Probably Correct Probability

Arjun Maneesh Agarwal

16 April 2025

The purpose of this handout is to do some interesting and useful things in probability which the course didn't cover or glossed over. We will try not to get too carried away and will probably try to provide value as in serving as a quick revision for the more confusing concepts. Basic familiarity with discrete and continuous probability is assumed.

We use $\text{dens}(X = x) = f_X(x) = \mathbb{P}(X = x)$ where X is an continuous random variable interchangeably throughout this handout. While this is an abuse of notation, it is a rather common one.

Furthermore, the subscript in f_X may be dropped at times when it is obvious which variable is being spoken about.

Table of contents

§1 Playing with Uniform Distribution

§1.1 Order Statistics

Take some n iid uniform variables named $X_{(1)}, X_{(2)}, ..., X_{(n)}$ with parameter a.

? Problem

What can be said about the distribution and density of $x_{(1)}$?

Solution

$$\begin{aligned} & f_{X_{(1)}}(x) \\ &= \mathbb{P}\Big(X_{(1)} \leq x\Big) \\ &= 1 - \mathbb{P}\Big(X_{(1)} > x\Big) \\ &= 1 - (\mathbb{P}(X > x))^n \\ &= 1 - \left(\frac{a-x}{a}\right)^n \end{aligned}$$

Thus, the density (by differentiation) is $\frac{n(a-x)^{n-1}}{a^n}$ for $0 \leq x \leq a$ and 0 elsewhere.

This however is much simpler than the real issue at hand.

? Problem

What can be said about the distribution and density of $x_{(k)}$?

Solution

Here we will use the fact $\lim_{h\to 0} \frac{\mathbb{P}(x \leq X \leq x+h)}{h} = f_X(x)$. Consider when only one element, $X_{(k)}$, is in the gap x to $x+h$. $k-1$ elements are in the area between 0 and x and $n-k$ between $x+h$ and a. This is clearly the dominant term in the expression of $\mathbb{P}(x \leq X \leq x+h)$ as the other terms have 2 or greater things in the gap x to $x+h$ and hence have a factor of h^2 or smaller. Thus,

$$\begin{aligned} & \mathbb{P}(x \leq X \leq x+h) \\ &= \binom{n}{1}\frac{h}{a}\binom{n-1}{k-1}\left(\frac{x}{a}\right)^{k-1}\binom{n-k}{n-k}\left(1-\frac{x+h}{a}\right)^{n-k} + \mathcal{O}(h^2) \\ &= n\binom{n-1}{k-1}\frac{h}{a}\left(\frac{x}{a}\right)^{k-1}\left(1-\frac{x+h}{a}\right)^{n-k} + \mathcal{O}(h^2) \end{aligned}$$

Using $\lim_{h\to 0} \frac{\mathbb{P}(x \leq X \leq x+h)}{h} = f_X(x)$,

$$
\begin{aligned}
&f_X(x) \\
&= \lim_{h\to 0}(\mathbb{P}(x \le X \le x+h)) \\
&= \lim_{h\to 0} \frac{n\binom{n-1}{k-1}\frac{h}{a}\left(\frac{x}{a}\right)^{k-1}\left(1-\frac{x+h}{a}\right)^{n-k} + \mathcal{O}(h^2)}{h} \\
&= \frac{n}{a}\binom{n-1}{k-1}\left(\frac{x}{a}\right)^{k-1}\left(1-\frac{x}{a}\right)^{n-k} \\
&= n\binom{n-1}{k-1}\left(\frac{x^{k-1}(a-x)^{n-k}}{a^n}\right)
\end{aligned}
$$

This is called the **beta distribution**. We can integrate it wrt x to get $F(x)$ but it leads to an incredibly messy form and should not be remembered or needed.

Corollary 1.1. $f_{X_{(n)}}(x) = n\frac{x^{n-1}}{a^n}$

Proof. Maybe try plugging n into the formula?! □

Exercise

Using the method shown above(or anything else you prefer), derive the joint distribution of $X_{(j)}, X_{(k)}$ i.e. $\mathbb{P}\Big(\big(X_{(j)} = x\big) \cap \big(X_{(k)} = y\big)\Big)$, given $x < y$ is

$$\frac{n!}{(j-1)!(k-j-1)!(n-k)!} \quad \frac{x^{j-1}(y-x)^{k-j-1}(a-y)^{n-k}}{a^n}$$

§1.2 Gaps in Uniform Process

Let the gap between 0 and $X_{(1)}$ be L_1 and those between $X_{(k)}$ and $X_{(k+1)}$ be L_{k+1}. This clearly implies

$$X_{(k)} = L_1 + L_2 + ... + L_k$$

If we find out about the distribution of these gaps then we can use linearity of expectations to get the expected value of $X_{(k)}$ without trying to integrate its distribution which seems too hard. Furthermore, we will see more uses for this later.

Claim 1.2. The gaps of the uniform process are equi-distributed.

We will see two proofs for this. One is a probabilistic one, the other is a more 'rigorous' one using conditional probability.

Probabilistic Proof. Imagine that we drop $n+1$ points on a circle of circumference a. It is obvious by symmetry that the gaps (measured along the circumference) so obtained are all equidistributed. On the other hand, this experiment is stochastically equivalent to the following experiment. Fix one point (call it 0) on the circle and then drop n more points at random. If we cut the circle at 0 and stretch it out over the interval $[0, a]$, then the gap distributions on $[0, a]$ are the same as those on the circle (the probability that another of the n points falls at the same place as 0 is zero). Therefore, the gap distributions on $[0, a]$ are all the same. This completes the proof. □

Rigrous Proof. After 'choosing' the first gap on an interval $[0, a]$ at say x, choosing the second gap becomes exactly the same as choosing the first gap on an interval $[x, a]$. Using Law of alternatives and the fact that L_1 is stochastically equivalent to $X_{(1)}$,

$$\begin{aligned}\mathbb{P}(L_2 = x_2) &= \int_0^{a-x_2} \mathbb{P}(L_2 = x_2 \mid L_1 = x)\mathbb{P}(L_1 = x)\,\mathrm{d}x \\ &= \int_0^{a-x_2} \mathbb{P}\Big(L_2 = x_2 \mid X_{(1)} = x\Big)\mathbb{P}\Big(X_{(1)} = x\Big)\,\mathrm{d}x \\ &= \int_0^{a-x_2} \mathbb{P}\Big(L_2 = x_2 \mid X_{(1)} = x\Big)\frac{n(a-x)^{n-1}}{a^n}\,\mathrm{d}x\end{aligned}$$

Let's now talk about the conditional probability. Consider,

$$\begin{aligned}&\mathbb{P}\Big(L_2 \le x_2 \mid X_{(1)} = x\Big) \\ &= 1 - \mathbb{P}\Big(L_2 > x_2 \mid X_{(1)} = x\Big) = 1 - \left(\frac{a - x_2 - x}{a - x}\right)^{n-1}\end{aligned}$$

And thus, $\mathbb{P}\Big(L_2 = x_2 \mid X_{(1)=x}\Big) = (n-1)\frac{(a-x_2-x)^{n-2}}{(a-x)^{n-1}}$

$$\begin{aligned}&\Rightarrow \mathbb{P}(L_2 = x_2) \\ &= \int_0^{a-x_2} \mathbb{P}\Big(L_2 = x_2 \mid X_{(1)} = x\Big)\frac{n(a-x)^{n-1}}{a^n}\,\mathrm{d}x \\ &= \int_0^{a-x_2} (n-1)\frac{(a-x_2-x)^{n-2}}{(a-x)^{n-1}}\frac{n(a-x)^{n-1}}{a^n}\,\mathrm{d}x \\ &= \frac{n(n-1)}{a^n}\int_0^{a-x_2} (a-x_2-x)^{n-2} \\ &= \frac{n(n-1)}{a^n}\left[-\frac{(a-x_2-x)^{n-1}}{n-1}\right]_0^{a-x_2} \\ &= \frac{n(a-x_2)^{n-1}}{a^n}\end{aligned}$$

which is same as L_1.

