Math 3332: Probability and Inference

Unit III: Continuous Random Variables

Mikhail Lavrov (mlavrov@kennesaw.edu)

Spring 2021

Our first continuous distribution

Example. A random variable $\mathbf{X}$ has the *Uniform*(a, b) distribution if it is equally likely to be any real number in the interval $[a, b]$.

We already know some things about this case. What we know:

- For any specific x, $\Pr[\mathbf{X} = x]$ is 0.
- For $x_1, x_2 \in [a, b]$, $\Pr[x_1 \leq \mathbf{X} \leq x_2] = \frac{x_2 - x_1}{b - a}$.
- We can guess that $\mathbb{E}[\mathbf{X}] = \frac{a+b}{2}$ by symmetry.

If we tried to define $\mathbf{X}$ by a PMF, it wouldn't work! The PMF would be 0 everywhere.

Cumulative distribution function

What is the minimum amount of information we need to describe the distribution of $\mathbf{X}$?

- Specifying $\Pr[x_1 \leq \mathbf{X} \leq x_2]$ for all $x_1 < x_2$ is enough.
- If we specified $\Pr[\mathbf{X} \leq t]$ for all $t \in \mathbb{R}$, that would also be enough!

  We can write $\Pr[x_1 \leq \mathbf{X} \leq x_2] = \Pr[\mathbf{X} \leq x_2] - \Pr[\mathbf{X} \leq x_1]$.

We define the **cumulative distribution function** or CDF of a random variable $\mathbf{X}$ to be the function $F_{\mathbf{X}}(t) = \Pr[\mathbf{X} \leq t]$.

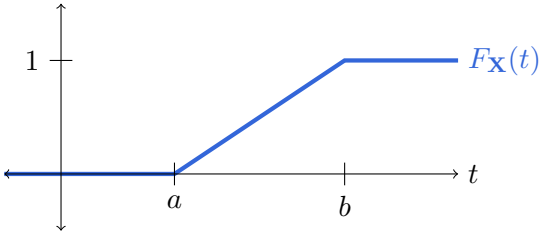
A random variable is **continuous** if the CDF is a continuous function $F_{\mathbf{X}} : \mathbb{R} \to [0, 1]$.

CDF of a Uniform distribution

Example: $\mathbf{X} \sim \textit{Uniform}(a, b)$ if it has CDF

$$F_{\mathbf{X}}(t) = \begin{cases} 0 & t < a \\ \frac{t-a}{b-a} & a \leq t \leq b \\ 1 & t > 0. \end{cases}$$

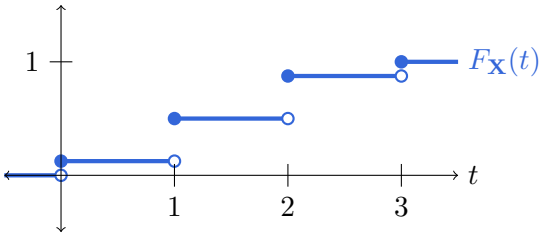
Here is a plot:



CDF of a discrete random variable

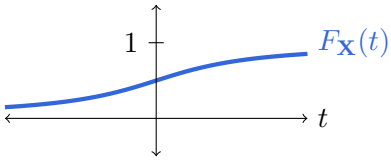
Discrete random variables can also be defined by a CDF, it's just not as common (since we can use the PMF instead). The function $F_{\mathbf{X}}(t) = \Pr[\mathbf{X} \leq t]$ still exists.

For example, here is a plot of a CDF of a random variable $\mathbf{X}$ with the *Binomial*$(3, \frac{1}{2})$ distribution:



The Cauchy distribution

A standard Cauchy distribution has CDF $F_{\mathbf{X}}(t) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}(t)$. (Don't worry about learning this distribution.) Here is a plot:



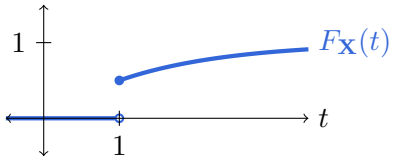
What can we deduce from this CDF?

- The CDF is continuous, so we have a continuous distribution.
- It only approaches, and never reaches 0 on the left or 1 on the right. So all real numbers must be in the range of $\mathbf{X}$.
- $\Pr[0 \leq \mathbf{X} \leq 1] = (\frac{1}{2} + \frac{1}{\pi}\tan^{-1}(1)) - (\frac{1}{2} + \frac{1}{\pi}\tan^{-1}(0)) = \frac{1}{4}$.

A jump in the CDF

Here is a CDF and plot of another distribution:

$$F_{\mathbf{X}}(t) = \begin{cases} 0 & t < 1 \\ 1 - 2^{-t} & t \geq 1 \end{cases}$$



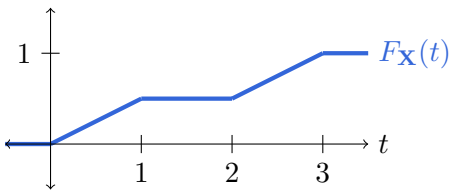
What can we deduce from this CDF?

- The distribution of $\mathbf{X}$ is not continuous: there is a jump in the CDF. (But not discrete, either!)
- $F_{\mathbf{X}}(t) = 0$ when $t < 1$, but it's never 1: the range of $\mathbf{X}$ is $[1, \infty)$.
- $\Pr[\mathbf{X} \leq 1] = \frac{1}{2}$, but for all $t < 1$, we get $\Pr[\mathbf{X} \leq t] = 0$. Therefore $\Pr[\mathbf{X} = 1] = \frac{1}{2}$.

Writing down a CDF

If $\mathbf{X}$ is uniformly sampled from $[0,1] \cup [2,3]$, what is the CDF of $\mathbf{X}$?

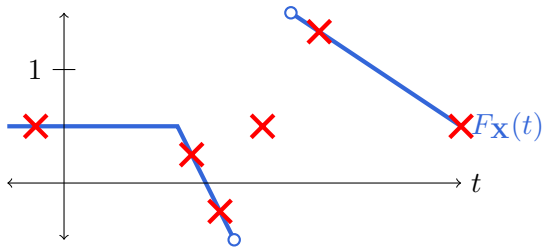
$$F_{\mathbf{X}}(t) = \begin{cases} 0 & t < 0 \\ \frac{1}{2}t & 0 \le t < 1 \\ \frac{1}{2} & 1 \le t < 2 \\ \frac{1}{2}t - \frac{1}{2} & 2 \le t < 3 \\ 1 & t \ge 3 \end{cases}$$



For $t < 0$, $\Pr[\mathbf{X} \le t] = 0$. Then, $F_{\mathbf{X}}(t)$ must increase at a constant rate from $(0,0)$ to $(1, \frac{1}{2})$, so $F_{\mathbf{X}}(t) = \frac{1}{2}t$ on that interval. $F_{\mathbf{X}}(t)$ is constant on $[1,2]$, since there is no chance of $\mathbf{X}$ landing there. On $[2,3]$, $F_{\mathbf{X}}(t)$ must increase at a constant rate again. Finally, for $t \ge 3$, $F_{\mathbf{X}}(t)$ stays at 1.

This is not a CDF

There are some restrictions on what a CDF can do. This function is not a CDF: it breaks all of them!



As $t \to -\infty$, $F_{\mathbf{X}}(t)$ must either reach or approach 0. The CDF cannot decrease or be negative. It must be defined everywhere! It cannot exceed 1 and as $t \to +\infty$, $F_{\mathbf{X}}(t)$ must either reach or approach 1.

The PDF of a random variable

There are two ways to describe a continuous distribution:

- the CDF (which we know about): a function $F_{\mathbf{X}}$ such that
$$\Pr[a \leq \mathbf{X} \leq b] = F_{\mathbf{X}}(b) - F_{\mathbf{X}}(a).$$
- the PDF (**probability density function**): a function $f_{\mathbf{X}}$ such that
$$\Pr[a \leq \mathbf{X} \leq b] = \int_a^b f_{\mathbf{X}}(t)\, dt.$$

This is a continuous version of the PMF: for discrete random variables,

$$\Pr[a \leq \mathbf{X} \leq b] = \sum_{k=a}^{b} P_{\mathbf{X}}(k).$$

Many other properties of the PMF also hold for the PDF.

When do PDFs exist?

All random variables have CDFs. Even discrete ones; for example, a *Bernoulli*(p) distribution has CDF

$$F(t) = \begin{cases} 0 & t < 0 \\ 1 - p & 0 \leq t < 1 \\ 1 & t = 1. \end{cases}$$

Only continuous random variables have PDFs. If $f_{\mathbf{X}}(t)$ exists, then

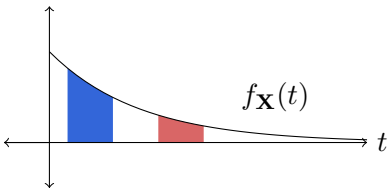
$$F_{\mathbf{X}}(t) = \Pr[\mathbf{X} \leq t] = \int_{-\infty}^{t} f_{\mathbf{X}}(t)\, dt$$

which is always continuous.

Intuition for PDFs

The integral of $f_{\mathbf{X}}(t)$ from a to b is the area under the graph of $f_{\mathbf{X}}(t)$ from a to b. This lets us see probabilities in the graph.

For example, below I have shaded the area representing $\Pr[0.2 \leq \mathbf{X} \leq 0.7]$ in blue and $\Pr[1.2 \leq \mathbf{X} \leq 1.7]$ in red:



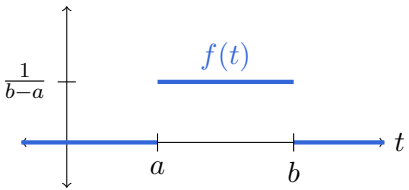
Suppose we took lots of samples from $\mathbf{X}$ and plotted them in a histogram. The histogram we'd get would roughly match the shape of the PDF (but the scale on the y-axis would be different).

