0 = 0$.

We next bring in the notion of change of variables with respect to expected value for continuous functions. As a reminder we saw when a distribution is continuous that:

$$\mathbb{P}(X \in A) = \int_{X^{-1}(A)} p(x)dx = \int_A p_X(t)dt$$

for some density $p_X(t)$.

Lemma 4.1.1: Change Of Variable For Integrals

Let X be a random variable such that $\mathbb{E}|X| < \infty$. Then:

$$\mathbb{E}X = \int_{\mathbb{R}} X(x)p(x)dx = \int_{\mathbb{R}} tp_X(t)dt$$

***Proof* :**
This proof is very long so I will do it later

Like in we saw in the discrete case in section **??**, we can quite easily represent $\mathbb{E}f(x)$. The same holds true for the continuous case:

Lemma 4.1.2: Change Of Variables For Integrals II

Let X be a random variable such that $\mathbb{E}|f(X)| < \infty$ for some [nice?] function f. Then:

$$\mathbb{E}f(X) = \int_{\mathbb{R}} f(t)p_X(t)dt$$

***Proof* :**
The proof is similar to the last according to the book with some minor modifications. I will fill this in later.

A closing word before continuing to the n-dimensional case: we can also generalize Chebyshev's and Markov's inequality to the continuous case. If X is a random variable with associated density $p_X(x)$, then as an exercise show that:

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

the proof essenitally mirrors the one for the discrete case.

4.2 Gaussian on $\mathbb{R}^n$

We will now define a probability on $\mathbb{R}^n$. First, recall (or define) the length of a vector $x \in \mathbb{R}^n$ by:

$$|x| = (x_1^2 + \cdots + x_n^2)^{1/2}$$

Since $|x| \in \mathbb{R}$, this will let us define the following

Definition 4.2.1: Guassisan Distribution On $\mathbb{R}^n$

Take $\mathbb{R}^n$ be be a the standard euclidean space. Then define $\mathbb{P}$ where:

$$\begin{gathered}\mathbb{P}(X \in A) \\ = \gamma_n(A) \\ = \int_A p_n(\vec{x})d\vec{x} \\ = \int_A p_n(x_1, ..., x_n)dx_1...dx_n \\ = \int_A \frac{1}{(2\pi)^{n/2}} e^{-|x|^2/2} dx_1 dx_2 \cdots x_n\end{gathered}$$

where $p_n(x)$ is called the *standard gaussian distribution on* $\mathbb{R}^n$

Just like in the 1-dimensional Gaussian case, most probability measures we will define will be with respect to the gaussian, and will be defined with respect to a *density*. We will get back to this after upgrading many previous concepts like independence, random vectors, expected value, and change of variables.

Remember that the gaussian on $\mathbb{R}$ was $p(x) = \frac{1}{(2\pi)^{1/2}} e^{-x^2/2}$. By definition of $p_n(x)$, notice that:

$$p_n(x) = p(x_1)p(x_2) \cdots p(x_n)$$

So $p_n(x)$ is the product of 1-dimensional Gaussians! We have defined this as a probability measure, however, we should quickly verify that it does indeed sastify that $\mathbb{P}(\mathbb{R}^n) = 1$, and indeed by Fubini's:

$$\gamma_n(\mathbb{R}^n) = \int_{\mathbb{R}^n} p(x_1) \cdots p(x_n) dx_1...dx_n = \prod_{i=1}^{n} \int_{\mathbb{R}} p(x_i)dx_i = 1$$

Notice hwo this plays an analogous role to independence in the discrete case! In particular, If we let $\gamma_i : \mathbb{R}^n \to \mathbb{R}$, $\gamma_i(\vec{x}) = x_i$ be the coordinate function (also known as projection map), then the collection $\{\gamma_i\}_{i=1}^n$ can be seen as *independent Gaussian random variables*: For any nice set $A_i \subseteq \mathbb{R}$, by Fubini's Theorem in Calculus:

$$\begin{aligned}\mathbb{P}(g_1 \in A_1,\ ...,\ g_n \in A_n) &= \gamma_n(A_1 \times A_2 \times \cdots \times A_n)\\ &= \int_{A_1 \times A_2 \times \cdots \times A_n} p_n(\vec{x})d\vec{x}\\ &= \prod_{i=1}^n \int_{A_i} p(x)dx\\ &= \prod_{i=1}^n \gamma(A_i)\end{aligned}$$

We want to conclue that $\prod_{i=1}^n \gamma(A_i) = \prod_{i=1}^n \mathbb{P}(\gamma_i \in A_i)$ however in order to do that we must make sure that $\mathbb{P}(\gamma_i \in A_i) = \gamma(A_i)$. This comes from by fixing A_i in the first equality in the previous string of equalities and letting all other $A_j = \mathbb{R}$, $i \neq j$. Then:

$$\mathbb{P}(g_i \in A_i) = \gamma(A_i)$$

Thus, we can conclude that

$$\mathbb{P}(g_1 \in A_1,\ ...,\ g_n \in A_n) = \prod_{i=1}^n \mathbb{P}(\gamma_i \in A_i)$$

And so we'd say that $g_1, ..., g_n$ are independent random variables with Gaussian distribution $N(0, 1)$.