And we are done! □

Corollary 1.3. *The expectation of a gap in the uniform process with n sampling points is* $\frac{a}{n+1}$

Proof. Notice, $L_1 + L_2 + ... + L_n + L_{n+1} = a \Rightarrow \mathbb{E}(a) = a = (n+1)\mathbb{E}(L_i) \Rightarrow \mathbb{E}(L_i) = \frac{a}{n+1}$ □

Corollary 1.4. *The expectation of k-th order statistic is* $\frac{ka}{n+1}$

Proof. Notice $X_{(k)} = L_1 + L_2 + ... + L_k \Rightarrow \mathbb{E}\Big(X_{(k)}\Big) = \frac{ka}{n+1}$ □

Now that we have established some theory let's look at some classic problems.

? Needles on a stick

If we drop a set of needles, each of length h, on a stick of length b, what is the probability that none of the needles overlap?

Solution

We first restate the problem in terms of the Uniform process. The position of a given needle is completely determined by its left endpoint. The process of dropping n needles of length h on a stick of length b is then the same as dropping n points on the interval $[0, b-h]$. Let $a := b - h$.

Define $A_{a,n} = \mathbb{P}(L_2 \geq h, L_3 \geq h, ..., L_n \geq h)$ be the probability that n points dropped on length a have distance between each other greater than h. This is rather hard to compute directly as the gaps of the uniform process are not independent.

We will define $B_{a,n} = \mathbb{P}(L_2 \geq h, L_2 \geq h, ..., L_n \geq h, L_{n+1} \geq h)$ which seems to be the same problem with an additional condition that the last point needs to be h away from the right end of the stick.

Now notice, $\mathbb{P}\Big(B_{a,n} \mid X_{(n)} = t\Big) = \mathbb{P}\big(B_{t,n-1}\big)$ for $(n-1)h \leq t \leq a - h$ as once we have the last point and it makes sense with the conditions of the problem, we are reduced to choosing $n-1$ points between $[0, t]$.

Define $p_n(a) := \mathbb{P}\big(B_{a,n}\big)$. Using law of alternatives,

$$\begin{aligned} p_n(a) &= \int_{(n-1)h}^{a-h} \mathbb{P}\Big(B_{a,n} \mid X_{(n)} = t\Big)\mathbb{P}\Big(X_{(n)} = t\Big)\,\mathrm{d}t \\ &= \int_{(n-1)h}^{a-h} p_{n-1}(t)\frac{n}{a^n}t^{n-1}\,\mathrm{d}t \end{aligned}$$

Observe $p_1(a) = \frac{a-h}{a}$ as it is the probability that a single point on $[0, a]$ falls farther than distance h from a.

This along with the recursive formula we just derived give us $p_n(a) = \left(\frac{a-hn}{a}\right)^n$.

Finally, notice $\mathbb{P}\Big(A_{a,n} \mid X_{(n)} = t\Big) = \mathbb{P}\big(B_{t,n-1}\big) = p_{n-1}(t)$. Using law of alternatives,

$$\begin{aligned} P_{a,n} &= \int_{(n-1)h}^{a} \mathbb{P}\Big(A_{a,n} \mid X_{(n)} = t\Big)\mathbb{P}\Big(X_{(n)} = t\Big)\,\mathrm{d}t \\ &= \int_{(n-1)h}^{a} \left(\frac{t-(n-1)h}{t}\right)^{n-1}\frac{n}{a^n}t^{n-1}\,\mathrm{d}t \\ &= \frac{n}{a^n}\int_{(n-1)h}^{a} (t-(n-1)h)^{n-1}\,\mathrm{d}t \\ &= \frac{(a-(n-1)h)^n}{a^n} \\ &= \left(\frac{b-nh}{a}\right)^n \end{aligned}$$

An almost smiler method is used for another classic problem, which is left as exercise.

Scimitars on a Circle

A scimitar is a sword shaped like a circular arc (at least for this problem). Suppose that during a Turkish festival, a group of n Turks throw their scimitars independently and at random along a circle of circumference a. Suppose that each scimitar has arc length h along this circle. What is the probability that none of the scimitars overlap?

There are also discrete versions of these problems. Formalizing them and solving them is also left up to the reader.

§2 Memoryless Continuous Waiting Times

This section will deal with some processes fulfilling the above requirements. These seem like strong, arbitrary requirements. There is the obvious examples of **radioactive decay** and **customers in a call center/gas station/barber shop.** but then there is the fact that **neuron firing**[1] and **chewing-gum and tobacco splatter on street** to **star formation** can all be modelled and predicted by such models.[2]

§2.1 Waiting for it

Let's say you have tossed a fair coin k times and have k tails. What is the probability you will get a head in next n tosses?

Obviously, the irrelevant fact that we have had gotten k heads is irrelevant and it is intuitively obvious that on the next toss there is the same probability for heads as ever: the coin does not remember what took place in the past.

To put it mathematically in terms of single waiting time W_1 as

$$\mathbb{P}(W_1 > k + n \mid W_1 > k) = \mathbb{P}(W_1 > n)$$

However, real life is often continuous and most events we deal with are not governed by an abstract entity flipping an abstract coin during small discrete time intervals determining when the incident is to occur. However, making these interval shorter will help us approach an interesting, continuous distribution.

Definition

A positive continuous random variable W is said to have the exponential distribution when

$$\mathbb{P}(W > t + s \mid W > s) = \mathbb{P}(W > t)$$

for all positive t, s.

Theorem 2.1. *The only distribution fulfilling this condition if* $F_X(x) = 1 - e^{-\alpha x}$

Proof. We will re-write this condition as

[1] under the Poisson spike model

[2] Checkout the excellent list on wikipedia.org/wiki/Poisson_distribution#Occurrence_and_applications

$$\begin{aligned}
&\mathbb{P}(W > t+s) = \mathbb{P}(W > t)\mathbb{P}(W > s) \\
&\Rightarrow (1 - F(t+s)) = (1 - F(t))(1 - F(s)) \\
&\Rightarrow G(t+s) = G(t)G(s) \quad \text{where } G(x) \coloneqq 1 - F(x)
\end{aligned}$$

Using the fact that the CDF is differentiable almost everywhere, and partially differentiating with respect to t and s will give us

$$\frac{\partial}{\partial t} G(t+s) = G'(t+s)\frac{\partial}{\partial t}(t+s) = G'(t+s)$$

And similarly $\frac{\partial}{\partial s} G(t+s) = G'(t+s)$. Also note,

$$\frac{\partial}{\partial t} G(t)G(s) = G'(t)G(s) + G(t) \cdot 0 = G'(t)G(s)$$

and similarly $\frac{\partial}{\partial s} G(t)G(s) = G(t)G'(s)$. Thus,

$$\begin{aligned}
G(t+s) &= G(t)G(s) \\
\Rightarrow G'(t)G(s) = G'(t+s) &= G(t)G'(s) \\
\Rightarrow G'(t)G(s) &= G(t)G'(s) \\
\Rightarrow \frac{G'(t)}{G(t)} &= \frac{G'(s)}{G(s)}
\end{aligned}$$

As this must hold no matter what t and s are.

$$\begin{aligned}
\frac{G'(t)}{G(t)} &= C \\
\Rightarrow \log|G(t)| &= Ct + D \\
\Rightarrow |G(t)| = e^{Ct+D} &= e^{Ct}e^{D} = Ke^{Ct}
\end{aligned}$$

Note, as $0 \leq G(t) \leq 1$ as it is $1 - F(t)$ and $F(t)$ is a CDF, we can drop the absolute brackets. This gives us

$$F(t) = 1 - Ke^{Ct}$$

As $F(t)$ is a CDF, $\lim_{t\to\infty} F(t) = 1$. Thus, $C < 0$. We will let $C \coloneqq -\alpha$ where $\alpha > 0$.