PDF of the uniform distribution

A random variable with the *Uniform*(a, b) distribution has PDF

$$f(t) = \begin{cases} \frac{1}{b-a} & a \le t \le b \\ 0 & \text{otherwise.} \end{cases}$$

Here is a plot:



Looking at this PDF tells us that all values in $[a, b]$ are equally likely, and that all values outside $[a, b]$ are impossible.

Properties of PDFs

Here are the things we can say about a PDF $f_{\mathbf{X}}(t)$:

- We must have $f_{\mathbf{X}}(t) \geq 0$ for all $t \in \mathbb{R}$.
- We must have $\int_{-\infty}^{\infty} f_{\mathbf{X}}(t)\, dt = 1$, since this corresponds to the total probability that $\mathbf{X}$ is anything.
- The values of $f_{\mathbf{X}}$ are **not probabilities!!!** In particular, they can be bigger than 1.

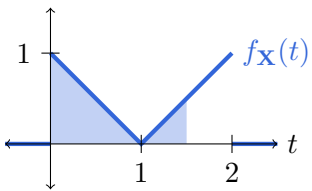
  Example: $\mathbf{X} \sim \mathit{Uniform}(0, 0.01)$ has $f_{\mathbf{X}}(t) = 100$ for $t \in [0, 0.01]$.
- However, larger or smaller values of $f_{\mathbf{X}}$ do correspond to more or less likely values of $\mathbf{X}$.

  (If $f_{\mathbf{X}}(a) > f_{\mathbf{X}}(b)$, then "$\mathbf{X} \approx a$" is likelier than "$\mathbf{X} \approx b$", once we settle on what "$\approx$" means.)

Using a PDF

If $\mathbf{X}$ has the PDF below, what is $\Pr[\mathbf{X} \leq 1.5]$?

$$f_{\mathbf{X}}(t) = \begin{cases} 0 & t < 0 \\ 1 - t & 0 \leq t \leq 1 \\ t - 1 & 1 \leq t \leq 2 \\ 0 & t > 2 \end{cases}$$



Using integrals:

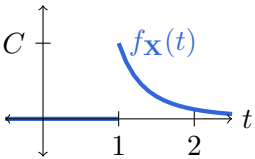
$$\Pr[\mathbf{X} \leq 1.5] = \int_{-\infty}^{0} 0\,dt + \int_{0}^{1} (1-t)\,dt + \int_{1}^{1.5} (t-1)\,dt.$$

We can also just find the area of the two shaded triangles: $\frac{1}{2} + \frac{1}{8} = \frac{5}{8}$.

Normalizing constant

If $\mathbf{X}$ has the PDF below, what should the value of C be?

$$f_{\mathbf{X}}(t) = \begin{cases} 0 & t < 1 \\ C/t^3 & t \geq 1 \end{cases}$$



All PDFs integrate to 1, so we must get

$$\int_{-\infty}^{\infty} f_{\mathbf{X}}(t)\,dt = \int_{1}^{\infty} \frac{C}{t^3}\,dt = 1.$$

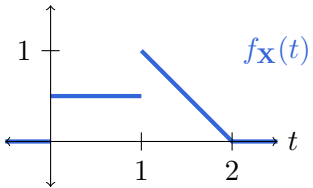
We get an antiderivative of $-\frac{C}{2t^2}$, which tends to 0 as $t \to \infty$, so we want

$$-\frac{C}{2t^2}\bigg|_{t=1}^{\infty} = 0 - \left(-\frac{C}{2(1)^2}\right) = 1 \implies C = 2.$$

From a PDF to a CDF

If $\mathbf{X}$ has the PDF below, what is its CDF?

$$f_{\mathbf{X}}(t) = \begin{cases} 0 & t < 0 \\ 1/2 & 0 \le t \le 1 \\ 2 - t & 1 \le t \le 2 \\ 0 & t > 2 \end{cases}$$

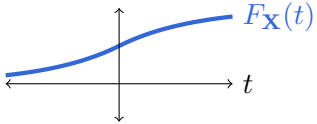


- For $u < 0$, we have $F_{\mathbf{X}}(u) = 0$.
- For $0 \le u < 1$, we can integrate $\int_0^u \frac{1}{2}\,dt$ or find the area: $F_{\mathbf{X}}(u) = \frac{1}{2}u$.
- For $1 \le u < 2$, integrating $\int_1^u (2 - t)\,dt$ gives $-\frac{1}{2}u^2 + 2u - \frac{3}{2}$. But don't forget $\Pr[0 \le \mathbf{X} \le 1]$! We get $F_{\mathbf{X}}(u) = -\frac{1}{2}u^2 + 2u - 1$.
- For $u \ge 2$, $F_{\mathbf{X}}(u) = 1$.

From a CDF to a PDF

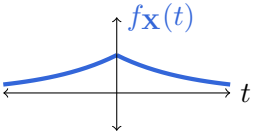
If $\mathbf{X}$ has the CDF below, what is its PDF?

$$F_{\mathbf{X}}(t) = \begin{cases} \frac{1}{2}e^{t} & t \leq 0 \\ 1 - \frac{1}{2}e^{-t} & t \geq 0 \end{cases}$$



To get the CDF, we integrated, so to get the PDF, we take a derivative: $f_{\mathbf{X}}(t) = F'_{\mathbf{X}}(t)$, separately for each case. We get

$$f_{\mathbf{X}}(t) = \begin{cases} \frac{1}{2}e^{t} & t \leq 0 \\ \frac{1}{2}e^{-t} & t \geq 0 \end{cases}$$



The radioactive decay problem

Nitrogen-13 is a radioactive isotope of nitrogen with a half-life of approximately 10 minutes. (Every 10 minutes, there is a 50% probability that a given ^{13}N atom will decay.)

Let $\mathbf{X}$ be the time until a particular ^{13}N atom decays. What is the distribution of $\mathbf{X}$?

Bad answer. We could say that it takes *Geometric*$(\frac{1}{2})$ ten-minute half-lives.

This is not entirely wrong. But $\mathbf{X}$ is a continuous distribution! The ^{13}N atom doesn't flip a coin once every 10 minutes; it could decay at any moment.

Zooming in

We could do a more fine-grained discrete approximation. Consider the model where, once every minute, the ^{13}N atom decays with probability $p = 1 - (\frac{1}{2})^{1/10} \approx 0.067$.

In other words, the atom decays in *Geometric*(p) minutes.

- If $\mathbf{X} \sim$ *Geometric*(p), measured in minutes, we have

$$\Pr[\mathbf{X} \leq 10] = 1 - (1-p)^{10} = \frac{1}{2}.$$

  So the half-life is "working as intended".

- This is still not continuous: we're only checking in on the atom once every minute.

The δ-step process

We can generalize this idea. Take a very small δ, and say that over any interval of δ seconds, the particle decays with probability p.

We want to choose p so that over $\frac{600}{\delta}$ intervals of length δ, the probability is $\frac{1}{2}$. So:

$$1 - (1-p)^{\frac{600}{\delta}} = \frac{1}{2} \implies p = 1 - \left(\frac{1}{2}\right)^{\frac{\delta}{600}}.$$

If $\mathbf{N} \sim \textit{Geometric}(p)$ and $\mathbf{X} = \delta \cdot \mathbf{N}$, that's still not exact, but we can make this approximation arbitrarily good.

What is the continuous distribution we're approaching?

The CDF of the decay time

Let's figure out the CDF of $\mathbf{X}$: the probability $\Pr[\mathbf{X} \leq t]$.

In the δ-step approximation, $\mathbf{X} = \delta \cdot \mathbf{N}$, where $\mathbf{N} \sim$ *Geometric*(p) for some tiny p, so this probability is

$$\Pr[\mathbf{X} \leq t] = \Pr\left[\mathbf{N} \leq \frac{t}{\delta}\right] = 1 - (1-p)^{t/\delta}$$

which is $1 - c^t$ for some constant c. It's more traditional to write $F_{\mathbf{X}}(t) = 1 - e^{-\lambda t}$ for some constant λ.

What is λ? If $\mathbf{X}$ is in seconds, then to respect the half-life, we want

$$F_{\mathbf{X}}(600) = \frac{1}{2} \implies 1 - e^{-600\lambda} = \frac{1}{2} \implies \lambda = \frac{\ln 2}{600} \approx 0.00116.$$

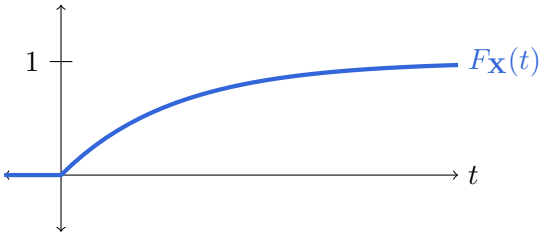
The exponential distribution

We say that $\mathbf{X}$ has the **exponential distribution** with **rate** λ if

$$F_{\mathbf{X}}(t) = \begin{cases} 0 & t < 0 \\ 1 - e^{-\lambda t} & t \geq 0 \end{cases}$$

Shorthand: $\mathbf{X} \sim \textit{Exponential}(\lambda)$.

Here is a plot of the CDF:

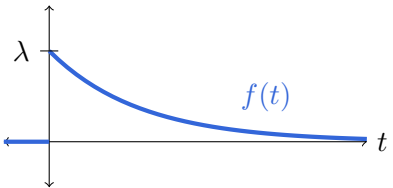


PDF of the exponential distribution

A random variable with the *Exponential*(λ) distribution has PDF

$$f(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0. \end{cases}$$

Here is a plot:



This has roughly the same shape as the geometric PMF. Decreasing values of $f(t)$ mean that larger values are less likely.