Just like in the discrete case, we can define random vectors in the continuous case. As a simple first example, let $g : \mathbb{R}^n \to \mathbb{R}^n$, with every component function $g_i : \mathbb{R}^n \to \mathbb{R}$ being $g_i(\vec{x}) = x_i$. Then any nice set $A \subseteq \mathbb{R}^n$ we can write:

$$\mathbb{P}(g \in A) = \gamma(A) = \int_A p_n(\vec{x})d\vec{x}$$

We would say that the random vector g *has density* $p_n(x)$. Since the density is $p_n(x)$, we call such random vectors *standard gaussian random vectors.*

Given any function $X : \mathbb{R}^n \to \mathbb{R}$, we will call it a random variable if it is a nice function (or if you know Lebesgue integration, it is an integrable funciton with respect to the Lebesgue measure). Thus, we may assume that $X^{-1}(A)$ will be a nice (or rectifiable) set for any interval A. Given a definition of a random variable, we can define the expected value:

Definition 4.2.2: Expected Value For $\mathbb{R}^n$ Gaussian

Let $(\mathbb{R}^n, \gamma_n)$ be the probability space and X a random variable. Then it's *expected value* is:

$$\mathbb{E}X = \int_{\mathbb{R}^n} X(\vec{x})p_n(\vec{x})d\vec{x}$$

and is considered well-defiend if $\mathbb{E}|X| < \infty$

Like in the 1-dimensional case, we would like it if X were defined as a distribution with respect to a density $p_X(t)$ ($p_X(t) \geq 0$, $\int_{\mathbb{R}} p_X(t)dt = 1$), so that we can compute (for some interval A):

$$\mathbb{P}(X \in A) = \int_{X^{-1}(A)} p(\vec{x})d\vec{x} = \int_A p_X(t)dt$$

This is certainly possible: as before, we can define $F(x) = \mathbb{P}(X \leq x)$ (sine the range of X is $\mathbb{R}$, $\leq$ is wel-defined), and find $p_X(t)$. we need to make sure that the change of variables will indeed return the same result, however this is a simple upgrade of the similar lemma from before:

Lemma 4.2.1: Change Of Variables For $\mathbb{R}^n$ Integral

Let $(\mathbb{R}^n, \gamma_n)$ be a probability space. and $f(X)$ a random variable. Then

$$\mathbb{E}f(X) = \int_{\mathbb{R}} f(t)p_X(t)dt$$

given $\mathbb{E}|f(x)| < \infty$

Proof **:**
This proof the same as in lemma 4.1.2 as long as f is defined such that

$$(f \circ X)^{-1}(A) = X^{-1}(f^{-1}(A))$$

is nice (rectifiable) for any interval A. If you know Lebesgue integration, it suffices that f be Borel measurable.

A quick side note: we have seen the change of variables in many settings at this point: discrete, $\mathbb{R}$ and $\mathbb{R}^n$. The proof seemed to work in all setting irregardless of the probability space. There is in fact a generalization of this proof that would be covered in a course on measure theory.

Before continuing to explore other distributions, we introduce an important result for computing them:

Theorem 4.2.1: Gaussian Stability

Let $a = (a_1, ..., a_n) \in \mathbb{R}^n$. and define the random variable

$$X = (a, g) = a_1g_1 + \cdots + a_ng_n$$

Then $X \sim N(0, |a|^2)$ with variance $|a|^2$.

Remember from example 4.4 that $a_ig_i \sim N(0, a_i^2)$, and so we are saying that the sum of independent Guassian random variable is another gaussian random variable, justifying the name "gaussian stability"

Proof **:**
If $a = 0$, then $X \sim N(0, 0) = 0$ (the zero function), and so this case is trivial. For $a \neq 0$, We compute the cdf:

$$\mathbb{P}(X \leq t) = \mathbb{P}((a, g) \leq t) = \int_H p_n(x)dt$$

where $t \in \mathbb{R}$ and $H = \{x \mid \sum_i a_i g_i(x) \leq t\}$ (i.e., we translated the restriction of $\mathbb{P}$ into a bound of $\int$). the set H is messy to understand, so we will simplify it using a clever change of variable. Let $q_1 = \frac{a}{|a|}$ be the normalized vector a, and let $q_2, ..., q_n$ be arbitrary vector that together with q_1 form an orthonormal basis for $\mathbb{R}^n$. Let Q be the matrix with rows $q_1, ..., q_n$. Since Q is orthogonal $\det(Q) = 1$, and $|Qx| = |x|$. This makes Q for a perfect candidate for a change of variables: Letting $y = Qx$, we get that the bound is now:

$$y_1 = \sum_i \frac{a_i}{|a|} x_i = \frac{1}{|a|} \sum a_i x_i$$

and the density becomes:

$$p_n(Qx) = \frac{1}{(2\pi)^{n/2}} e^{\frac{-|Qx|^2}{2}} = \frac{1}{(2\pi)^{n/2}} e^{\frac{-|y|^2}{2}} = p_n(y)$$

Finally, putting this all together, we get:

$$\begin{aligned}
\int_H p_n(x)dx &= \int_H p_n(y)dy \\
&= \int_{|a|y_1 \leq t} p(y_1) \cdots p(y_n) dy_1 \cdots dy_n \\
&= \int_{|a|y_1 \leq t} p(y_2) \cdots p(y_n) dy_2 \cdots dy_n p(y_1) dy_1 && \text{Fubini} \\
&\overset{!}{=} \int_{|a|y_1 \leq t} p(y_1) dy_1 \\
&= \mathbb{P}(|a|g_1 \leq t)
\end{aligned}$$

where the $\overset{!}{=}$ inequality comes from the constraint $|a|y_1 \leq t$ only applying to the first variable, and so all the other iterated integrals range freely over $\mathbb{R}$, and so their integral is 1. Thus, we have reduced the cdf to

$$\mathbb{P}(X \leq t) = \mathbb{P}(|a|g_1 \leq t)$$

the last cdf defines the distribution with density $|a|g_1$, which in example 4.4 we showed was $N(0, |a|^2)$, which is what we sought to show.

Example 4.5: Stability And Change Of Variables

Combining the gaussian stability with the change of variables for $\mathbb{R}^n$, we get that:

$$\begin{aligned}
\mathbb{E}(f(a, g)) &= \mathbb{E}(f(a_1 g_1 + \cdots + a_n g_n)) \\
&= \int_{\mathbb{R}} f(t) p_|a|(t) dt \\
&= \frac{1}{\sqrt{2\pi}|a|} \int_{\mathbb{R}} f(t) e^{\frac{-x^2}{2|a|^2}}
\end{aligned}$$

This formula will come back in the proof of the CLT since $f^{-1}(A)$ will be an interval and, if we set $X = (a, g)$, then $(f \circ X)^{-1}(A)$ will be a region bounded by parrallel hyperplanes.