As t is waiting time, $t \geq 0$, thus, $\lim_{t\to 0} F(t) = 0 \Rightarrow K = 1$.

Thus, $F(t) = 1 - e^{-\alpha t}$. And we are done. □

α here acts almost like the probability of the event occurring per some unit time aka the bias of our coin. It is normally called the rate parameter or intensity.

Corollary 2.2. *The PDF of exponential distribution is* $f(x) = \alpha e^{-\alpha t}$

Corollary 2.3. *The mean and variance of exponential distribution is* $\frac{1}{\alpha}$ *and* $\frac{1}{\alpha^2}$ *respectively.*

§2.2 Sum of Waiting Times

Given that we know amount of time to see one event, it is natural to ask the amount of time to see, say n events.

Claim 2.4.

$$\mathbb{P}(W_k = t) = \frac{\alpha^k t^{k-1}}{(k-1)!} e^{-\alpha t}$$

Proof. (B) This is obviously true for $k = 1$ as that is just the exponential distribution. While we don't need to , I will prove the case for $k = 2$ to make the induction make slightly more sense. Using the law of alternatives and the memory-less property,

$$\begin{aligned}
\mathbb{P}(W_2 = t) &= \int_0^t \mathbb{P}(W_2 = t \mid W_1 = x)\mathbb{P}(W_1 = x)\,\mathrm{d}x \\
&= \int_0^t \alpha e^{-\alpha(t-x)} \alpha e^{-\alpha x}\,\mathrm{d}x \\
&= \alpha^2 \int_0^t e^{-\alpha t}\,\mathrm{d}x \\
&= \alpha^2 t e^{-\alpha t}
\end{aligned}$$

(S) Let $\mathbb{P}(W_n = t) = \frac{\alpha^n t^{n-1}}{(n-1)!} e^{-\alpha t}$ for some $n > 0$. Then

$$\begin{aligned}
\mathbb{P}(W_{n+1} = t) &= \int_0^t \mathbb{P}(W_{n+1} = t \mid W_n = x)\mathbb{P}(W_n = x)\,\mathrm{d}x \\
&= \int_0^t \alpha e^{-\alpha(t-x)} \frac{\alpha^n x^{n-1}}{(n-1)!} e^{-\alpha x}\,\mathrm{d}x \\
&= \frac{\alpha^{n+1} e^{-\alpha t}}{(n-1)!} \int_0^t x^{n-1}\,\mathrm{d}x \\
&= \frac{\alpha^{n+1} e^{-\alpha t}}{(n-1)!} \left[\frac{x^n}{n}\right]_0^t \\
&= \frac{\alpha^{n+1} e^{-\alpha t}}{n!} t^n \\
&= \frac{\alpha^{n+1} t^{n+1-1}}{(n+1-1)!} e^{-\alpha t}
\end{aligned}$$

As required. Thus, by induction. we are done. □

This distribution is called the **Erlang distribution** after Agner Krarup Erlang, the man behind queuing theory. Notice that it has two parameters: α and k, where k is a positive integer. α denotes the rate, probability of event concurring in unit time or expected events per unit time as in exponential and k denotes the number of events observed.

The thing is, despite our insistence that the events are discrete, they happen or they don't; a generalization is possible. All terms of the formula make sense for arbitrary positive reals except the factorial. So, with great personal pain, I present to you **Gamma Function**.

Definition

A continuous random variable with positive real parameters k, α is Gamma distributed if

$$f(t) = \frac{\alpha^k t^{k-1}}{\Gamma(k)} e^{-\alpha t}$$

where α denotes the expected events per unit time and k is the number of events we wish to observe. Note, k is often called the shape parameter in this context.

In the definition of Gamma function, some people will observe some weirdly circular logic. I am not responsible for how this function was defined(and yes, this is exactly how it was conceived and the definition reached).

Definition

$$\Gamma(x) = \int_0^\infty \alpha^k t^{k-1} e^{-\alpha t} \, \mathrm{d}t$$

And by the substitution $u = \alpha t$

$$\Gamma(x) = \int_0^\infty u^{k-1} e^{-u} \, \mathrm{d}u$$

Remark

The gamma function should ideally be the solution to the functional equation $f(x+1) = xf(x)$ with $f(1) = 1$. But as it turns out, it is not the unique solution.

So after a lot of gymnastics, it was decided this ideal function will also have to satisfy the condition $\log(f(x))$ is convex. This was enough to make the gamma function the only solution. This is called Bohr–Mollerup theorem after Harold Bohr(Neils Bohr's brother) and Johannes Mollerup.

However, it is often believed that it was a reverse hunt. We already had reason to want the solution to be the gamma function and looked for conditions which would do so. The 'circler' probabilistic definition is due to Laplace and exists since 1836 while Bohr-Mollerup was proven in 1922.

Corollary 2.5. *Mean and Variance of Gamma distribution is $\frac{k}{\alpha}$ and $\frac{k}{\alpha^2}$ respectively.*

§2.3 Number of events in Time

We now ask the other obvious question,

Problem

Given we observed something for time t, how many events will we get to see, given α of them happen per unit time?

Solution. Let the lifespan of the universe, considering our observation starts at time 0 and goes till time t is a. Let the number of events to ever happen in the universe be n.

Notice, $\frac{n}{a} \to \alpha$ as $a, n \to \infty$. It is intuitive that the events are uniformly distributed in the universe, thanks to their memorylessness. So what we are asking is the limit of the distribution of some variable defined on the uniform process as $a, n \to \infty$.

Define $O_{a,n}(t)$ as the number of points in $[0, t]$ when n points are sampled from $[0, a]$ in an independent and uniform manner, and $\frac{n}{a} = \alpha$ and $t < a$.

$$\begin{aligned}\mathbb{P}(O_{a,n}(t) = k) &= \binom{n}{k}\left(\frac{t}{a}\right)^k\left(1 - \frac{t}{a}\right)^{n-k} \\ &= \frac{n!}{(n-k)!k!}\left(\frac{\alpha t}{n}\right)^k\left(1 - \frac{\alpha t}{n}\right)^{n-k} \\ &= \frac{n!}{(n-k)!n^k}\frac{(\alpha t)^k}{k!}\left(1 - \frac{\alpha t}{n}\right)^{n-k}\end{aligned}$$

As $n \to \infty$, $\left(1 - \frac{\alpha t}{n}\right)^{n-k} = \left(1 - \frac{\alpha t}{n}\right)^n\left(1 - \frac{\alpha t}{n}\right)^{-k} \to e^{-\alpha t} \cdot 1^{-k} = e^{-\alpha t}$ as k is a fixed constant.