A conditional probability

Q. Suppose $\mathbf{X} \sim \textit{Exponential}(\lambda)$. What is $\Pr[\mathbf{X} \leq t+1 \mid \mathbf{X} \geq t]$?

(For example, if our ^{13}N atom has lasted for 1 hour, what's the probability it decays in the next minute?)

A. We can write this in terms of the CDF:

$$\Pr[\mathbf{X} \leq t+1 \mid \mathbf{X} \geq t] = \frac{\Pr[t \leq \mathbf{X} \leq t+1]}{\Pr[\mathbf{X} \geq t]} = \frac{F_{\mathbf{X}}(t+1) - F_{\mathbf{X}}(t)}{1 - F_{\mathbf{X}}(t)}.$$

For the exponential distribution, we get:

$$\Pr[\mathbf{X} \leq t+1 \mid \mathbf{X} \geq t] = \frac{(1 - e^{-\lambda(t+1)}) - (1 - e^{-\lambda t})}{1 - (1 - e^{-\lambda t})} = 1 - e^{-\lambda}.$$

Notably, this doesn't depend on t.

Half-life

Q. Suppose we have a radioactive particle, and we start with the time-to-decay distribution: $\mathbf{X} \sim \textit{Exponential}(\lambda)$, for some λ. What is its half-life?

A. The half-life $t_{1/2}$ is the time after which the particle has a $\frac{1}{2}$ chance to decay: $F_{\mathbf{X}}(t_{1/2}) = \frac{1}{2}$. In general, this is called the **median** of a probability distribution.

In our case: $1 - e^{-\lambda t_{1/2}} = \frac{1}{2}$ gives us a half-life of $t_{1/2} = \frac{\ln 2}{\lambda}$.

Note that this is different from the **mean** or **expected value** of our random variable! We will see later than $\mathbb{E}[\mathbf{X}] = \frac{1}{\lambda}$: off by a factor of $\ln 2 \approx 0.693$.

Expected value

For discrete random variables:

$$\mathbb{E}[\mathbf{X}] = \sum_{k \in R_{\mathbf{X}}} k \cdot P_{\mathbf{X}}(k).$$

For continuous random variables:

$$\mathbb{E}[\mathbf{X}] = \int_{-\infty}^{\infty} t \cdot f_{\mathbf{X}}(t)\, dt.$$

Intuition: $f_{\mathbf{X}}(t)\, dt$ is like the probability that $\mathbf{X}$ is in the infinitesimally short interval $[t, t + dt]$.

In this interval, $\mathbf{X}$ is approximately t, so we multiply t by this probability, and "sum over all the infinitesimally short intervals".

Expected value of a uniform

Suppose that $\mathbf{X} \sim \textit{Uniform}(a, b)$. What is $\mathbb{E}[\mathbf{X}]$? (We expect $\frac{a+b}{2}$.)

The PDF of $\mathbf{X}$ is $f_{\mathbf{X}}(t) = \frac{1}{b-a}$ when $a \leq t \leq b$, and 0 otherwise. Therefore

$$\mathbb{E}[\mathbf{X}] = \int_{-\infty}^{\infty} t \cdot f_{\mathbf{X}}(t)\, dt = \int_a^b t \cdot \frac{1}{b-a}\, dt.$$

The antiderivative of $\frac{t}{b-a}$ is $\frac{t^2/2}{b-a} + C$. So we get

$$\mathbb{E}[\mathbf{X}] = \left.\frac{t^2/2}{b-a}\right|_{t=a}^{b} = \frac{b^2/2}{b-a} - \frac{a^2/2}{b-a}.$$

Since $b^2 - a^2$ factors as $(b+a)(b-a)$, $\mathbb{E}[\mathbf{X}]$ simplifies to $\frac{a+b}{2}$.

Expected value of an exponential

Suppose that $\mathbf{X} \sim \textit{Exponential}(\lambda)$. What is $\mathbb{E}[\mathbf{X}]$?

The PDF of $\mathbf{X}$ is $f_{\mathbf{X}}(t) = \lambda e^{-\lambda t}$ when $t \geq 0$, and 0 otherwise. Therefore

$$\mathbb{E}[\mathbf{X}] = \int_{-\infty}^{\infty} t \cdot f_{\mathbf{X}}(t)\, dt = \int_{0}^{\infty} t \cdot \lambda e^{-\lambda t}\, dt.$$

Integrate by parts with $u = t$ and $dv = \lambda e^{-\lambda t}\, dt$ to get

$$\mathbb{E}[\mathbf{X}] = t \cdot -e^{-\lambda t}\Big|_{t=0}^{\infty} - \int_{0}^{\infty} -e^{-\lambda t}\, dt = 0 + \int_{0}^{\infty} e^{-\lambda t}\, dt.$$

The antiderivative of $e^{-\lambda t}$ is $-\frac{1}{\lambda} e^{-\lambda t} + C$, giving us

$$\mathbb{E}[\mathbf{X}] = -\frac{1}{\lambda} e^{-\lambda t}\Big|_{t=0}^{\infty} = 0 - \left(-\frac{1}{\lambda}\right) = \frac{1}{\lambda}.$$

Expected value of a function

Suppose we know the distribution of $\mathbf{X}$, and a function $h : \mathbb{R} \to \mathbb{R}$. (Or just $R_{\mathbf{X}} \to \mathbb{R}$.) How can we find $\mathbb{E}[h(\mathbf{X})]$?

Method 1: Use the definition of expected value.

1. Find the PDF of $h(\mathbf{X})$.

   (We will discuss how to do this in the next lecture.)

2. Use the PDF of $h(\mathbf{X})$ to find its expected value.

Method 2: LOTUS for continuous random variables.

$$\mathbb{E}[h(\mathbf{X})] = \int_{-\infty}^{\infty} h(t) \cdot f_{\mathbf{X}}(t)\, dt.$$

This is usually easier.

Moments of *Uniform*(a, b)

Suppose that $\mathbf{X} \sim$ *Uniform*(a, b). What is $\mathbb{E}[\mathbf{X}^k]$?

Using LOTUS:

$$\mathbb{E}[\mathbf{X}^k] = \int_{-\infty}^{\infty} t^k \cdot f_{\mathbf{X}}(t)\, dt = \int_a^b t^k \cdot \frac{1}{b-a}\, dt.$$

The antiderivative of $\frac{t^k}{b-a}$ is $\frac{t^{k+1}/(k+1)}{b-a} + C$, so we get

$$\mathbb{E}[\mathbf{X}^k] = \left.\frac{t^{k+1}/(k+1)}{b-a}\right|_{t=a}^{b} = \frac{b^{k+1} - a^{k+1}}{(k+1)(b-a)}.$$

In particular: $\mathrm{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2 = \frac{b^3-a^3}{3(b-a)} - (\frac{a+b}{2})^2 = \frac{(a-b)^2}{12}$.

Scaling uniform random variables

Here's another approach to finding $\mathbb{E}[\mathbf{X}]$ and $\mathrm{Var}[\mathbf{X}]$ when $\mathbf{X} \sim \textit{Uniform}(a, b)$.

It's enough to use the method from the previous slide to show that when $\mathbf{X} \sim \textit{Uniform}(0, 1)$, we have $\mathbb{E}[\mathbf{X}] = \frac{1}{2}$ and $\mathrm{Var}[\mathbf{X}] = \frac{1}{12}$. Then:

1. If we multiply by a constant, we get $c\mathbf{X} \sim \textit{Uniform}(0, c)$.

   Then $\mathbb{E}[c\mathbf{X}] = \frac{c}{2}$ and $\mathrm{Var}[\mathbf{X}] = \frac{c^2}{12}$.

2. If we add a constant, we get $a + c\mathbf{X} \sim \textit{Uniform}(a, a + c)$.

   Then $\mathbb{E}[a + c\mathbf{X}] = a + \frac{c}{2}$ and $\mathrm{Var}[a + c\mathbf{X}] = \frac{c^2}{12}$.

Set $c = b - a$ to get the formulas for $\textit{Uniform}(a, b)$.

Variance of *Exponential*(λ)

Suppose that $\mathbf{X} \sim$ *Exponential*(λ). What is $\mathbb{E}[\mathbf{X}^2]$?

Using LOTUS:

$$\mathbb{E}[\mathbf{X}^2] = \int_{-\infty}^{\infty} t^2 \cdot f_{\mathbf{X}}(t)\, dt = \int_{0}^{\infty} t^2 \cdot \lambda e^{-\lambda t}\, dt.$$

This requires applying integration by parts **twice** to $\lambda t^2 e^{-\lambda t}$. We get

$$\int \lambda t^2 e^{-\lambda t} = \left(-t^2 - \frac{2t}{\lambda} - \frac{2}{\lambda^2}\right) e^{-\lambda t} + C$$

This leads to $\mathbb{E}[\mathbf{X}^2] = \frac{2}{\lambda^2}$. $\mathrm{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2 = \frac{1}{\lambda^2}$.

Moment generating functions

Much of what we did with the PGF $G_{\mathbf{X}}(z) = \mathbb{E}[z^{\mathbf{X}}]$ for discrete random variables can be done for a continuous random variable $\mathbf{X}$ with the **moment generating function**

$$M_{\mathbf{X}}(s) = \mathbb{E}[e^{-s\mathbf{X}}].$$

We won't go into details, but we can look at an example to practice.

Example. If $\mathbf{X} \sim$ *Exponential*(λ), what is $M_{\mathbf{X}}(s)$?

We compute

$$\int_0^\infty e^{-st} \cdot \lambda e^{-\lambda t}\, dt = -\frac{\lambda}{\lambda + s} e^{-(s+\lambda)t}\Big|_{t=0}^{\infty} = \frac{\lambda}{\lambda + s}.$$

Claim. Another way to find $\mathbb{E}[\mathbf{X}^2]$ is to compute $M''_{\mathbf{X}}(0)$.