Finally, note that $\mathbb{E}f(X)$ does *not* depend on a chosen distribution for X or or probability space on which X is defined. Because of this, some authors write:

$$\mathbb{E}(f(a,g)) = \mathbb{E}f(|a|g_1)$$

since $|a|g_1$ also has distribution $N(0,|a|^2)$.

4.2.1 General Gaussian distribution on $\mathbb{R}^n$

We will now consider more general Gaussian random variables, in particular gaussian random vectors (or gaussian vectors for short). Let A be an n by n matrix and $g = (g_1, ..., g_n)^T$ be a guassian vector. Let:

$$X = (X_1, ..., X_n)^T = Ag = A(g_1, ..., g_n)^T$$

We will show that the distribution of X will de dependent on:

$$C = AA^t$$

To get there, we start by defining some terms:

Definition 4.2.3: Covariance Matrix

Let X be a random vector. Then the *covariance matrix* is:

$$\text{Cov}(X) = [\text{Cov}(X_i, X_j)]_{i,j\leq n}$$

assuming all the entries are well-defined.

The covariance matrix is linked to the usual covariance of a random variable ($\text{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X)) \cdot (Y - \mathbb{E}(Y)])$ if we take $\vec{\mu} = \mathbb{E}(X) = (\mathbb{E}(X_1), ..., \mathbb{E}(X_n))^T$. Then:

$$\text{Cov}(X) = \mathbb{E}[(X - \vec{\mu})(X - \vec{\mu})^T]$$

As we've seen in example 4.4, $\mathbb{E}X_{ij} = 0$, so $\mathbb{E}X_i = 0$ given $X = Ag$. And so, in this special case, we get that:

$$\text{Cov}(Ag) = \mathbb{E}Ag(Ag)^T = \mathbb{E}Agg^TA = A(\mathbb{E}gg^T)A^T \overset{!}{=} AA^T = C$$

where $\mathbb{E}gg^T = [\mathbb{E}g_ig_j]_{i,j\leq n} = I$ using example 4.4. With this, we can explore the cases where $\det C \neq 0$ and $\det C = 0$

Lemma 4.2.2: Distribution Of $\det(C)$

Let $X = Ag$, $C = AA^T$, and $\det(C) \neq 0$. Then the distribution of C has density:

$$p_C(\vec{x}) = \frac{1}{(2\pi)^{n/2}} \frac{\sqrt{\det(C)}}{e}^{\frac{-1}{2}\vec{x}^TC^{-1}\vec{x}}$$

Meaning:

$$\mathbb{P}(X \in \Delta) = \int_\Delta p_C(\vec{x})d\vec{x}$$

for nice set Δ

***Proof* :**
Let Δ be any nice set of $\mathbb{R}^n$ (for ex. a rectangle). Since $X = Ag$, and $\det(C) \neq 0$, then A is invertible, and so we can simply write:

$$\mathbb{P}(Ag \in \Delta) = \mathbb{P}(g \in A^{-1}) = \frac{1}{(2\pi)^{n/2}} \int_{A^{-1}\Delta} e^{\frac{-|x|^2}{2}} d\vec{x}$$

Since A is invertible, we can do the change of variables $y = Ax$, so $x = A^{-1}y$, and get:

$$\mathbb{P}(Ag \in \Delta) = \int_{\Delta} \frac{1}{(2\pi)^{n/2}} e^{\frac{-|A^{-1}y|^2}{2}} \frac{1}{|(A)|} d\vec{y}$$

at this point, we just manipulate notation. Notice that:

$$\det(C) = \det(AA^T) = \det(A)\det(A)^T = |\det(A)|^2$$

so $\sqrt{\det(C)} = |\det(A)|$. Next

$$|A^{-1}y|^2 = (A^{-1}y)(A^{-1}y)^T = y^T(A^{-1})^T A^{-1}y = y^T(AA^T)^{-1}y = y^T C^{-1} y$$

and so:

$$\mathbb{P}(Ag \in \Delta) = \int_{\Delta} \frac{1}{(2\pi)^{n/2}} \frac{1}{\sqrt{(C)}} e^{\frac{y^T C^{-1} y}{2}} d\vec{y} = \int_{\Delta} p_C(\vec{y}) d\vec{y}$$

which is $N(0, C)$, completing the proof.

Example 4.6: Orthogonal Maps
If we take Q be orthogonal matrix (i.e. $QQ^T = I$) and take $X = Qg$, then lemma 4.2.2 gives us:

$$\mathbb{P}(X \in \Delta) = \int_{\Delta} p_I(y) dy = \int_{\Delta} \frac{1}{(2\pi)^{n/2}} e^{\frac{-|\vec{x}|^2}{2}} d\vec{x}$$

i.e. $p_I(x) = p_n(x)$.