Finally, we deal with

$$\begin{aligned}\frac{n!}{(n-k)!n^k} &= \frac{n(n-1)...(n-k+1)}{n^k} \\ &= \frac{n}{n}\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)...\left(1 - \frac{k-1}{n}\right)\end{aligned}$$

Which approaches 1 as $n \to \infty$ as every term approaches 1 and there are a finite number of terms.

Define $U_\alpha(t)$ as the number of points in $[0, t]$ when n points are sampled from $[0, a]$ in an independent and uniform manner, and $\frac{n}{a} = \alpha$ and $a, n \to \infty$.

Thus, $\mathbb{P}(U_\alpha(t) = k) = \frac{(\alpha t)^k}{k!}e^{-\alpha t}$ □

Often $\lambda := \alpha t$ is taken as the parameter where it represents the average number of events observed in time t.

Definition

An integer random variable X is said to have the **Poisson distribution** with parameter λ if

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

when $k \geq 0$ and 0 otherwise.

Remark

One can derive the Poisson distribution directly using the definition, considering the number of independent identical exponentially distributed events in time t. It is just a much less elegant method.

Corollary 2.6. *The mean and variance of Poisson distribution is λ*

Theorem 2.7. *Sum of two Poisson distributed variables with average a and b is a Poisson distributed variable with average $a + b$*

§3 Analysis in probability?!

I write this section with a lot of pain. While, I do like the methods and tricks analytic probability provides. However, the proofs, especially the measure theoretic ones, make one question why one does this.

§3.1 Randomized Exponential Processes

Consider a exponential process T with intensity A which is a positive continuos random variable. Using law of alternatives,

$$\begin{aligned}\mathbb{P}(T > t) &= \int_0^\infty \mathbb{P}(W_1 > t \mid A = \alpha) f(\alpha)\,\mathrm{d}\alpha \\ &= \int_0^\infty e^{-\alpha t} f(\alpha)\,\mathrm{d}\alpha \\ &= \mathbb{E}\left(e^{-At}\right)\end{aligned}$$

This is called the **Laplace Transform** of A or $\hat{f}(t)$ where f is the PDF of A. The Laplace Transform (and its perturbations) have a variety of uses, as we shall soon see.

Definition

For all these definitions, X is a random variable.

- Moment Generating function is $\mathbb{E}(e^{Xt})$ (defined on positive RVs)
- Characteristic function is $\mathbb{E}(e^{Xit})$
- Laplace Transform is $\mathbb{E}(e^{-Xt})$ (defined on positive RVs)
- Fourier Transform is $\mathbb{E}(e^{-Xit})$

Definition

For general functions $f(x)$, Laplace transform is

$$\hat{f}(t) = \int_0^\infty f(x) e^{-xt}\,\mathrm{d}x$$

Remark

Note that there are other definitions of the Fourier transform. In one definition, the expectation is multiplied by $\frac{1}{2\pi}$, and in another it is multiplied by $\frac{1}{\sqrt{2\pi}}$.

We can now prove one of the most important theorems with respect the transforms.

Theorem 3.1. *The transform of the convolution of functions is the product of their transforms i.e.*

$$\widehat{f * g} = \hat{f} \cdot \hat{g}$$

Proof. This proof being this short is another reason for using probabilistic reasoning over analytic bash.

We will only need to prove for one of the transforms here are all other are just perturbations of the Laplace transform.

Let A be a rv with PDF f and Bbe a rb with PDF g. As transforms are exponential processes, $\widehat{f * g}$ denotes the first waiting time for either of two exponential processes, with parameter A and B respectively, being observed. As they are independent,

$$\mathbb{P}(T_{A+B} > t) = \mathbb{P}(T_A > t)\mathbb{P}(T_B > t)$$
$$\Rightarrow \widehat{f * g} = \hat{f} \cdot \hat{g}$$

And we are done! □

§3.2 The Dreaded Proof

Theorem 3.2. *(Levy's Convergence Theorem) X_n, X be random variables in $\mathbb{R}$.*

$$F_{X_n}(x) \to F_X(x) \iff \varphi_{X_n}(t) \to \varphi_X(t) \quad \forall t \in \mathbb{R}$$

In proving this, we will use a bunch of theorems which we shall prove later.

Proof. ($\Longrightarrow$)Using the fact that the weak convergence of X_n implies that $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(x))$ for continuous increasing f and note, $x \mapsto e^{ixt}$ is continuos and increasing, we are done.

($\Longleftarrow$) I will write this later. Please send me a cute proof of this, if you find one!

I will otherwise use the one I have. □

§3.3 Renewal Process and some more properties of Characteristic function

Problem 3.3. How long it takes to cross the street. We will model this problem with a Poisson process with intensity α. Now one can cross the street only if there is a pause or gap in traffic that is long enough for one to cross. Let b be the size of the gap necessary for crossing the street safely.

What is the expected of waiting time before one can cross the street?

Solution

Let's say the variable W denotes our waiting time. Using law of alternatives on the first time we see a car, let is be T_1.

$$F_W(t) = \mathbb{P}(W \leq t) = \int_0^\infty \mathbb{P}(W \leq t \mid T_1 = u)\mathbb{P}(T_1 = u)\,\mathrm{d}u$$
$$= \int_0^\infty \mathbb{P}(W \leq t \mid T_1 = u)\alpha e^{-\alpha u}\,\mathrm{d}u$$

If $u \geq b$ then $\mathbb{P}(W \leq t) = 1$ as we can cross the road on the first try. Otherwise, $\mathbb{P}(W \leq t \mid T_1 = u) = \mathbb{P}(W \leq t - u)$. Note, $\mathbb{P}(W < k) = 0$ for all $k < 0$. Thus,

$$\begin{aligned} F_W(t) = \mathbb{P}(W \leq t) &= \int_0^\infty \mathbb{P}(W \leq t \mid T_1 = u)\alpha e^{-\alpha u}\,\mathrm{d}u \\ &= \int_0^b \mathbb{P}(W \leq t \mid T_1 = u)\alpha e^{-\alpha u}\,\mathrm{d}u + \int_b^\infty \alpha e^{-\alpha u}\,\mathrm{d}u \\ &= \int_0^b \mathbb{P}(W \leq t - u)\alpha e^{-\alpha u}\,\mathrm{d}u + e^{-\alpha b} \end{aligned}$$

This is clearly a convolution between F_W and

$$G(t) = \begin{cases} \alpha e^{-\alpha u} & 0 \leq u \leq b \\ 0 & \text{otherwise} \end{cases}$$

Thus,

$$F_W(t) = (F_W * G)(t) + e^{-\alpha b}$$

Taking the Laplace transform,

$$\widehat{F_W}(t) = \widehat{F_W * G}(t) + \frac{e^{-\alpha b}}{t}$$

The latter part comes from the fact that the Laplace transform(wrt to t) for a constant c is $\frac{c}{t}$. Thus,

$$\begin{aligned} \widehat{F_W}(t) &= \widehat{F_W}(t)\hat{G}(t) + \frac{e^{-\alpha b}}{t} \\ \Rightarrow \widehat{F_W}(t) &= \frac{e^{-\alpha b}}{\big(1 - \hat{G}(t)\big)t} \end{aligned}$$

Now,

$$\begin{aligned} \hat{G}(t) &= \int_0^\infty e^{-xt} G(x)\,\mathrm{d}x \\ &= \int_0^b e^{-xt}\alpha e^{-\alpha x}\,\mathrm{d}x \\ &= \alpha \int_0^b e^{-x(\alpha+t)}\,\mathrm{d}x \\ &= \frac{\alpha}{\alpha + t}\big(1 - e^{-b(\alpha+t)}\big) \end{aligned}$$

We will now prove some more theorems regarding characteristic functions, before going ahead.