Transforming random variables

General question. Suppose we know the distribution of $\mathbf{X}$. Given a function $h : \mathbb{R} \to \mathbb{R}$ (or $R_{\mathbf{X}} \to \mathbb{R}$), what is the distribution of $h(\mathbf{X})$?

- We have already discussed this for discrete random variables, where the key is understanding the set $A_k = \{x \in R_{\mathbf{X}} : h(x) = k\}$.
- We have seen one example for continuous random variables: if $\mathbf{X} \sim \textit{Uniform}(0, 1)$, then $a + (b - a)\mathbf{X} \sim \textit{Uniform}(a, b)$.
- Today, we will see ways to solve this problem in general.

Transformations and the CDF

If we know the CDF $F_{\mathbf{X}}(t)$ of $\mathbf{X}$, then we can try to use it to solve for

$$F_{h(\mathbf{X})}(t) = \Pr[h(\mathbf{X}) \leq t].$$

This will describe the distribution of $h(\mathbf{X})$ completely.

Example. Let $\mathbf{X} \sim \textit{Uniform}(0, 1)$. What is the distribution of $2\mathbf{X} - 1$?

We have $F_{2\mathbf{X}-1}(t) = \Pr[2\mathbf{X} - 1 \leq t] = \Pr[\mathbf{X} \leq \frac{t+1}{2}]$, so

$$F_{2\mathbf{X}-1}(t) = \begin{cases} 0 & \frac{t+1}{2} < 0 \\ \frac{t+1}{2} & 0 \leq \frac{t+1}{2} < 1 \\ 1 & \frac{t+1}{2} \geq 1 \end{cases} = \begin{cases} 0 & t < -1 \\ \frac{t+1}{2} & -1 \leq t < 1 \\ 1 & t \geq 1 \end{cases}$$

We can recognize this as the CDF of the $\textit{Uniform}(-1, 1)$ distribution.

A more difficult transformation

Example 2. Let $\mathbf{X} \sim \textit{Uniform}(-1, 2)$. What is the distribution of $\mathbf{Y} = \mathbf{X}^2$?

We have

$$F_{\mathbf{Y}}(t) = \Pr[\mathbf{X}^2 \leq t] = \begin{cases} 0 & t \leq 0 \\ \Pr[|\mathbf{X}| \leq \sqrt{t}] & t \geq 0. \end{cases}$$

Things get tricky here:

- When $0 \leq \sqrt{t} \leq 1$, $\Pr[|\mathbf{X}| \leq \sqrt{t}] = \Pr[-\sqrt{t} \leq \mathbf{X} \leq \sqrt{t}] = \frac{2\sqrt{t}}{3}$.
- When $1 \leq \sqrt{t} \leq 2$, $\Pr[|\mathbf{X}| \leq \sqrt{t}] = \Pr[-1 \leq \mathbf{X} \leq \sqrt{t}] = \frac{1+\sqrt{t}}{3}$.
- When $\sqrt{t} \geq 2$, $\Pr[|\mathbf{X}| \leq \sqrt{t}] = 1$.

We can put this together to write down a 4-part formula for $F_{\mathbf{Y}}(t)$.

Uniform to exponential

Example 3. Let $\mathbf{X} \sim \textit{Uniform}(0, 1)$ and $\mathbf{Y} = -\ln \mathbf{X}$. Then $\mathbf{Y} \sim \textit{Exponential}(1)$.

We have

$$F_{\mathbf{Y}}(t) = \Pr[\mathbf{Y} \leq t] = \Pr[-\ln \mathbf{X} \leq t] = \Pr[\mathbf{X} \geq e^{-t}].$$

For $t \leq 0$, $e^{-t} \geq 1$, so this probability is 0.

For $t \geq 0$, $0 \leq e^{-t} \leq 1$, so this probability is $1 - e^{-t}$.

Putting this together,

$$F_{\mathbf{Y}}(t) = \begin{cases} 1 - e^{-t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

which is the CDF of an $\textit{Exponential}(1)$ random variable.

Method of transformations

As a special case, suppose $h(t)$ is **strictly increasing on the range of $\mathbf{X}$**. Then

$$F_{h(\mathbf{X})}(t) = \Pr[h(\mathbf{X}) \leq t] = \Pr[\mathbf{X} \leq h^{-1}(t)] = F_{\mathbf{X}}(h^{-1}(t)).$$

The derivative of $h^{-1}(t)$ is $\frac{1}{h'(h^{-1}(t))}$. Therefore we can take the derivative of the equation above to get

$$f_{h(\mathbf{X})}(t) = f_{\mathbf{X}}(h^{-1}(t)) \cdot \frac{1}{h'(h^{-1}(t))}.$$

(This assumes $h^{-1}(t)$ exists, but if it doesn't, then t is not in the range of $h(\mathbf{X})$, and so $f_{h(\mathbf{X})}(t)$ should be 0.)

This lets us compute the PDF of $h(\mathbf{X})$ directly from the PDF of $\mathbf{X}$.

Example of the method of transformations

Suppose $\mathbf{X} \sim \textit{Uniform}(0, 2)$. What is the distribution of $\mathbf{Y} = \mathbf{X}^2$?

First step: the range of $\mathbf{X}$ is $[0, 2]$, $h(t) = t^2$ is increasing on that range, and the range of $\mathbf{Y}$ is $[0^2, 2^2] = [0, 4]$.

Next, we use the formula from the previous slide:

$$f_{h(\mathbf{X})}(t) = f_{\mathbf{X}}(h^{-1}(t)) \cdot \frac{1}{h'(h^{-1}(t))}.$$

Here, $f_{\mathbf{X}}(t) = \frac{1}{2}$, $h(t) = t^2$, $h^{-1}(t) = \sqrt{t}$, and $h'(t) = 2t$.

This gives us

$$f_{\mathbf{Y}}(t) = \begin{cases} \frac{1}{4\sqrt{t}} & 0 \leq t \leq 4 \\ 0 & \text{otherwise.} \end{cases}$$

The PDF and scaling random variables

Suppose we know the PDF of $\mathbf{X}$. What is the PDF of $a\mathbf{X} + b$? (Assume $a > 0$.)

We have a general (and scary) rule: for a strictly increasing transformation $h(t)$,

$$f_{h(\mathbf{X})}(t) = f_{\mathbf{X}}(h^{-1}(t)) \cdot \frac{1}{h'(h^{-1}(t))}.$$

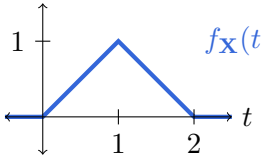
In our case, $h(t) = at + b$, so $h^{-1}(t) = \frac{t-b}{a}$ and $h'(t) = a$.

This tells us that

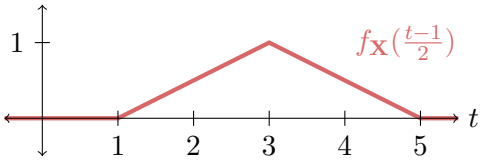
$$f_{a\mathbf{X}+b}(t) = \frac{1}{a} f_{\mathbf{X}}\left(\frac{t-b}{a}\right).$$

Intuition: scaling the graph of a PDF

Suppose that this is $f_{\mathbf{X}}(t)$. How can we find the PDF of $2\mathbf{X}+1$?



Going to $f_{\mathbf{X}}(\frac{t}{2})$ will stretch it, and $f_{\mathbf{X}}(\frac{t-1}{2})$ will shift it right by 1:



We take $\frac{1}{2}f_{\mathbf{X}}(\frac{t-1}{2})$ to return the area under the curve to 1.

Scaling the exponential

If $\mathbf{X} \sim \textit{Exponential}(\lambda)$, then

$$f_{\mathbf{X}}(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t \leq 0. \end{cases}$$

What is the distribution of $a\mathbf{X}$?

The rule says: $f_{a\mathbf{X}}(t) = \frac{1}{a} f_{\mathbf{X}}(\frac{t}{a})$. This gives us

$$f_{a\mathbf{X}}(t) = \begin{cases} \frac{\lambda}{a} e^{-\lambda t/a} & t \geq 0 \\ 0 & t \leq 0. \end{cases}$$

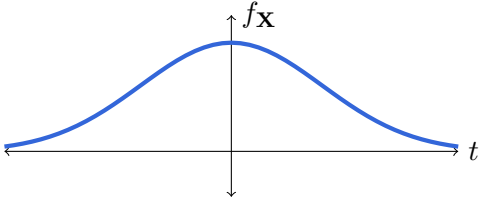
Comparing, we conclude $a\mathbf{X} \sim \textit{Exponential}(\frac{\lambda}{a})$. We say λ is an "inverse scale" parameter.

Incomprehensible definition

We say that $\mathbf{X} \sim \mathit{Normal}(0, 1)$, or $\mathbf{X}$ has the **standard normal distribution**, if

$$f_{\mathbf{X}}(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}.$$

Here is a plot:



Why is this distribution interesting? More on this later.

General normal distribution

If $\mathbf{X} \sim \textit{Normal}(0, 1)$, then $\mathbb{E}[\mathbf{X}] = 0$ and $\mathrm{Var}[\mathbf{X}] = 1$.

Let $\mathbf{Y} = \sigma\mathbf{X} + \mu$. (We may assume $\sigma > 0$.) Then $\mathbb{E}[\mathbf{Y}] = \mu$ and $\mathrm{Var}[\mathbf{Y}] = \sigma^2$. We say that $\mathbf{Y} \sim \textit{Normal}(\mu, \sigma^2)$.

What is the PDF of $\mathbf{Y}$?