If $\det(C) = 0$ or $\det(A) = 0$, then this would be our first example of a random variable with no density (in $\mathbb{R}^n$)! We can still describe this random variable in terms of C with a bit more work. We first review some linear algebra facts:

1. Show that $\mathrm{Cov}(X)$ is symmetric and non-negative
2. If $C = AA^T$ and $C = QDQ^T$ is the eigen-decomposition of C given appropriate orthogonal matrices Q and $D = \mathrm{diag}(\lambda_1, ..., \lambda_n)$. Then $A = QD^{1/2}R$ for appropriate orthogonal matrix R and $D^{1/2} = \mathrm{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n})$ (hint: use that $B = Q^T A$ satisfies $BB^T = D$ to represent $B = D^{1/2}R$)

At this point Pachenko ellaborated on how to use the aforementioned inforamtion, which I'll ommit for nwo

4.3 Central Limit Theorem

Before stating the theorem, we need to be able to define independent random variables with arbitrary distributions on the same probability space. Let $X_1, ..., X_n$ be random variables on $(\Omega, \mathbb{P})$ with respective *independent* distribution $\mathbb{P}_1, ..., \mathbb{P}_n$, meaning

$$\mathbb{P}(X_1 \in A_1, ..., X_n \in A_n) = \prod \mathbb{P}(X_i \in A_i) = \prod \mathbb{P}_i(A_i)$$

We have shown such random variables exist when we took the product space of random variables (recall definition 1.5.6). This definition worked well since all the variables were discrete, however we may now work with both continuous and discrete random variables. To update our idea of a product space to this more general scenario, let $X_1, ..., X_n$ be random variables (discrete or continuous). If X_i is discrete, let Ω_i be the range of X_i. If X_i is continuous, let $\Omega_i = \mathbb{R}$. Let

$$\Omega = \Omega_1 \times \cdots \times \Omega_n$$

with $\mathbb{P}$ defined as:

$$\mathbb{P}(A_1 \times \cdots \times A_n) = \prod \mathbb{P}_i(A_i)$$

where A_i can be any subset for discrete random variable, and A_i is the union of intervals in the continuous case. As before, we can define coordinate functions $X_i : \Omega \to \mathbb{R}$ to be:

$$X_i(\omega) = X_i(\omega_1, ..., \omega_n) = \omega_i$$

and we see that:

$$\mathbb{P}(X_1 \in A_1, ..., X_n \in A_n) = \mathbb{P}(\omega_1 \in A_1, ..., \omega_n \in A_n) = \mathbb{P}(A_1 \times \cdots \times A_n) = \prod_i \mathbb{P}_i(A_i)$$

showing us that the coordinate functions are still independent random variables with distributiosn $\mathbb{P}_1, ..., \mathbb{P}_n$.

Acutally, there is a slight problem in how we defined $\mathbb{P}$: we only showed the result for rectangles. To extend to general sets A, we would take the integral:

$$\mathbb{P}(A) = \int_A 1 d\mathbb{P}_1(\omega_1) \cdots d\mathbb{P}_n(\omega_n)$$

where we are automatically going to always apply Fubini's theorem. If we are integrating over a discrete space, $\int_{\Omega_i} \cdots d\mathbb{P}_i(\omega_i)$, then we would be taking the *sum* $\sum ...d\mathbb{P}_i(\omega_i)$. When X is continuous, we take the integral.

With this new knowledge we are ready to prove the Central Limit Theorem (commonly abbreviated to CLT). This proof of the CLT is called the *Lindeberg's method.* It is a particularly enlightening proof of the CLT since it shows how the Gaussian distribution emerges in the limit due to its stability property.

For the CLT, the random variables can either be i.i.d. or indepenent. If $X_1, ..., X_n$ are independent with distributions $\mathbb{P}_1, ..., \mathbb{P}_n$, then denote

$$\mathbb{E}X_i = \mu_i, \quad \text{Var}(X_i) = \sigma_i^2 < \infty$$

and assume that

$$\frac{\mathbb{E}|X_i|^3}{\sigma_i^3} \leq K, \infty$$

for all $i \leq n$ and some constant K. This is the Lindeberg propety. Now define:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i - \mu_i}{\sigma_i}$$

which is the "normalized sum" . If all the variable are i.i.d., $\mu = \mu_i$, $\sigma = \sigma_i$, then the sum becomes

$$Z_n = \frac{S_n - \mu}{\sqrt{n}\sigma} = \frac{\sqrt{n}(\overline{X} - \mu)}{\sigma}$$

and the Lindeberg condition reduces to

$$\mathbb{E}|X_1|^3 < \infty$$

With these, we get the main result of this chapter:

Theorem 4.3.1: Central Limit Theorem (CLT)

Let $X_1, ..., X_n$ be independent [identical?] random variables, with the Lindeberg condition, and let Z_n, μ_i, σ_i be defned as before. Then for all $t \in \mathbb{R}$:

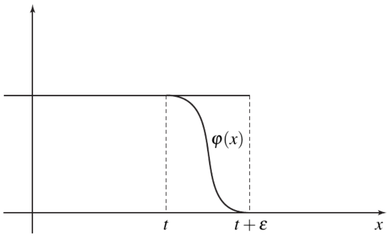
$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq t) = \frac{1}{2\pi} \int_{-\infty}^{t} e^{x^2/2} dx$$

which is also known as *converge in distribution* since it is the distribution's that are convering to each other.

Proof **:**
We will first reduce the proof which will work better for the proof technique we'll use. We can start by taking

$$\mathbb{P}(Z_n \leq t) = \mathbb{E}I(Z_n t)$$

Next, we know that we can make I "arbitrarily smooth" like so:



We will want this later to use Taylor's Theorem. As in the image, let φ be this smooth approximation. In particular, we'll want $|\varphi'''| \leq c$ for some c that is allowed to depend on ε

and

$$I(Z_n \leq t) \leq \varphi(x) \leq I(Z_n \leq t + \varepsilon)$$

Now, it is equivalent to show that:

$$\lim_{n\to\infty} \varphi(Z_n) = \varphi(g) = \int_{\mathbb{R}} \varphi(x)p(x)dx \tag{4.2}$$

where g is the standard gaussian variable and $p(x)$ is the usual density. The reason it is equivalent come down to the following: by monotonicity of applying the expected value:

$$\mathbb{E}I(Z_n \leq t) \leq \mathbb{E}\varphi(Z_n) \leq \mathbb{E}I(Z_n \leq t + \varepsilon)$$