Theorem 3.4. *Given $\hat{F}(s)$ is the Laplace transform of the CDF of some positive random variable X, then $t\hat{F}(s)$ is the Laplace transform of PDF of X.*

Proof.

$$\begin{aligned} s\hat{F}(s) &= s\int_0^\infty F(t)e^{-st}\,\mathrm{d}t \\ &= \int_0^\infty sF(t)e^{-st}\,\mathrm{d}t \end{aligned}$$

Using integration by parts on $F(t)$ and se^{-st}

$$\begin{aligned} s\hat{F}(s) &= [-F(t)e^{-st}]_0^\infty + \int_0^\infty F'(t)e^{-st}\,\mathrm{d}t \\ &= (0 + F(0)) + \int_0^\infty f(t)e^{-st}\,\mathrm{d}t \\ &= F(0) + \hat{f}(s) \end{aligned}$$

As X is a positive random variable, $F(0) = 0$. Thus, $t\hat{F}(s) = \hat{f}(s)$ □

Laplace Transform Moment Formula 3.5. *If $\hat{f}(t)$ is the Laplace transform of the density of a positive random variable,*

$$\mathbb{E}(X^n) = (-1)^n \left[\frac{\mathrm{d}^n}{\mathrm{d}t^n}\hat{f}(t)\right]_{t=0}$$

The proof follows from direct computation.

With these, we now complete our solution to the road crossing problem.

Solution

Using Theorem 3.4 and Theorem 3.5

$$\begin{aligned} \mathbb{E}(W) &= -\hat{f}'(0) \\ &= -\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(x\hat{F}(x)\right)\right]_{x=0} \\ &= -\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(x \cdot \frac{e^{-\alpha b}}{\left(1-\hat{G}(x)\right)x}\right)\right]_{x=0} \\ &= -e^{-\alpha b}\left[\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{1}{1-\hat{G}(x)}\right)\right]_{x=0} \\ &= -e^{-\alpha b}\left[\left(1-\hat{G}(x)\right)^{-2}\hat{G}'(x)\right]_{x=0} \end{aligned}$$

Now, note $\hat{G}(t) = \frac{\alpha}{\alpha+t}\left(1-e^{-b(\alpha+t)}\right) \Rightarrow \hat{G}(0) = 1 - e^{-\alpha b}$ and $\hat{G}'(0) = \frac{e^{-\alpha b}(1+\alpha b - e^{\alpha b})}{\alpha}$.

Thus,

$$\begin{aligned}\mathbb{E}(W) &= -e^{-\alpha b}\left[\left(1-\hat{G}(x)\right)^{-2}\hat{G}'(x)\right]_{x=0} \\ &= -e^{-\alpha b}\cdot\frac{1}{e^{-2\alpha b}}\cdot\frac{e^{-\alpha b}\left(1+\alpha b-e^{\alpha b}\right)}{\alpha} \\ &= \frac{e^{-\alpha b}-\alpha b-1}{\alpha}\end{aligned}$$

Remark

One can also compute variance of W and find it to be $\frac{e^{2\alpha b}-1-2\alpha b e^{\alpha b}}{\alpha^2}$. Try plugging some normal values for α and b to see a surprising(and predictive) consequence.

§4 Towards Normal Distribution

Definition

A random variable X with mean μ and variance σ^2 is said to have the normal distribution if

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{-\left(\frac{(x-\mu)^2}{2\sigma^2}\right)} = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

Remark

We will abstain from calling the Normal Distribution the Gaussian distribution. The reason is simply the fact that this is one of the few rare exceptions in probability where a distribution isn't named after the person who introduced it. If any name were to be attached, it would make far more historical sense to call name it after Abraham de Moivre who first derived it in the context of the binomial distribution, long before Gauss ever came near it.

Sure, de Moivre didn't have the full modern idea of probability or a density function, but that hasn't stopped us before—Zipf was a linguist, and he still got a distribution named after him.

In the spirit of historical fidelity, I might be tempted to call it the Moivrian distribution. But in the interest of not swimming too hard against the current, I'll settle for normal,

We will also state without proof the following

Theorem 4.1.

$$\int_{-\infty}^{\infty} e^{-x^2}\,\mathrm{d}x = \sqrt{\pi}$$

The fact that we are studying such a weird, alien and (dare I say) abnormal distribution raises some questions regarding the point of the subject. We shall take a small, completely optional, highly recommended detour to information theory.

§4.1 The Detour

Information theory deals with the idea of surprise. When something we expect will happen, happens, we are neither very surprised nor do we gain too much information. However, when something we don't expect occurs, we are surprised and learn more about how the world works. We will begin with the discrete entropy and move to continuos.

Let's say $I(p)$ is the information we gain from an event which occurs with probability p. Some natural conditions are

- $I(p)$ is monotonically decreasing in p.
- I is continuos as an event with likeliness p and $p+\varepsilon$ can't somehow have magnitudes of difference in information conveyed for small ε, i.e. the definition of continuity.
- $I(1) = 0$ as events that always occur do not communicate information.
- $I(p_1 \cdot p_2) = I(p_1) + I(p_2)$ which says that the information learned from independent events is the sum of the information learned from each event.

Theorem 4.2. *The only function satisfying these conditions is* $I(p) = \log\left(\frac{1}{p}\right) = -\log(p)$ *where the base of logarithm can be any positive real.*

Proof. We begin by defining $f(x) = I(e^x)$. This implies

$$f(x+y) = I(e^{x+y}) = I(e^x e^y) = I(e^x) + I(e^y) = f(x) + f(y)$$

This is the Cauchy functional equation which has only continuous solutions of the form $f(x) = kx$ where k is real.

This implies $f(x) = I(e^x) = kx \Rightarrow f(\ln x) = I(x) = k \ln x$.

As $I(x)$ is monotonically decreasing $k < 0$. And as we can convert the positive part to 1 by changing the base of the log. Thus, $I(p) = -\log(p)$ for some base of the logarithm. □

We define entropy as the expected surprise/information gained as the result of some experiment. For an experiment with n possible outcomes, It is denoted as

$$H(A) = \mathbb{E}(I(A)) = \sum_i \mathbb{P}(A_i) I(A_i) = -\sum_{i=1}^{n} \mathbb{P}(A_i) \log(\mathbb{P}(A_i))$$

Where we take $0 \log(0) = 0$ as 0 by convention. This is justified as the probability event that never occurs should not contribute to entropy.

This all seems nice, but extending this to a continuos case is rather non-trivial.

> 💡 **Idea**
>
> What about simply extending the definition to continuous like we do for expectation(actually we don't but...) or variance? As $\mathbb{P}(X = x) \to f(x)\,\mathrm{d}x$ and let S_x be the state space of X.
>
> $$\Rightarrow H(X) = -\int_{S_X} f(x)\,\mathrm{d}x \log(f(x)\,\mathrm{d}x)$$
>
> $$= -\int_{S_X} f(x) \log(f(x))\,\mathrm{d}x - \int_{S_X} f(x) \log(\mathrm{d}x)\,\mathrm{d}x$$
>
> Unfortunately, $\log(\mathrm{d}x)$ explodes and our result basically loses all semblance of meaning.