Applying the rule $f_{a\mathbf{X}+b}(t) = \frac{1}{a} f_{\mathbf{X}}(\frac{t-b}{a})$, we get:

$$f_{\mathbf{Y}}(t) = \frac{1}{\sigma} f_{\mathbf{X}}\left(\frac{t-\mu}{\sigma}\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-\mu)^2/(2\sigma^2)}.$$

The graph of the PDF is a bell curve centered at μ. It is narrow and tall when σ is small, and wide and flat when σ is large.

Sums of normal random variables

A boring property that follows from the definition: if $\mathbf{X}$ is normally distributed, then so is $a\mathbf{X} + b$, for any $a, b \in \mathbb{R}$ (maybe with $a \neq 0$).

A vitally important and surprising property: if $\mathbf{X}, \mathbf{Y}$ are normally distributed and independent, then so is $\mathbf{X} + \mathbf{Y}$.

As a result, if $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are independent normals, then so is

$$a_0 + a_1\mathbf{X}_1 + a_2\mathbf{X}_2 + \cdots + a_n\mathbf{X}_n.$$

It is often the case that if you sum many **arbitrary** independent random variables, their sum will be **approximately** normal.

(In all of the cases above: which normal? We can work out μ and σ^2 from properties of mean and variance.)

Applications

Many real-world random variables are the sum of many independent factors, and so tend to be approximately normal. Examples:

- The time it takes me to grade 60 exams.
- Amount of rainfall over a year in a particular location.

Also, many discrete random variables have normal approximations (under some assumptions), because:

- *Binomial*(n, p) is the sum of n independent *Bernoulli*(p)'s;
- when $\lambda \in \mathbb{N}$, *Poisson*(λ) is the sum of λ independent *Poisson*(1)'s;
- and so on.

(We'll see more how and when these approximations hold.)

Convergence to a normal distribution

When we talk about a sequence of random variables $\mathbf{S}_1, \mathbf{S}_2, \ldots$ converging to a normal distribution, what do we mean?

First, we should standardize these, replacing each $\mathbf{S}_i$ by $\mathbf{Z}_i := \frac{\mathbf{S}_i - \mathbb{E}[\mathbf{S}_i]}{\mathrm{SD}[\mathbf{S}_i]}$. Then we can compare them to a random variable $\mathbf{Z} \sim \mathit{Normal}(0, 1)$.

In particular: if $\mathbf{S}_n$ is a sum $\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n$, where each $\mathbf{X}_i$ has mean μ and variance σ^2, then

$$\mathbf{Z}_n = \frac{\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n - n\mu}{\sqrt{n}\sigma}$$

might hopefully be approximately normal.

A concrete statement: we can hope that each probability $\Pr[\mathbf{Z}_n \leq t]$ converges to $\Pr[\mathbf{Z} \leq t]$ as $n \to \infty$.

The Central Limit Theorem

Theorem. For any sequence of i.i.d. random variables $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3, \ldots$ with mean μ and variance σ^2 (both finite),

$$\mathbf{Z}_n = \frac{\mathbf{X}_1 + \mathbf{X}_2 + \cdots + \mathbf{X}_n - n\mu}{\sqrt{n}\sigma}$$

converges to $\mathbf{Z} \sim \textit{Normal}(0, 1)$ as $n \to \infty$.

That is, for all $u \in \mathbb{R}$,

$$\lim_{n\to\infty} \Pr[\mathbf{Z}_n \leq u] = \Pr[\mathbf{Z} \leq u] = \int_{-\infty}^{u} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt.$$

Continuity correction

Suppose that $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are **discrete, integer-valued** random variables, $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_n$, and we want to know $\Pr[a \leq \mathbf{S} \leq b]$.

By the Central Limit Theorem, this is approximately $\Pr\left[\frac{a-n\mu}{\sqrt{n}\sigma} \leq \mathbf{Z} \leq \frac{b-n\mu}{\sqrt{n}\sigma}\right]$ for $\mathbf{Z} \sim \textit{Normal}(0, 1)$.

But the same event $\Pr[a \leq \mathbf{S} \leq b]$ can be described in many ways:

$$\Pr[1 \leq \mathbf{S} \leq 5] = \Pr[0.5 \leq \mathbf{S} \leq 5.5] = \Pr[0.001 \leq \mathbf{S} \leq 5.999].$$

We usually get the best approximation for the probability in the middle: if we take a and b to be halfway between two integers.

Example 1: Approximating a binomial

Suppose we flip a fair coin 100 times and it lands heads 37 times. Is this unusual for a fair coin?

If $\mathbf{X} \sim \textit{Binomial}(100, \frac{1}{2})$, then $\mathbb{E}[\mathbf{X}] = 50$ and $\mathrm{SD}[\mathbf{X}] = 5$, so $\frac{\mathbf{X}-50}{5}$ can be approximated by a standard normal.

The chances of $\mathbf{X}$ being **closer to the mean than 37** are about

$$\Pr\left[\frac{37.5-50}{5} \leq \mathbf{Z} \leq \frac{62.5-50}{5}\right] \approx 0.9876.$$

A fair coin should only be this far from the mean about 1.24% of the time.

(In a few lectures, we'll see more null hypothesis testing, and the rationale behind asking the question in this way.)

Example 2: Stirling's formula

If $\mathbf{X} \sim \textit{Poisson}(n)$, then $\mathbf{X}$ is the sum of n $\textit{Poisson}(1)$'s. By the central limit theorem,

$$\Pr[\mathbf{X} = n] = \Pr[n - \tfrac{1}{2} \leq \mathbf{X} \leq n + \tfrac{1}{2}] \approx \Pr[-\tfrac{1}{2\sqrt{n}} \leq \mathbf{Z} \leq \tfrac{1}{2\sqrt{n}}].$$

Integrating the normal PDF is hard, so let's make a further approximation:

$$\Pr[\mathbf{X} = n] \approx \frac{1}{\sqrt{n}} f_{\mathbf{Z}}(0) = \frac{1}{\sqrt{2\pi n}}.$$

Why bother, when we have an exact formula: $\Pr[\mathbf{X} = n] = e^{-n}\frac{n^n}{n!}$? Because

$$e^{-n}\frac{n^n}{n!} \approx \frac{1}{\sqrt{2\pi n}} \implies n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

This is an excellent approximation to $n!$ known as Stirling's formula.

Berry–Esseen theorem (very optional)

The drawback of the Central Limit Theorem is that it makes no concrete promises: how quickly does the probability converge? How bad is the error in the normal approximation?

More concrete statements **do exist**. For example, the Berry–Esseen theorem.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be independent and $\mathbf{S} = \mathbf{X}_1 + \cdots + \mathbf{X}_n$. Assume everything is already adjusted so that $\mathbb{E}[\mathbf{X}_1] = \cdots = \mathbb{E}[\mathbf{X}_n] = \mathbb{E}[\mathbf{S}] = 0$ and $\mathrm{Var}[\mathbf{S}] = 1$. Let $\mathbf{Z} \sim \textit{Normal}(0, 1)$.

Then

$$\Big| \Pr[\mathbf{S} \leq u] - \Pr[\mathbf{Z} \leq u] \Big| \leq \sum_{i=1}^{n} \mathbb{E}[|\mathbf{X}_i|^3].$$

In the typical case, this upper bound is proportional to $\frac{1}{\sqrt{n}}$.

Bernoulli and Poisson processes

Some of our discrete random variables lived in a common experiment called a **Bernoulli process**, where we do a sequence of trials that independently succeed with probability p.

- The number of successes in a block of n trials is *Binomial*(n, p).
- The position of the first success is *Geometric*(p).
- The position of the m^{th} success is *Pascal*(m, p).

A **Poisson process** is a continuous analogue.

First, imagine a Bernoulli process with one trial per minute. Then, make the trials more frequent while preserving the rate of successes: N trials per minute with a $\frac{p}{N}$ probability of success. As $N \to \infty$, we get a Poisson process.

More on the Poisson process

Poisson processes model individual events we can count ("arrivals") that happen continuously over time. Some examples:

- visits to a website;
- earthquakes in a region;
- customers visiting a store.

The two key properties that make the process Poisson are:

1. A constant rate $\lambda > 0$ such that in any time period of length t, the expected number of events is λt.
2. Independent increments: the inter-arrival times (waiting times between the i^{th} and $(i+1)^{\text{th}}$ event) are independent.

The three random variables

We can describe the arrivals by a **counting process**: for all real numbers $t > 0$, $\mathbf{N}(t)$ is the number of arrivals in the interval $[0, t]$.

This has the distribution $\mathbf{N}(t) \sim \textit{Poisson}(\lambda t)$. (This does what the Binomial did in the discrete process.)

We can also describe the arrivals by their arrival times: $\mathbf{T}_1, \mathbf{T}_2, \mathbf{T}_3, \ldots$ where $\mathbf{T}_k$ is the time of the k^{th} arrival.

The inter-arrival times $\mathbf{X}_i = \mathbf{T}_i - \mathbf{T}_{i-1}$ (with $\mathbf{X}_1 = \mathbf{T}_1$) are independent, and $\mathbf{X}_i \sim \textit{Exponential}(\lambda)$. (This does what the Geometric did in the discrete process.)

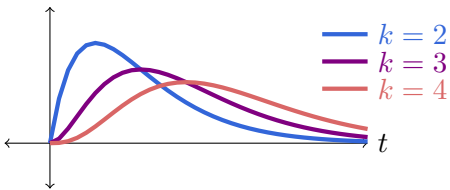
The arrival times have a new distribution: $\mathbf{T}_k \sim \textit{Gamma}(k, \lambda)$. (This does what the Pascal did in the discrete process.)