The first inequality implies, assuming 4.2:

$$\begin{aligned}\limsup_{n\to\infty} \mathbb{E}I(Z_n \leq t) &\leq \lim_{n\to\infty} \varphi(Z_n)\\ &= \int_{\mathbb{R}} \varphi(x)p(x)dx\\ &\leq \int_{-\infty}^{t+\varepsilon} p(x)dx\end{aligned}$$

where the last inequality is true since $\varphi(x) \leq I(Z_n \leq t + \varepsilon)$. Similarly:

$$\begin{aligned}\liminf_{n\to\infty} \mathbb{E}I(Z_n \leq t) &\geq \lim_{n\to\infty} \varphi(Z_n)\\ &= \int_{\mathbb{R}} \varphi(x)p(x)dx\\ &\geq \int_{-\infty}^{t} p(x)dx\\ &\geq \int_{-\infty}^{t-\varepsilon} p(x)dx\end{aligned}$$

where the 2nd last inequality is true since $\varphi(x) \geq I(Z_n \leq t)$. Thus, we get $\mathbb{E}I(Z_n \leq t)$ is squizzed in the limit:

$\int_{-\infty}^{t-\varepsilon} p(x)dx \leq \lim_{n\to\infty} \mathbb{E}I(Z_n \leq t) \leq \int_{-\infty}^{t+\varepsilon} p(x)dx$and since ε was arbitrary, we get that as $\varepsilon \to 0$, we get the statement in the theorem. Thus, we'll prove equation 4.2

First, let's compactify notation. Let:

$$Y_i = \frac{X_i - \mu_i}{\sigma_i}$$

so we get:

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i$$

The limitations on the X_i's gets translated to Y_i's like so:

$$\mathbb{E}Y_i = 0,\ \mathbb{E}Y_i^2 = 1,\ \mathbb{E}|Y_i|^3 \leq K < \infty$$

Now, suppose $X_1, ..., X_n$ are random variables constructed on the same probability space and $g_1, g_2, ..., g_n$ are i.i.d. standard gaussian random variables. If we wanted to, we can think of $g_1, g_2, ..., g_n$ as as $X_{n+1}, X_{n+2}, ..., X_{2n}$ and use the product space with $2n$ coordinates instaed of n, but we will do as we said in the first sentence. Consider the random variable:

$$g = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g_i$$

By the stability of the gaussian, we have that $g \sim N(0,1)$, and so by the change of variables formula

$$\mathbb{E}\varphi\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\right) = \int_{\mathbb{R}} \varphi(x)p(x)dx$$

Thus, this g agrees with the g in the equation 4.2, and so to prove equation 4.2, it suffices to show that:

$$\mathbb{E}\varphi\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i\right) - \mathbb{E}\varphi\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i\right) \tag{4.3}$$

get's small as $n \to \infty$. The way we will do this is through Lindeberg's method. In it, we compare Z_n and g, or more precisely

$$Z_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i \qquad \text{and} \qquad g = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i$$

by taking "small steps" through replacing $\frac{Y_i}{\sqrt{n}}$ with $\frac{g_i}{\sqrt{n}}$ one by one. Here is where the stability requirement was so important since we can replace g as sums of smaller g_i's. For $1 \leq i \leq n+1$, define:

$$T_i := \frac{1}{\sqrt{n}}(g_1 + \cdots + g_{i-1} + Y_i + \cdots + Y_n)$$

so that $T_1 = Z_n$ and $T_{n+1} = g$. With this, we can write equation 4.3 as

$$\mathbb{E}\varphi(Z_n) - \mathbb{E}\varphi(g) = \sum_{i=1}^{n}\left(\mathbb{E}\varphi(T_i) - \mathbb{E}\varphi(T_{i+1})\right)$$

as a telescoping sum of "small increments" (similar to the proof of Azuma's). Notice that the terms T_i and T_{i+1} differ in only 1 term, namely $\frac{Y_i}{\sqrt{n}}$. Since all the other terms are shared, we can represent them as

$$T_i := \frac{1}{\sqrt{n}}(g_1 + \cdots + g_{i-1} + Y_{i+1} + \cdots + Y_n)$$

so we get that consecutive terms are:

$$T_i = S_i + \frac{Y_i}{\sqrt{n}} \qquad T_{i+1} = S_i + \frac{g_i}{\sqrt{n}}$$

Now, applying absolute values and using the triangle inequality, we get that:

$$|\mathbb{E}\varphi(Z_n) - \mathbb{E}\varphi(g)| \leq \sum_{i=1}^{n} |\mathbb{E}\varphi(T_i) - \mathbb{E}\varphi(T_{i+1})|$$

We'll be gaining more information of the right hand side using Taylor expansion. Recall that $|\varphi^{'''}(x)| \leq c$. Thus, doing the Taylor polynomial of order 2, we get:

$$\left|\varphi(T_i) - \varphi(S_i) - \varphi'(S_i)\frac{Y_i}{\sqrt{n}} - \varphi''(S_i)\frac{Y_i^2}{2n}\right| \leq \frac{c|Y_i|^3}{6n^{3/2}}$$

$$\left|\varphi(T_{i+1}) - \varphi(S_i) - \varphi'(S_i)\frac{g_i}{\sqrt{n}} - \varphi''(S_i)\frac{g_i^2}{2n}\right| \leq \frac{c|g_i|^3}{6n^{3/2}}$$