So what do we do? Let's define a few new things.

Definition

Cross Entropy is a function which takes two probability distributions A and B and tells us our expected surprise observing a random variable with actual distribution A and while believing its distribution is B. It has the formula

$$H(A,B) = -\sum_s \mathbb{P}(A=s)\log(\mathbb{P}(B=s))$$

Corollary 4.3. $H(A,A) = H(A)$

Corollary 4.4. $H(A,B) \geq H(A)$

The corollaries are simply justified by the fact that when our belief about a model doesn't match reality, their are two sources of surprise. The first being due to the inherent uncertainty and the other due to our belief in the wrong model.

Definition

Kullback-Leibler Divergence or **KL Divergence** is the measure of extra surprise we get due to believing in the wrong model. It is defined as

$$\begin{aligned} D_{KL}(A,B) &= H(A,B) - H(A) \\ &= \sum_s \mathbb{P}(A=s)\log\left(\frac{\mathbb{P}(A=s)}{\mathbb{P}(B=s)}\right) \end{aligned}$$

Another interpretation of D_{KL} is how much more surprising is B in comparison to A.

One may ask why would we want to measure this? Well, as this is already a detour, I will not go in too much detail but the point is to allow us to model complex distributions. A key idea in machine is making programmes to choose B to minimize $D_{KL}(A,B) = H(A,B) - H(A)$ where A is the distribution of some parameters wrt to the training data set.

There is also KL minimizing regression as well as uses in most likely estimator statistics. We will not explore them here.

What is the point of all this then?

Idea contd.

We return to our task of extending the definition of entropy to continuous case.

While we can't really talk about the entropy of some continuous random variable X as we saw above, can we make a comparison which does a good enough job? The uniform distribution seems like a good candidate. Let U be a uniformly distributed variable in the same state space as X, say $[0, a]$.

$$\begin{aligned} D_{KL}(X, U) &= \int_0^a f(x)\,\mathrm{d}x \log\left(\frac{f(x)\,\mathrm{d}x}{\left(\frac{1}{a}\right)\mathrm{d}x}\right) \\ &= \int_0^a f(x)\log(f(x))\,\mathrm{d}x + \log(a) \end{aligned}$$

Hence, we can effectively take $H(X) = -D_{KL}(X, U) = -\int_{-\infty}^{\infty} f(x)\log(f(x))\,\mathrm{d}x$ where $H(X)$ here represents(notice the change of signs and limits) the how much less surprising/random is X in comparison to the uniform distribution/complete randomness. For more context check this out.[3]

Finally, for this notion of entropy, the case of total randomness will be represented by an entropy of 0. More uncertain random variables can have a negative relative entropy. Continuous random variables can have arbitrarily large negative entropy.

§4.2 The Natural-ity of Normality

Using information theory, we will now show that normal distribution is indeed natural as it is the 'entropy maximizing distribution' for some conditions.

Claim 4.5. $H(cX) = H(X) + \log(c)$ for $c \in \mathbb{R}^+$

Proof. Using the change of variables formula,

$$\begin{aligned} f_{cX}(x) &= \frac{f_X\left(\frac{x}{c}\right)}{c} \\ \Rightarrow H(cX) &= -\int_{-\infty}^{\infty} f_{cX}(x)\log(f_{cX}(x))\,\mathrm{d}x \\ &= -\frac{1}{c}\int_{-\infty}^{\infty} f_X\left(\frac{x}{c}\right)\log\left(\frac{f_X\left(\frac{x}{c}\right)}{c}\right)\mathrm{d}x \\ &= -\frac{1}{c}\int_{-\infty}^{\infty} f_X\left(\frac{x}{c}\right)\log\left(f_X\left(\frac{x}{c}\right)\right)\mathrm{d}x + \frac{1}{c}\log(c)\int_{-\infty}^{\infty} f_X\left(\frac{x}{c}\right)\mathrm{d}x \\ &= -\int_{-\infty}^{\infty} f_X(y)\log(f_X(y))\,\mathrm{d}y + \log(c)\int_{-\infty}^{\infty} f_X(y)\,\mathrm{d}y \end{aligned}$$

making the substitution $y = \frac{x}{c}$.

$\therefore H(cX) = H(X) + \log(c)$ □

[3] Introduction to Continuous Entropy, Charles Marsh

It is also easy to observe that entropy is not position variant as $H(X+k) = H(X)$ is not generally true. So

This leads us to define $SH(X) = H\left(\frac{X-\mu}{\sigma}\right)$ or the the standard entropy of X is the entropy of standardization of X.

Theorem 4.6. *Standard entropy is maximized for a normally distributed variable.*

Proof. A standard method to solve optimization problems with constraints is **Lagrange multipliers**. We will use them without proof here.[4]

We wish to maximize

$$-\int_{-\infty}^{\infty} f(x)\log(f(x))\,\mathrm{d}x$$

subject to

$$\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x = 1$$
$$\int_{-\infty}^{\infty} xf(x)\,\mathrm{d}x = 0$$
$$\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x = 1$$

Using α, β as our multipliers(we don't need three as the second equation is 0), we define our Lagrangian as

$$\begin{aligned}\mathcal{L} &= -\int_{-\infty}^{\infty} f(x)\log(f(x))\,\mathrm{d}x + \alpha\left(\int_{-\infty}^{\infty} f(x)\,\mathrm{d}x\right) + \beta\left(\int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x\right)\\ &= \int_{-\infty}^{\infty} f(x)(-\log(f(x)) + \alpha + \beta x^2)\,\mathrm{d}x\\ &= \int_{-\infty}^{\infty} f(x)\log\left(\frac{e^{\alpha+\beta x^2}}{f(x)}\right)\mathrm{d}x\end{aligned}$$

Using $1 + \log(x) \leq x$ for all $x \in \mathbb{R}^+$ with equality at only $x = 1$,

$$\begin{aligned}\mathcal{L} &= \int_{-\infty}^{\infty} f(x)\log\left(\frac{e^{\alpha+\beta x^2}}{f(x)}\right)\mathrm{d}x\\ &\leq \int_{-\infty}^{\infty} f(x)\left(\frac{e^{\alpha+\beta x^2}}{f(x)} - 1\right)\mathrm{d}x\\ &= \int_{-\infty}^{\infty} e^{\alpha+\beta x^2}\,\mathrm{d}x - 1\end{aligned}$$

As the final equation has no variables, this is the maximum value of $\mathcal{L}$ and hence,

[4] I belive they were supposed to be done by a certain someone in a certain class.

$$\log\left(\frac{e^{\alpha+\beta x^2}}{f(x)}\right) = 1$$
$$\Rightarrow f(x) = e^{-\alpha-\beta x^2}$$

maximized $\mathcal{L}$ and by effect $H(X)$. We just need to find the multipliers wrt constraints. Notice,

$$\begin{aligned}1 &= \int_{-\infty}^{\infty} f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} e^{-\alpha-\beta x^2}\,\mathrm{d}x \\ &= e^{-\alpha}\sqrt{\frac{\pi}{\beta}}\end{aligned}$$

and

$$\begin{aligned}1 = \int_{-\infty}^{\infty} x^2 f(x)\,\mathrm{d}x \\ &= \int_{-\infty}^{\infty} x^2 e^{-\alpha-\beta x^2}\,\mathrm{d}x \\ &= \frac{1}{2\beta}e^{-\alpha}\frac{\pi}{\beta}\end{aligned}$$

Thus,

$$f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$$

is the entropy maximizing distribution.