The Gamma distribution

We say that $\mathbf{X} \sim \textit{Gamma}(k, \lambda)$ if, for all $t \geq 0$,

$$f_{\mathbf{X}}(t) = \frac{\lambda^k}{(k-1)!} t^{k-1} e^{-\lambda t}$$

with $f_{\mathbf{X}}(t) = 0$ for $t < 0$. A few plots:



We get $\mathbb{E}[\mathbf{X}] = \frac{k}{\lambda}$ and $\mathrm{Var}[\mathbf{X}] = \frac{k}{\lambda^2}$ because $\mathbf{X}$ is the sum of k independent $\textit{Exponential}(\lambda)$'s.

The CDF of the Gamma distribution

We can relate $\mathbf{T}_k \sim \textit{Gamma}(k, \lambda)$ to $\mathbf{N}(t) \sim \textit{Poisson}(\lambda t)$. The idea: in the Poisson process, "It takes at most t seconds to see k arrivals" is the same as "In t seconds, there are at least k arrivals". Therefore

$$F_{\mathbf{T}_k}(t) = \Pr[\mathbf{T}_k \leq t] = \Pr[\mathbf{N}(t) \geq k] = 1 - \sum_{i=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^i}{i!}.$$

For example, if we take $\mathbf{T}_3 \sim \textit{Gamma}(3, \lambda)$, we have

$$F_{\mathbf{T}_3}(t) = 1 - e^{-\lambda t} - e^{-\lambda t} \frac{(\lambda t)^1}{1!} - e^{-\lambda t} \frac{(\lambda t)^2}{2!}.$$

In general, the CDF of the Gamma looks messy; the PDF is much nicer, because lots of things cancel.

The PDF of the Gamma distribution

Let's first see what happens for $\mathbf{T}_3 \sim \mathit{Gamma}(3, \lambda)$, and take the derivative of

$$F_{\mathbf{T}_3}(t) = 1 - e^{-\lambda t} - e^{-\lambda t}\frac{(\lambda t)^1}{1!} - e^{-\lambda t}\frac{(\lambda t)^2}{2!}.$$

We get

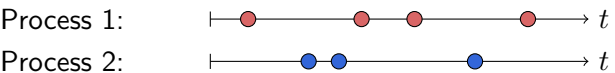
$$f_{\mathbf{T}_3}(t) = \lambda e^{-\lambda t} + \lambda e^{-\lambda t}\frac{(\lambda t)^1}{1!} - \lambda e^{-\lambda t} + \lambda e^{-\lambda t}\frac{(\lambda t)^2}{2!} - \lambda e^{-\lambda t}\frac{(\lambda t)^1}{1!}.$$

Lots of terms cancel, and we're left with $f_{\mathbf{T}_3}(t) = \lambda e^{-\lambda t}\frac{(\lambda t)^2}{2!}$ only.

In general, $\mathbf{T}_k \sim \mathit{Gamma}(k, \lambda)$ has the PDF $\lambda e^{-\lambda t}\frac{(\lambda t)^{k-1}}{(k-1)!}$, which we get in the same way.

Merging Poisson processes

Say we have two independent Poisson processes with rates λ_1, λ_2:



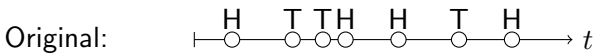
We can **merge** them to get a Poisson process with rate $\lambda_1 + \lambda_2$:

The merged process just has an arrival whenever either of the original processes does.

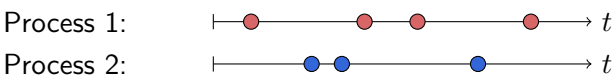
$\mathbf{N}(t) = \mathbf{N}_1(t) + \mathbf{N}_2(t)$ for the merged process, since the number of arrivals to it by time t is the number of arrivals to process 1, plus the number of arrivals to process 2.

Splitting Poisson processes

Say we have a single Poisson process with rate λ.



We can split it into two processes with rates λp and $\lambda(1-p)$. For each arrival, flip a coin to put it in the first process (with probability p) or the second (with probability $1-p$).



It is not quite intuitive, but these two processes are **independent**!

(Analogy: imagine tracking red and blue cars driving down a highway.)

Properties we can deduce

There's many properties of random variables that follow from thinking about Poisson processes.

- If $\mathbf{X} \sim \textit{Poisson}(\lambda_1)$ and $\mathbf{Y} \sim \textit{Poisson}(\lambda_2)$ and they're independent, then $\mathbf{X} + \mathbf{Y} \sim \textit{Poisson}(\lambda_1 + \lambda_2)$.

  (Count the arrivals in 1 second when we merge two processes.)

- If $\mathbf{X} \sim \textit{Exponential}(\lambda_1)$ and $\mathbf{Y} \sim \textit{Exponential}(\lambda_2)$ and they're independent, then $\min\{\mathbf{X}, \mathbf{Y}\} \sim \textit{Exponential}(\lambda_1 + \lambda_2)$.

  (Time until the first arrival when we merge two processes.)

- If $\mathbf{X} \sim \textit{Poisson}(\lambda)$ and $\mathbf{Y} \mid \mathbf{X} \sim \textit{Binomial}(\mathbf{X}, p)$, then $\mathbf{Y} \sim \textit{Poisson}(\lambda p)$.

  (Count the arrivals in 1 second when we split two processes.)

How to specify joint distributions?

Say we have two **continuous** random variables $\mathbf{X}, \mathbf{Y}$. How can we specify their joint distribution?

- We can define the **joint CDF** $F_{\mathbf{XY}}(s,t) = \Pr[\mathbf{X} \leq s \text{ and } \mathbf{Y} \leq t]$.

  This works, but isn't as helpful as it was in one dimension. We can find probabilities such as $\Pr[a \leq \mathbf{X} \leq b \text{ and } c \leq \mathbf{Y} \leq d]$ but even something basic like $\Pr[\mathbf{X} \geq \mathbf{Y}]$ is hard.

- We can define the **joint PDF** $f_{\mathbf{XY}}(s,t)$.

  This has the property that for any set $A \subseteq \mathbb{R}^2$,

$$\Pr[(\mathbf{X}, \mathbf{Y}) \in A] = \iint_A f_{\mathbf{XY}}(s,t)\, dt\, ds.$$

Summary of the joint PDF

Key facts to know about the joint PDF:

- The total probability is 1, which means

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{\mathbf{XY}}(s,t)\,dt\,ds = 1.$$

- To find the **marginal PDFs** of $\mathbf{X}$ and $\mathbf{Y}$, integrate:

$$f_{\mathbf{X}}(s) = \int_{-\infty}^{\infty} f_{\mathbf{XY}}(s,t)\,dt \qquad f_{\mathbf{Y}}(t) = \int_{-\infty}^{\infty} f_{\mathbf{XY}}(s,t)\,ds.$$

- When $\mathbf{X}$ and $\mathbf{Y}$ are independent,

$$f_{\mathbf{XY}}(s,t) = f_{\mathbf{X}}(s) f_{\mathbf{Y}}(t).$$

Example 1: Another interpretation of the PDF

Let $f(t)$ be the PDF of some distribution, and let

$$A = \{(x, y) \in \mathbb{R}^2 : 0 \leq y \leq f(x)\}$$

be the set of all points between the x-axis and the curve $y = f(x)$.

If $(\mathbf{X}, \mathbf{Y})$ is chosen uniformly from A, what is the distribution of $\mathbf{X}$?

The joint PDF here is $f_{\mathbf{XY}}(s, t) = \begin{cases} 1 & (s, t) \in A \\ 0 & \text{otherwise.} \end{cases}$

We integrate and get $f_{\mathbf{X}}(s) = \displaystyle\int_{-\infty}^{\infty} f_{\mathbf{XY}}(s, t)\, dt = \int_0^{f(s)} 1\, dt = f(s).$

This is a method for sampling from a distribution with a known PDF!

Example 2: Independence

Suppose $(\mathbf{X}, \mathbf{Y})$ have the joint PDF $f_{\mathbf{XY}}(s,t) = \frac{2}{\pi} e^{-s^2-4t^2}$. Are they independent?

Yes: a quick way to see this is that the PDF factors as $C_1 e^{-s^2}$ times $C_2 e^{-4t^2}$. We can choose C_1, C_2 so that $f_{\mathbf{X}}(s) = C_1 e^{-s^2}$ is a PDF for $\mathbf{X}$ and $f_{\mathbf{Y}}(t) = C_2 e^{-4t^2}$ is a PDF for $\mathbf{Y}$.

Another example: consider $(\mathbf{X}, \mathbf{Y})$ with the joint PDF

$$f_{\mathbf{XY}}(s,t) = \begin{cases} 8st & 0 \le s \le t \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

This looks like it factors, but the condition $s \le t$ prevents it. Two variables whose joint range is a triangle like it is here can't possibly be independent!

Functions of two variables

What do we do if we have a function $\mathbf{Z} = g(\mathbf{X}, \mathbf{Y})$ of two random variables, and want to know its distribution?

Here is a general approach, based on what we did for discrete random variables:

1. Define the set $A_u = \{(x, y) \in \mathbb{R}^2 : g(x, y) \leq u\}$.
2. Find the integral

$$\Pr[(\mathbf{X}, \mathbf{Y}) \in A_u] = \iint_{A_u} f_{\mathbf{XY}}(s, t)\, dt\, ds.$$

3. Obtain the CDF $F_{\mathbf{Z}}(u) = \Pr[(\mathbf{X}, \mathbf{Y}) \in A_u]$, which we can deal with however we like.

Exponential(λ) and *Gamma*$(2, \lambda)$

If $\mathbf{X}, \mathbf{Y} \sim$ *Exponential*(λ) are independent, what is the CDF of $\mathbf{X} + \mathbf{Y}$?