Now, opening up the absolute value (i.e. $-b \leq a \leq b$ instead of $|a| \leq b$) we tan take the absolute value on all sides of both. When taking the absolute value of the 1st and 2nd order term:

$$\mathbb{E}\left[\varphi'(S_i)\frac{Y_i}{\sqrt{n}}\right],\ \mathbb{E}\left[\varphi''(S_i)\frac{Y_i^2}{2n}\right],\ \mathbb{E}\left[\varphi'(S_i)\frac{g_i}{\sqrt{n}}\right],\ \mathbb{E}\left[\varphi''(S_i)\frac{g_i^2}{2n}\right]$$

notice that S_i is a function of the coordinates that do *not* include Y_i and g_i. By independence, we use the Fubini formula to calculate the expected value to get:

$$\mathbb{E}\left[\varphi'(S_i)\frac{Y_i}{\sqrt{n}}\right] = \mathbb{E}\varphi'(S_i)\mathbb{E}\frac{Y_i}{\sqrt{n}} = 0$$

since $\mathbb{E}\frac{Y_i}{\sqrt{n}} = 0$. Similarly for the other one:

$$\mathbb{E}\left[\varphi'(S_i)\frac{g_i}{\sqrt{n}}\right] = \mathbb{E}\varphi'(S_i)\mathbb{E}\frac{g_i}{\sqrt{n}} = 0$$

since $\mathbb{E}g_i = 0$. Since we have subtracted the expectation in $Y_i = \frac{X_i - \mu_i}{\sigma_i}$, the first order terms in the Tailor expansions match. Similarly:

$$\mathbb{E}\left[\varphi''(S_i)\frac{Y_i^2}{\sqrt{n}}\right] = \mathbb{E}\varphi''(S_i)\mathbb{E}\frac{Y_i^2}{\sqrt{n}} = \frac{1}{2n}\mathbb{E}\varphi''(S_i)$$

Sicne $\mathbb{E}Y_i^2 = 1$ and

$$\mathbb{E}\left[\varphi''(S_i)\frac{g_i^2}{\sqrt{n}}\right] = \mathbb{E}\varphi''(S_i)\mathbb{E}\frac{g_i^2}{\sqrt{n}} = \frac{1}{2n}\mathbb{E}\varphi''(S_i)$$

Sicne $\mathbb{E}g_i^2 = 1$. Since we scaled by σ in $Y_i = \frac{X_i - \mu_i}{\sigma_i}$, the second order terms in the taylor expantions match. Since the 0th order terms are the same, namely $\mathbb{E}\varphi(S_i)$, the difference is only in the third order error terms:

$$|\mathbb{E}\varphi(T_i) - \mathbb{E}\varphi(T_{i+1})| \leq \frac{c(\mathbb{E}|Y_i|^3 + \mathbb{E}|g_i|^3)}{6n^{3/2}}$$

As we've shown in the compactification of notation, we have that $\mathbb{E}|Y_i|^3 \leq K$ and $\mathbb{E}|g_i|^3$ is the same for all $i \leq n$, and so it is in fact an another constant so

$$|\mathbb{E}\varphi(T_i) - \mathbb{E}\varphi(T_{i+1})| \leq \frac{c'}{6n^{3/2}}$$

given appropriate c'. Notice that these third order terms have an order smaller than $n^{3/2}$ than the number of "steps" in the telescoping sum (namely n) and so we get the final result of :

$$|\mathbb{E}\varphi(Z_n) - \mathbb{E}\varphi(g)| \leq \sum_{i=1}^{n} |\mathbb{E}\varphi(T_i) - \mathbb{E}\varphi(T_{i+1})| \leq \frac{c'}{\sqrt{n}}$$

and since c' is independent of n, as $n \to \infty$, we see that $|\mathbb{E}\varphi(Z_n) - \mathbb{E}\varphi(g)| \to 0$, which we showed is equivalent to our original result, completing the proof

Using the CLT is all about how to setup the scenario correctly:

Example 4.7: CLT Use Case

Let's say that all we know about the checkout time of a grocery store is that the mean per customer is $\mu = 5$ minutes, and the variance if $\sigma = 2$ minutes. We want to estimate the probability that a cashier checkout at laest 50 customers during a 4h shift. In other words, what is the probability that 50 customers will be surved in under 240 minutes?

Let T_i be the time it takes to server the ith customer, and $S_n = \sum_i^n T_i$. Then we want to estimate $\mathbb{P}(S_n \leq 240)$. Since each T_i are (essentially) idependent and i.i.d., we have that:

$$Z_n = \frac{240 - 50 \cdot 5}{2\sqrt{50}} \leq \frac{240 - 250}{2\sqrt{50}} \approx 0.7$$

and pulging it into the approximation:

$$\mathbb{P}(Z_n \leq -0.7) \approx \frac{1}{2\pi} \int_{-\infty}^{-0.7} e^{\frac{-x^2}{2}} dx \approx 0.242$$

so, there is approximately 24% chance that 50 customers are served.

There are many different variation of the CLT. A common one is where the distributions of $X_1, ..., X_n$ also depend in some way on n. The random variables $X_{n1}, ..., X_{nn}$ is still independent, but the sequences varies with respect to n. This is usually called the *triangular array*. For notational simplicity, let:

$$\nu_{ni} = \mathbb{E}X_{ni},\ \sigma_{ni}^2 = \text{Var}(X_{ni}) < \infty,\ D_n^2 = \sum_i^n \sigma_{ni}^2$$

Instead of usual normalizing Z_n, we will consider:

$$Z_n = \frac{S_n - \mathbb{E}S_n}{\sqrt{\text{Var}(S_n)}} = \frac{1}{D_n} \sum_i^n (X_{ni} - \mu_{ni})$$

If the X_{ni}'s are i.i.d., then this and the previous definition of Z_n match. Just like for the previous Z_n, we have $\mathbb{E}Z_n = 0$, $\text{Var}(Z_n) = 1$, but this time, the variance of each summand (i.e. σ_{ni}^2/D_{ni}^2) is not necessarily $1/n$. So now that we've changed Z_n, what condition do we impose on it so that CLT applies? It turns out that a sufficient condition is:

$$\lim_{n\to\infty} \frac{1}{D^3} \sum_i^n \mathbb{E}|X_{ni} - \mu_{ni}|^3 = 0$$

and is called the *Lyapunov condition*.

Theorem 4.3.2: CLT For Triangular Arrays

The CLT holds for Z_n that satisfy the Lyapunov condition.

***Proof* :**
The book states that the proof is very similar, where we would change $a_{ni} = \sigma_{ni}/D_n$ and represent $g = \sum_i^n a_{ni} g_i$. Some more hints where given but I will leave to when I prove CLT in the notes.