Notice, this is the standardized normal distribution. Thus, Standard entropy is maximized for a normally distributed variable. And we are done! □

This tells us that the universal fact of maximization of entropy pushes events towards the normal distribution. Another reason for this is the famed **Central Limit Theorem** .

Theorem 4.7. *For iid's $X_1, X_2, ..., X_n$ with mean 0 and variance σ^2,*

$$\lim_{n\to\infty}\left(\frac{X_1 + X_2 + ... + X_n}{\sqrt{n}}\right) \sim \mathcal{N}(0, \sigma^2)$$

where $\mathcal{N}(\mu, \sigma^2)$ is a variable normally distributed with mean μ and variance σ^2.

Inoformation Theoretic Proof. [5]

[5] What? Did you think I brought in Information theory just for one thing?

Definition

Fisher Information is a measure of information gained about the parameter(in our case, it is the location parameter or mean[6]) of a random variable from observations of it. It is defined as the variance of score function $\rho_{U(u)} = \frac{f_U{}'(u)}{f_U(u)} = \frac{\mathrm{d}}{\mathrm{d}u}\log(f_U(u))$ and the Fisher information is

$$J(X) = \mathbb{E}(\rho_U^2(U)) = \int \frac{1}{f(u)}\left(\frac{\mathrm{d}}{\mathrm{d}u}p(u)\right)^2 \mathrm{d}u$$

The idea is that we measure how high the peaks are in the log-likelihood function - that is, how much small changes in the parameters affect the likelihood. For more intuition this video out.

As part of our proof, we will use one of the most important results in probability(and perhaps the 20th century), I will outline it for the interested.

Lemma 4.8. *Given a random variable with a suitably well behaved smooth density f and score function ρ, $\mathbb{E}(f(X)\rho(X)) = -\mathbb{E}(f'(X))$*

Hint: Use integration by parts.

Stein's Lemma 4.9. *Given $X \sim \mathcal{N}(\mu, \sigma^2)$ and a well behaved continuous g for which $\mathbb{E}(g(X)(X-\mu))$ and $\mathbb{E}(g'(X))$ exist, then*

$$\mathbb{E}(g(X)(X-\mu)) = \sigma^2\mathbb{E}(g'(X))$$

Hint: Use integration by parts

Theorem(Crammer-Rao Lower Bound) 4.10. *Given a random variable U with mean μ and variance σ^2,*

$$J(U) \geq \frac{1}{\sigma^2}$$

with equality if and only if $U \sim \mathcal{N}(\mu, \sigma^2)$

This result is what is believed to have put ISI on the world map with the Rao referring to the legendary late C.R. Rao. Anyways, continuing onwards.

Lemma 4.11. *If U, V are independent random variables and $W = U + V$ with score functions ρ_U, ρ_V, ρ_W then*

$$\rho_W(w) = \mathbb{E}(\rho_U(U) \mid W = w) = \mathbb{E}(\rho_V(V) \mid W = w)$$

Proof. Let U, V have density u, v. Then the density w of $W = U + V$ is $w(x) = \int u(y)v(x-y)\,\mathrm{d}y$ then

$$w'(x) = \int u(y)\frac{\partial}{\partial x}v(x-y)\,\mathrm{d}y$$

[6]The complete definition and formula can deal with a wider range of a parameters. What we have defined here is technically called **Fisher information with respect to location parameter** to distinguish it from the more general one.

To use integration by parts, we need to transform ∂x into ∂y. For that notice, $z = x - y \Rightarrow \frac{\partial z}{\partial x} = 1$ and $y = x - z \Rightarrow 1 = -\frac{\partial z}{\partial y} \Rightarrow \frac{\partial (y)}{\partial (x)} = -1$, hence

$$w'(x) = -\int u(y)\frac{\partial}{\partial y}v(x-y)\,\mathrm{d}y = \int u'(y)v(x-y)\,\mathrm{d}y$$

The final equality follows from integration by parts. Thus,

$$\frac{w'(x)}{w(x)} = \int \frac{u'(y)v(x-y)}{w(x)}\,\mathrm{d}y = \int \frac{u'(y)}{u(y)}\frac{u(y)v(x-y)}{w(x)}\,\mathrm{d}y = \mathbb{E}(\rho_U(U) \mid W = w)$$

Similarly for V. And we are done. □

Lemma 4.12. *If U, V are independent then for any $\beta \in [0,1]$*

(1) $J(U+V) \leq \beta^2 J(U) + (1-\beta)^2 J(V)$

(2) $J\left(\sqrt{\beta}U + \sqrt{1-\beta}V\right) \leq \beta J(U) + (1-\beta)J(V)$ with equality if U and V are gaussian

Proof. Using Lemma 4.11, $\rho_W(w) = \mathbb{E}(\beta\rho_U(U) + (1-\beta)\rho_V(V) \mid W)$, then by Jensen

$$J(U+V) \leq \beta^2 J(U) + (1-\beta)^2 J(V)$$

substituting $\sqrt{\beta}U$ and $\sqrt{1-\beta}V$ for U and V respectively, we recover the second result.

As we have used Jensen, the equality will occur if $\rho_U(x)$ is linear in x for all $x \in \mathbb{R}$. Thus,

$$\frac{\mathrm{d}}{\mathrm{d}x}\log(f_U(x)) = ax + b$$

$$\int \frac{\mathrm{d}}{\mathrm{d}x}\log(f_U(x))\,\mathrm{d}x = \int ax + b\,\mathrm{d}x$$

$$\log(f_U(x)) = \frac{a}{2}x^2 + bx + C$$

$$f_{U(x)} = e^{\frac{a}{2}x^2+bx+C} = e^C e^{\frac{a}{2}x^2+bx} = Ae^{\frac{a}{2}x^2+bx}$$

It is easy to show that f_U normalizes to a normal distribution. Thus, the equality condition is also proved. □

Finally, we are now ready to prove the cental limit theorem.

We first notice $J\left(X_1 + \frac{X_2}{\sqrt{2}}\right) < J(U)$ by Lemma 4.12, given X is not already normal.

As $J(U)$ is bounded below by Theorem 4.10(CRLB) and thus by monotone bounded $J\left(\frac{X_1+\ldots+X_n}{\sqrt{n}}\right)$ converges to $\frac{1}{\sigma^2}$ which can only occur if $\frac{X_1+\ldots+X_n}{\sqrt{n}} \sim \mathcal{N}(0,\sigma^2)$

And we are done! □

And obviously, for the sake of completeness, we will present the classical 'analytical' proof for the same. Note, I will not be proving Levy Continuity or the uniqueness of characteristic function or as part of proof, but for any comparisons regarding the length of proofs, it must be noted that the last proof is very much self contained.