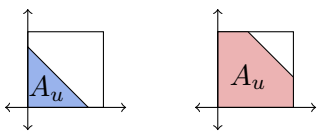
1. Find the joint PDF: $f_{\mathbf{XY}}(s,t) = f_{\mathbf{X}}(s) f_{\mathbf{Y}}(t) = \lambda^2 e^{-\lambda s - \lambda t}$ for $s, t \geq 0$.
2. Find $A_u = \{(x,y) : x \geq 0, y \geq 0, x + y \leq u\}$.
3. Integrate to find $F_{\mathbf{X+Y}}(u) = \displaystyle\int_0^u \int_0^{u-s} \lambda^2 e^{-\lambda s - \lambda t}\, dt\, ds$.
4. Obtain $F_{\mathbf{X+Y}}(u) = 1 - (1 + \lambda u) e^{-\lambda u}$: the CDF of the *Gamma*$(2, \lambda)$ distribution.

   Its derivative is $f_{\mathbf{X+Y}}(u) = \lambda^2 u e^{-\lambda u}$.

Adding up uniform distributions

If $\mathbf{X}, \mathbf{Y} \sim \textit{Uniform}(0,1)$ are independent, what is the CDF of $\mathbf{X}+\mathbf{Y}$?

1. Find the joint PDF: $f_{\mathbf{XY}}(s,t) = 1$ for $0 \le s,t \le 1$, and 0 otherwise.
2. Find $A_u = \{(x,y) : 0 \le x \le 1, 0 \le y \le 1, x+y \le u\}$.



3. Rather than integrating over A_u, $F_{\mathbf{X}+\mathbf{Y}}(u)$ can just be found as the area of A_u, since the distribution is uniform.

   This is $\frac{1}{2}u^2$ for $0 \le u \le 1$, but $1 - \frac{1}{2}(2-u)^2$ for $1 \le u \le 2$.

Finding the PDF

What if we want to find the PDF of a function of $\mathbf{X}$ and $\mathbf{Y}$?

In general, this is messy. It generalizes our one-variable rule

$$f_{h(\mathbf{X})}(t) = f_{\mathbf{X}}(h^{-1}(t)) \cdot \frac{1}{h'(h^{-1}(t))}$$

but $\frac{1}{h'(h^{-1}(t))}$ is replaced by a determinant of partial derivatives. (See Theorem 5.1 in section 5.2.4 of the textbook for details.)

The sum $\mathbf{X} + \mathbf{Y}$ is an unusually nice case. Here, we can use the rule

$$f_{\mathbf{X}+\mathbf{Y}}(t) = \int_{-\infty}^{\infty} f_{\mathbf{X}\mathbf{Y}}(s, t-s)\, ds$$

which is similar to taking a marginal PDF, but along the line $x + y = t$.

Conditional continuous distributions

If $\mathbf{X}$ is a random variable and A is an event in the same random experiment, then $\mathbf{X} \mid A$ is another random variable.

We've seen this in action with discrete random variables already.

How do we deal with $\mathbf{X} \mid A$ when $\mathbf{X}$ is a continuous random variable? There are two ways:

1. Using the CDF. The definition is:

$$F_{\mathbf{X}|A}(t) = \Pr[\mathbf{X} \leq t \mid A] = \frac{\Pr[\mathbf{X} \leq t \text{ and } A]}{\Pr[A]}.$$

   This is not always easy, but the setup always stays the same.

2. Using the PDF. We'll see this in action first for one random variable, then for a joint distribution.

An example with the CDF

Suppose a random variable $\mathbf{X}$ has

$$F_{\mathbf{X}}(t) = \begin{cases} 0 & t < -1 \\ \frac{1}{2} + \frac{1}{2}t^3 & -1 \le t < 1 \\ 1 & 1 \le t \end{cases} \qquad f_{\mathbf{X}}(t) = \begin{cases} \frac{3}{2}t^2 & -1 \le t \le 1 \\ 0 & t < -1 \text{ or } t > 1. \end{cases}$$

What is the distribution of $\mathbf{X} \mid \mathbf{X} \ge 0$?

For $0 \le t \le 1$, we have

$$F_{\mathbf{X}|\mathbf{X}\ge 0}(t) = \frac{\Pr[\mathbf{X} \le t \text{ and } \mathbf{X} \ge 0]}{\Pr[\mathbf{X} \ge 0]} = \frac{F_{\mathbf{X}}(t) - F_{\mathbf{X}}(0)}{1 - F_{\mathbf{X}}(0)} = t^3.$$

Taking the derivative, $f_{\mathbf{X}}(t) = 3t^2$ for $0 \le t \le 1$.

Using the PDF

You may have noticed that the basic shape of the PDF stayed the same throughout this process. We started with $f_{\mathbf{X}}(t) \propto t^2$ on the range of $\mathbf{X}$, and we got $f_{\mathbf{X}|\mathbf{X}\geq 0}(t) \propto t^2$ on the range of $\mathbf{X} \mid \mathbf{X} \geq 0$.

This always happens when conditioning on an event A that's **all about** $\mathbf{X}$. We simply:

1. Set the PDF to 0 everywhere that A does not occur.
2. Divide by $\Pr[A]$; equivalently, rescale so that the PDF integrates to 1.

This is an extension of how we found the PMF of $\mathbf{X} \mid A$ in such cases, for discrete random variables. We discarded values of $\mathbf{X}$ inconsistent with A, and rescaled to make the sum of the probabilities 1 again.

An example with the PDF

If $\mathbf{X} \sim \textit{Exponential}(2)$, what is $\mathbb{E}[\mathbf{X} \mid 1 \leq \mathbf{X} \leq 2]$?

We know that $f_{\mathbf{X}|1\leq\mathbf{X}\leq 2}(t)$ is going to be proportional to e^{-2t} on $[1, 2]$; specifically,

$$f_{\mathbf{X}|1\leq\mathbf{X}\leq 2}(t) = \frac{2e^{-2t}}{\Pr[1 \leq \mathbf{X} \leq 2]} = \frac{2}{e^{-2} - e^{-4}} e^{-2t}.$$

To find the expected value of a random variable $\mathbf{Y}$, we always integrate $t \cdot f_{\mathbf{Y}}(t)$. In this case,

$$\mathbb{E}[\mathbf{X} \mid 1 \leq \mathbf{X} \leq 2] = \int_1^2 t \cdot \frac{2}{e^{-2} - e^{-4}} \cdot e^{-2t}\, dt.$$

With integration by parts, we get $\mathbb{E}[\mathbf{X} \mid 1 \leq \mathbf{X} \leq 2] = \frac{5-3e^2}{2-2e^2} \approx 1.343$.

The two-variable procedure

The PDF of $\mathbf{X} \mid A$ gets trickier to find when A is not solely dependent on $\mathbf{X}$.

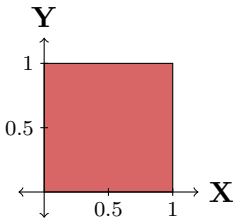
We'll just take things one step further: events A that depend on $\mathbf{X}$ and another random variable $\mathbf{Y}$.

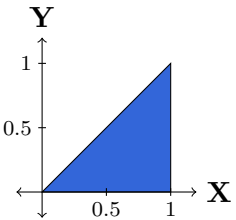
The way to solve problems like this is the same as before, but with joint distributions thrown in.

1. Describe the joint distribution of **all** the random variables involved in the event A.
2. Condition on A: throw out all the outcomes incompatible with A, and rescale (dividing by $\Pr[A]$).
3. Find the marginal distribution of $\mathbf{X} \mid A$.

Joint distributions and conditioning

Suppose $\mathbf{X}, \mathbf{Y} \sim \textit{Uniform}(0, 1)$ and $\mathbf{X}$ and $\mathbf{Y}$ are independent. What happens when we condition on $\mathbf{X} \geq \mathbf{Y}$?





The joint distribution: $(\mathbf{X}, \mathbf{Y})$ is uniformly chosen from the red square. (The joint PDF $f_{\mathbf{X},\mathbf{Y}}(s, t)$ is 1 for points (s, t) in this square.)

The conditional joint PDF $f_{\mathbf{X},\mathbf{Y}|\mathbf{X}\geq\mathbf{Y}}(s, t)$ will be constant on the blue triangle. (Specifically, $f_{\mathbf{X},\mathbf{Y}|\mathbf{X}\geq\mathbf{Y}}(s, t) = 2$ there, to integrate to 1.)

Finding a conditional distribution

Suppose $\mathbf{X}, \mathbf{Y} \sim \textit{Uniform}(0,1)$ and $\mathbf{X}$ and $\mathbf{Y}$ are independent.

What is the conditional PDF $f_{\mathbf{X}|\mathbf{X}\geq\mathbf{Y}}$?

A summary of the previous slide:

$$f_{\mathbf{X},\mathbf{Y}|\mathbf{X}\geq\mathbf{Y}}(s,t) = \begin{cases} 2 & 0 \leq t \leq s \leq 1 \text{ (in the blue triangle)} \\ 0 & \text{otherwise.} \end{cases}$$

To find the marginal probability of $\mathbf{X} \mid \mathbf{X} \geq \mathbf{Y}$, integrate away the second coordinate:

$$f_{\mathbf{X}|\mathbf{X}\geq\mathbf{Y}}(s) = \int_{-\infty}^{\infty} f_{\mathbf{X},\mathbf{Y}|\mathbf{X}\geq\mathbf{Y}}(s,t)\,dt = \int_{0}^{s} 2\,dt$$

which gives $f_{\mathbf{X}|\mathbf{X}\geq\mathbf{Y}}(s) = 2s$ for $0 \leq s \leq 1$.