Example 4.8: Records In A Random Permutation
Let $(\pi_1, ..., \pi_n)$ be an n-tuple with elements from $\{1, ..., n\}$. We say that π_k is a *record* if it is greater than every π_j: $\pi_j < \pi_i$ for $j < i$. From here, Pachenko proceeds to carefully use CLT.

4.4 Distributions Related to Gaussian

In this section, we study many distributions related to the Gaussian distribution:

1. Gamma distribution
2. chi-squared distribution
3. F-distribution
4. Studient's t-distribution

These will be the focus of attention for simple linear regression. Linear regression was not covered in my year, but these still have interesting properties worth studying for their own right.

Definition 4.4.1: Gamma Function

For all $\alpha > 0$:

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$$

TO make this a distribution, we can divide by $\Gamma(\alpha)$ on both sides. If we further do the change of variable sof $x = \beta y$ for some $\beta > 0$, we get:

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

Labelling the function inside as $f_{a,b}(x)$, we give it a name:

Definition 4.4.2: Gamma Distribution

Let $f_{a,b}(x)$ be defined as above. Then it defines the *Gamma distribution with parameters α and β*, usually denoted $\gamma(\alpha, \beta)$ and it is the density of this distribution.

Using integration by parts, we can see that for any $n \in \mathbb{N}_{>0}$ we have

$$\Gamma(n) = (n-1)!$$

Example 4.9: Moments Of Gamma Distribution

Show that if $X \sim \Gamma(\alpha, \beta)$,

$$\mathbb{E}X^k = \frac{(\alpha+k-1)\cdots\alpha}{\beta^k}$$

In particular, the important values to remember are :

$$\mathbb{E}X = \frac{\alpha}{\beta} \qquad \mathbb{E}X^2 = \frac{(\alpha+1)\alpha}{\beta^2} \qquad \mathrm{Var}(X) = \frac{\alpha}{\beta^2}$$

Next, we prove a really important result about the density of adding two distributions (I really should move this sooner!)

Proposition 4.4.1: Density Of Sums Of Random Variables

Let X_1 and X_2 be independent random variables with densities $f(x)$ and $g(x)$. Then $X_1 + X_2$ has density

$$h(x) = \int_{-\infty}^{\infty} f(x-y)g(y)dy$$

For those who knwo some analysis, h is the convolution of f and g: $h = f * g$.

Proof **:**

We compute the cumultaive distribution of X_1+X_2. By independence and the Fubini representation, we get that:

$$\begin{aligned}
\mathbb{P}(X_1 + X_2 \leq t) &= \int_{\mathbb{R}}\int_{\mathbb{R}} I(x+y<t)f(x)g(y)dxdy \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} I(x<t-y)f(x)g(y)dxdy \\
&= \int_{\mathbb{R}}\left[\int_{-\infty}^{t-y} f(x)g(y)dx\right]dy \\
&= \int_{\mathbb{R}}\left[\int_{-\infty}^{t} f(z-y)g(y)dz\right]dy \\
&= \int_{-\infty}^{t}\left[\int_{\mathbb{R}} f(z-y)g(y)dy\right]dz \\
&= \int_{-\infty}^{t} h(z)dz
\end{aligned}$$

showing that h is indeed the density of $X_1 + X_2$, as we sought to show

Corollary 4.4.1: Sums Of Gaussians

Let $X_1, ..., X_n$ be independ where $X_1 \sim \Gamma(\alpha_1, \beta)$, ..., $X_n \sim \Gamma(\alpha_n, \beta)$. Then

$$\sum X_i = \Gamma(\alpha_1, ..., \alpha_n, \beta)$$

***Proof* :**

5

Markov Chains

Definition 5.0.1: Random Variable Sequence

Let $\{X_i\}$ be some collection of random variables. Then if we put them in an order, we will say they form a *sequence*.

Recall that we can write:

$$\mathbb{P}(X_1 = x_2, ..., X_n = x_n) = \prod_{k=0}^{n-1} \mathbb{P}(X_{k+1} = x_{k+1} | X_1 = x_1, ..., X_k = x_k)$$

where for $k = 0$ we have $\mathbb{P}(X_1 = x_1)$.

Definition 5.0.2: Markov Chain

Let $X_1, ..., X_n, ...$ be a sequence of random variables. Then this sequence is called a *markov chain* if

$$\mathbb{P}(X_{k+1} = x_{k+1} | X_1 = x_1, ..., X_k = x_k) = \mathbb{P}(X_{k+1} = x_{k+1} | X_k = x_k) \tag{5.1}$$

that is, the conditional propability only depends on the last random variable.

The property demonstrated in equation 5.1 is somtimes called the *Markov property* or *memoryles property*. In this chapter, we will limit ourselves to *finite state* markov chains, that is Markov chains where all the X_i's take values form a finite set:

$$X_i \in S = \{s_1, ..., s_n\}$$

A

Change of Variables

It is important to be comfortable with change of variables of integrals once you start learning about continuous random variables, and so here is an appendix on it.