Analytical proof. WLOG, let X be standard and Let $S_n = X_1 + X_2 + \ldots + X_n$ and we will show $\frac{S_n}{\sqrt{n}}$ converges to a normal distribution. If $\varphi(t)$ is the characteristic function of X, then

$$\varphi(0) = E(X^0) = 1$$
$$\varphi'(0) = iE(X^1) = 0$$
$$\varphi''(0) = -E(X^2) = -1$$

Thus, by Taylor Expansion

$$\begin{aligned}\varphi(t) &= \varphi(0) + \varphi'(0)t + \frac{\varphi''(0)}{2}t^2 + O(t^3) \\ &= 1 - \frac{t^2}{2} + O(t^3)\end{aligned}$$

Using the convolution theorem and a change of variables, S_n has characteristic function $\varphi\left(\frac{s}{\sqrt{n}}\right)^n$, thus

$$\begin{aligned}\varphi\left(\frac{s}{\sqrt{n}}\right)^n &= \left(1 - \frac{1}{2}\left(\frac{s}{\sqrt{n}}\right)^2 + O\left(\left(\frac{s}{\sqrt{n}}\right)^3\right)\right)^n \\ &= \left(1 - \frac{1}{2}\frac{s^2}{n} + O\left(\frac{s^3}{n\sqrt{n}}\right)\right)^n\end{aligned}$$

Taking limit as n goes to ∞,

$$\begin{aligned}\lim_{n\to\infty} \varphi\left(\frac{s}{\sqrt{n}}\right)^n &= \lim_{n\to\infty} \left(1 - \frac{1}{2}\frac{s^2}{n} + O\left(\frac{s^3}{n\sqrt{n}}\right)\right)^n \\ &= \lim_{n\to\infty} \left(1 - \frac{1}{2}\frac{s^2}{n}\right)^n \\ &= e^{-\frac{s^2}{2}}\end{aligned}$$

which is the characteristic function of $\mathcal{N}(0,1)$. As different probability distributions have different characteristic functions and convergence in characteristic functions is also convergence in distributions, $\frac{S_n}{\sqrt{n}} \to \mathcal{N}(0,1)$ as desired. □

In general use, we are much better off using an equivalent but often more useful form of CLT.

Theorem 4.13. *For iid's $X_1, X_2, ..., X_n$ with mean μ and variance σ^2,*

$$\lim_{n\to\infty}\left(\frac{X_1 + X_2 + ... + X_n}{n}\right) \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\mathcal{N}(\mu, \sigma^2)$ is a variable normally distributed with mean μ and variance σ^2.

We will present the following useful facts about normal distribution here.

Theorem 4.14. *If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ then $X + Y \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$*

Corollary 4.15. *If $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ then $X - Y \sim \mathcal{N}(\mu_1 - \mu_2, \sigma_1^2 + \sigma_2^2)$*

Theorem 4.16. *If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $X + c \sim \mathcal{N}(\mu + t, \sigma^2)$ where c is a real constant.*

Theorem 4.17. *If* $X \sim \mathcal{N}(\mu, \sigma^2)$ *then* $cX \sim \mathcal{N}(c\mu, c^2\sigma^2)$

§5 Exercises

Please note, these are a bunch of cute things I found here and there. The questions, while I belive are nice, are not meant to be indicative or predictive in any way.

Exercise 5.1. Cantelli's inequality states that for a random variable X with mean μ and variance σ^2,

$$\mathbb{P}(X \geq \mu + k\sigma) \leq \frac{1}{k^2+1}$$

and/or

$$\mathbb{P}(X \leq \mu - k\sigma) \leq \frac{1}{k^2+1}$$

or probability mass higher than k standard deviations from the mean is at most $\frac{1}{1+k^2}$.

Prove Cantelli's inequality.

Exercise 5.2. Chebyshev's inequality states that for a random variable X with mean μ and variance σ^2,

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

or probability mass higher than k standard deviations from the mean is at most $\frac{1}{k^2}$.

Prove Chebyshev's inequality.

Exercise 5.3. How many times must one toss a fair coin in order to have 95% confidence that it really is fair? What does Cantelli, Chebyshev and CLT each say?

> ***i*** **Remark**
>
> Can you think why both Cantelli and Chebyshev are useful? When should which one be used?

Exercise 5.4. Markov's Inequality states If X is a nonnegative random variable and $a > 0$, then the probability that X is at least a is at most the expectation of X divided by a

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

Exercise 5.5. There's a line of ten people waiting to order at a café, when the proprietor says,

"The first person in line, whose birthday this year is on the same weekday as anyone in front of them, will get a free muffin with their order."

You're standing in eighth place. What's the probability that you get a free muffin, and which person should you ask to swap with, to maximize your probability? What would your chances be if you manage to swap?

Exercise 5.6. There exists a coin(which may be biased for any $p \in [0, 1]$ with uniform probability) which has landed k heads. What is the probability that the next flip will be a head?[We claim the answer is $\frac{k+1}{k+2}$.]

Exercise 5.7. Laplace's rule of succession says that if we have observed S successes over T trials we should estimate the probability of success in the next trial as $\frac{S+1}{T+2}$. Prove Laplace's rule of succession.

Exercise 5.8. Let's suppose we are studying the likelihood of an earthquake. We have observed the seismograph of the Example-topia region for a decade, and observed no earthquakes so far. The guild of architects of Example-topia wants to know if the good luck will continue onto the next decade.

Alpha, the chief seismologist wastes no time in answering. We have seen no earthquakes over a decade. Using Laplace, What is the likelihood of no earthquakes happening in the next decade?

What is Alpha had reported this information is pieces, for example no earthquakes in first 5 years and no earthquakes in second 5 years? Can you see a problem?

> ***i*** **Remark**
>
> A method to fix this is by defining a continuous variant of Laplace's rule by replacing the binomial by Poisson. The prior distribution for λ of the Poisson is taken improperly as $f(\lambda) = \frac{1}{\lambda}$. The exact details for this can be found at here.

Exercise 5.9. Show that the Cauchy distribution has the property that the sample mean of a sequence of independent, equidistributed Cauchy random variables has the same distribution as a single random variable in the sequence. Does this affect the convergence of Cauchy under CLT?

Exercise 5.10. In a telephone network, it is known that phone calls are made with a Poisson distribution having intensity α. It is also found that the length of time of any given phone conversation is exponentially distributed with parameter β. The phone company charges customers $c + kt$ dollars for a phone call lasting t units of time, where c and k are constants. What is expected amount of money the company receive during a billing period of length T?

Exercise 5.11. Let $Y_1, Y_2, ..., Y_n$ be n independent, exponentially distributed random variables, each of which has intensity α. Compute the order statistics $Y_{(1)}, Y_{(2)}, ..., Y_{(n)}$ of these random variables.

Compute $\mathbb{E}\left(Y_{(i)}\right)$

Exercise 5.12. Three persons, A, B and C, arrive at a post office simultaneously. There are two counters, and these are taken immediately by A and B. Assume the service time of a given individual is exponentially distributed with rate parameter α. Assume also that different individuals

have independent service times. Now C will take whichever of the two counters becomes free first. Answer the following questions:

1. What is the probability that C is the last to leave the post office?
2. What is the distribution of the total time (i.e., both waiting time and service time) spent by C in the post office? How long does C spend on average?
3. When does the last of the three persons leave the post office? When on average?

Exercise 5.13. In a college cafeteria, ice cream is available for the evening meal in servings that vary in weight according to a normal distribution with a standard deviation of 10 gm. The cafeteria workers maintain about 15 servings for students to choose from. Every day student A chooses the smallest serving available, while student D chooses the largest. Over the school year (200 meals), how much more ice cream does student D eat than student A?

Exercise 5.14. During World War II, Allied forces captured or observed a number of German tanks with sequentially marked serial numbers. Intelligence analysts sought to estimate the total production of these tanks based on the serial numbers of the captured or observed samples.

Formally, Given $\{x_1, x_2, ..., x_i\}$ sampled without replacement from a finite population $1, 2, ..., N$, where N is unknown; estimate N. Why is this a good estimate?

That's all for now! I will add some more problems as and when I find cute problems.