B

Glossary

B.1 Distributions

1. Bernoulli Distribution: $X \sim B(p)$, $\mathbb{P}(X = 0) = (1 - p)$, $\mathbb{P}(X = 1) = p$. A fair Bernoulli distribution has $p = 1/2$

$$\mathbb{E}X = p \qquad \mathbb{E}X^2 = p \qquad \mathrm{Var}(X) = p(1-p) \qquad \mathbb{E}X^k = p \qquad \mathbb{E}e^{tX} = (1-p) + pe^t$$

2. Radamacher distribution: $X \sim \varepsilon(p)$, $\mathbb{P}(\varepsilon = -1) = \mathbb{P}(\varepsilon = 1) = 1/2$

$$\mathbb{E}X = 0 \qquad \mathbb{E}X^2 = 1 \qquad \mathrm{Var}(X) = 1 \qquad \mathbb{E}X^k = \begin{cases} 1 & k \text{ is even} \\ 0 & k \text{ is even} \end{cases} \qquad \mathbb{E}e^{tX} = \cosh(t)$$

3. Binomial Distribution: $X \sim B(n, p)$. $\mathbb{P}(X = k) = \binom{n}{p}^k (1-p)^{n-k}$

$$\mathbb{E}X = np \qquad \mathbb{E}X^2 = np(1-p) + n^2p^2 \qquad \mathrm{Var}(X) = np(1-p) \qquad \mathbb{E}X^k = \sum_{i=1}^{k} \left\{ {k \atop i} \right\} n^{\underline{i}} p^i \qquad \mathbb{E}e^{tX} = ((1-p) + pe^t)^n$$

where $\left\{ {k \atop i} \right\}$ is the *sterling number* and $n^{\underline{i}} = n(n-1)\cdots(n-k+1)$ Note: $X = \sum_i^n X_i$ where the X_i's are i.i.d. Bernoulli random variables, giving us easy $\mathbb{E}X$ and $\mathrm{Var}(X)$

4. Geometric Distribution: $X \sim \mathrm{Geo}(p)$ $\mathbb{P}(X = n) = (1-p)^{n-1}p$

$$\mathbb{E}X = \frac{1}{p} \qquad \mathbb{E}X^2 = p \qquad \mathrm{Var}(X) = \frac{1-p}{p^2} \qquad \mathbb{E}X^k = \sum_{k=0}^{\infty} (1-p)^k p \cdot k^n \qquad \mathbb{E}e^{tX} = (1-p) + \frac{pe^t}{1-(1-p)e^t}$$

where $t < -\ln(1-p)$ for the MGT. Note that higher moments of the geometric distribution don't seem to admit a nice simplification.

5. negative geometric distribution: seems complicated, not sure if needed, will pass for now

6. Poisson Distribution: $X \sim \text{Poiss}(\lambda)$ $\mathbb{P}(X=k) = \frac{\lambda^k}{k!}e^{-\lambda}$

   $$\mathbb{E}X = \lambda \qquad \mathbb{E}X^2 = \lambda + \lambda^2 \qquad \text{Var}(X) = \lambda \qquad \mathbb{E}X^k = \sum_{i=1}^{k} \lambda^k \left\{ {k \atop i} \right\} \qquad \mathbb{E}e^{tX} = \exp(\lambda(e^t - 1))$$

   Note that $\lim_{n\to\infty} B\left(n, \frac{\lambda}{n}\right) = \text{Poiss}(n)$, that the sum of two Poisson with λ_1 and λ_2 is Poisson with $\lambda_1 + \lambda_2$

7. Gaussian Distribution: Given $p(x) = \frac{1}{\sqrt{2\pi}}e^{\frac{-x^2}{2}}$ and $\mathbb{P}(A) = \int_A p(x)dx = \gamma(A)$ then $X \sim g$, $g(x) = x$, then:

   $$\mathbb{P}(g \in A) = \gamma(\{x \mid x \in A\}) = \gamma(A)$$

   Other notation: $X \sim N(0,1)$. More generally, if $P_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$, then $X \sim N(\mu, \sigma^2)$

   $$\mathbb{P}(X \le t) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \qquad \mathbb{E}X = \mu \qquad \text{Var}(X) = \sigma^2 \qquad \mathbb{E}e^{tX} = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$

   where $\text{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt$

8. Uniform Distribution: Let $p_X(x) = \frac{1}{b-a}I(a \le x \le b)$ be the density. Then $\mathbb{P}(X \in A)$ is called the uniform distribution.

   $$\mathbb{P}(X \le t) = \begin{cases} 0 & t < a \\ \frac{x-a}{b-a} & a \le t \le b \\ 1 & t > b \end{cases} \qquad \mathbb{E}X = \frac{1}{2}(a+b) \qquad \mathbb{E}X^2 = \frac{b^3 - a^3}{3b - 3a} \qquad \text{Var}(X) = \frac{1}{12}(b-a)^2$$

   $$\mathbb{E}X^n = \frac{b^{n+1} - a^{n+1}}{n+1(b-a)} = \frac{1}{n+1}\sum_{k=0}^{n} a^k b^{n-k} \qquad \mathbb{E}e^{tX} = \begin{cases} \frac{e^{tb} - e^{ta}}{t(b-a)} & t \ne 0 \\ 1 & t = 0 \end{cases}$$

   note that the MGF reduces to $\mathbb{E}e^{tX} = \frac{\sinh(bt)}{bt}$.

9. Exponential Distribution: Let $p_X(x) = \lambda e^{-\lambda}I(x \ge 0)$ be the density. Then $\mathbb{P}(X \in A)$ is called the exponential distribution

   $$\mathbb{P}(X \le t) = 1 - e^{-\lambda t} \qquad \mathbb{E}X = \frac{1}{\lambda} \qquad \mathbb{E}X^2 = 0 \qquad \text{Var}(X) = \frac{1}{\lambda^2} \qquad \mathbb{E}X^n = \frac{n!}{\lambda^n}$$

10. Gamma Distribution: Let $f_{\alpha,\beta}(x) = \int_0^\infty \frac{\beta^\alpha}{\Gamma(\alpha)}y^{\alpha-1}e^{-\beta y}dy$ then $\mathbb{P}(X \in A)$ is called the Gamma distribution

    $$\mathbb{E}X = \frac{\alpha}{\beta} \qquad \mathbb{E}X^2 = \frac{(\alpha+1)\alpha}{\beta^2} \qquad \text{Var}(X) = \frac{\alpha}{\beta^2} \qquad \mathbb{E}X^k = \frac{(\alpha+k-1)\cdots\alpha}{\beta^k} \qquad \mathbb{E}e^{tX} = \left(1 - \frac{t}{\beta}\right)^{-\alpha}$$