The conditional density function

Discrete variables had the conditional PMF $P_{\mathbf{X}|\mathbf{Y}}$, defined as

$$P_{\mathbf{X}|\mathbf{Y}}(a,b) = \Pr[\mathbf{X} = a \mid \mathbf{Y} = b] = \frac{P_{\mathbf{XY}}(a,b)}{P_{\mathbf{Y}}(b)}.$$

For conditional random variables, we have a similar thing: the conditional PDF

$$f_{\mathbf{X}|\mathbf{Y}}(s,t) = \frac{f_{\mathbf{XY}}(s,t)}{f_{\mathbf{Y}}(t)}.$$

If everything goes well, this is useful, and we have

$$\Pr[a \le \mathbf{X} \le b \mid \mathbf{Y} = t] = \int_a^b f_{\mathbf{X}|\mathbf{Y}}(s,t)\,ds.$$

You should be worried conditioning on $\mathbf{Y} = t$, since $\Pr[\mathbf{Y} = t] = 0$. This can cause paradoxes if you're not careful.

Two binomials

In your spare time, you play ranked tic-tac-toe games online. You are matched against a random opponent and play a 5-game match.

- 50% of all opponents are beginners, and you win $\frac{2}{3}$ of the time against a beginner; in this case, you win $\mathbf{X}_1 \sim \textit{Binomial}(5, \frac{2}{3})$ games.
- The other 50% are experts, and you win $\frac{1}{3}$ of the time against an expert; in this case, you win $\mathbf{X}_2 \sim \textit{Binomial}(5, \frac{1}{3})$ games.

What is the overall distribution of $\mathbf{X}$, the number of games you win against a randomly chosen opponent?

Despite appearances, $\mathbf{X}$ is **not** $\textit{Binomial}(5, \frac{1}{2})$! You have a $\frac{1}{2}$ overall chance of winning an individual game, but the wins are positively correlated.

Mixture distributions

"$\mathbf{X}$ is *Binomial*$(5, \frac{1}{3})$ half the time, and *Binomial*$(5, \frac{2}{3})$ half the time"

1. One way to understand this setup is via joint distributions.

   "$\mathbf{P}$ is uniformly chosen from $\{\frac{1}{3}, \frac{2}{3}\}$, and $\mathbf{X} \mid \mathbf{P} \sim$ *Binomial*$(5, \mathbf{P})$"

2. Today, we will introduce mixture distributions to describe random variables like this.

   "$\mathbf{X}$ follows a **mixture distribution** of $\mathbf{X}_1 \sim$ *Binomial*$(5, \frac{1}{3})$ and $\mathbf{X}_2 \sim$ *Binomial*$(5, \frac{2}{3})$ with **weights** $w_1 = \frac{1}{2}$ and $w_2 = \frac{1}{2}$."

Using mixture distributions to describe a problem makes more sense when there's only a few cases, and no nice pattern.

For example: $\mathbf{X}_1 \sim$ *Binomial*$(10, \frac{1}{2})$ and $\mathbf{X}_2 \sim$ *Geometric*$(\frac{1}{5})$ with weights $w_1 = w_2 = \frac{1}{2}$.

Law of total probability

Recall the law of total probability: if $B_1, \ldots, B_n$ are disjoint events that form a partition of the sample space, then for any event A,

$$\Pr[A] = \sum_{i=1}^{n} \Pr[A \mid B_i] \Pr[B_i].$$

Now imagine that $\mathbf{X}$ is a mixture distribution and events $B_1, \ldots, B_n$ tell us which of the possible cases we're in.

- In the discrete case, $P_{\mathbf{X}}(k) = \sum_{i=1}^{n} P_{\mathbf{X}|B_i}(k) \Pr[B_i]$.
- In the continuous case, $F_{\mathbf{X}}(t) = \sum_{i=1}^{n} F_{\mathbf{X}|B_i}(t) \Pr[B_i]$.
- Taking the derivative, $f_{\mathbf{X}}(t) = \sum_{i=1}^{n} f_{\mathbf{X}|B_i}(t) \Pr[B_i]$.

Expected value

"$\mathbf{X}$ follows a **mixture distribution** of $\mathbf{X}_1 \sim$ *Binomial*$(5, \frac{1}{3})$ and $\mathbf{X}_2 \sim$ *Binomial*$(5, \frac{2}{3})$ with **weights** $w_1 = \frac{1}{2}$ and $w_2 = \frac{1}{2}$."

What is the expected value of $\mathbf{X}$?

1. Good but very specific answer: we win an individual game with probability $\frac{1}{2}$, so by linearity of expectation, $\mathbb{E}[\mathbf{X}] = \frac{5}{2}$.
2. Roundabout answer: by the law of total probability, $P_{\mathbf{X}}(k) = \frac{1}{2}\binom{5}{k}(\frac{1}{3})^k(\frac{2}{3})^{5-k} + \frac{1}{2}\binom{5}{k}(\frac{2}{3})^k(\frac{1}{3})^{5-k}$, and then we can use the definition of $\mathbb{E}[\mathbf{X}]$.
3. Direct answer: by the law of total expectation, $\mathbb{E}[\mathbf{X}] = w_1\mathbb{E}[\mathbf{X}_1] + w_2\mathbb{E}[\mathbf{X}_2]$, which is $\frac{1}{2} \cdot \frac{5}{3} + \frac{1}{2} \cdot \frac{10}{3} = \frac{5}{2}$.

Moments and variance

"$\mathbf{X}$ follows a **mixture distribution** of $\mathbf{X}_1 \sim$ *Binomial*$(5, \frac{1}{3})$ and $\mathbf{X}_2 \sim$ *Binomial*$(5, \frac{2}{3})$ with **weights** $w_1 = \frac{1}{2}$ and $w_2 = \frac{1}{2}$."

What is $\mathrm{Var}[\mathbf{X}]$?

- It is **not true** that we can write this as $w_1\mathrm{Var}[\mathbf{X}_1] + w_2\mathrm{Var}[\mathbf{X}_2]$!

  The true variance will always be at least as big. (Intuitively, the random choice between $\mathbf{X}_1$ and $\mathbf{X}_2$ is part of the variance.)

- We can find $\mathbb{E}[\mathbf{X}^2] = w_1\mathbb{E}[\mathbf{X}_1^2] + w_2\mathbb{E}[\mathbf{X}_2^2]$, because $\mathbf{X}^2$ is a mixture of $\mathbf{X}_1^2$ and $\mathbf{X}_2^2$ with weights w_1, w_2.

- To find $\mathrm{Var}[\mathbf{X}]$, use the formula $\mathrm{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$.

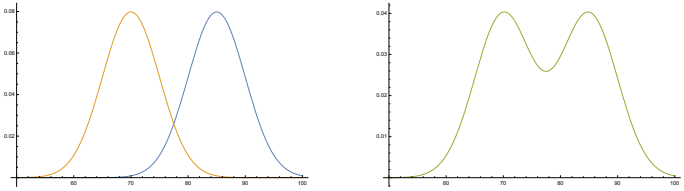
Here, $\mathbb{E}[\mathbf{X}^2] = \frac{1}{2}(\frac{35}{9} + \frac{110}{9}) = \frac{145}{18}$; $\mathrm{Var}[\mathbf{X}] = \frac{145}{18} - (\frac{5}{2})^2 = \frac{65}{36} \approx 1.81$.

Visualizing a mixture

A student that studies for an exam gets a grade $\mathbf{X}_1 \sim \textit{Normal}(85, 5^2)$.
A student that does not study gets a grade $\mathbf{X}_2 \sim \textit{Normal}(70, 5^2)$.

If these are equally likely, what is the grade distribution?

This is a mixture of $\mathbf{X}_1$ and $\mathbf{X}_2$ with weights $w_1 = w_2 = \frac{1}{2}$.



On the left are the PDFs of $\mathbf{X}_1$ and $\mathbf{X}_2$; on the right is the PDF of $\mathbf{X}$.

Note that $\mathbf{X}$ is **not** the same as $\frac{1}{2}\mathbf{X}_1 + \frac{1}{2}\mathbf{X}_2 \sim \textit{Normal}(77.5, 12.5)$.

Lifespan of a computer

You buy a new computer.

- With probability $\frac{9}{10}$, it's a high-quality model that will work for *Exponential*$(\frac{1}{5})$ years. (Average 5.)
- With probability $\frac{1}{10}$, a manufacturing defect means it will break after *Exponential*(1) years. (Average 1.)

Let $\mathbf{X}$ be the lifespan of your new computer.

As a mixture distribution: $\mathbf{X}$ is the mixture of $\mathbf{X}_1 \sim$ *Exponential*$(\frac{1}{5})$ and $\mathbf{X}_2 \sim$ *Exponential*(1) with weights $w_1 = \frac{9}{10}$ and $w_2 = \frac{1}{10}$.

This mixture distribution is a continuous distribution. It has PDF

$$f_{\mathbf{X}}(t) = \frac{9}{10}\left(\frac{1}{5}e^{-t/5}\right) + \frac{1}{10}e^{-t}.$$

Questions about the lifespan of a computer

$\mathbf{X}$ is the mixture of $\mathbf{X}_1 \sim \textit{Exponential}(\frac{1}{5})$ and $\mathbf{X}_2 \sim \textit{Exponential}(1)$ with weights $w_1 = \frac{9}{10}$ and $w_2 = \frac{1}{10}$.

Q1. What is the probability that the computer lasts more than 1 year?

A1. $\Pr[\mathbf{X}_1 \geq 1] = e^{-1/5} \approx 0.818$; $\Pr[\mathbf{X}_2 \geq 1] = e^{-1} \approx 0.368$. Taking a weighted average, we get $\Pr[\mathbf{X} \geq 1] = \frac{9}{10}e^{-1/5} + \frac{1}{10}e^{-1} \approx 0.773$.

Q2. What is the expected lifespan of the computer?

A2. $\mathbb{E}[\mathbf{X}_1] = 5$ and $\mathbb{E}[\mathbf{X}_2] = 1$, so $\mathbb{E}[\mathbf{X}] = 5 \cdot 0.9 + 1 \cdot 0.1 = 4.6$.

Q3. What is the variance?

A3. $\mathbb{E}[\mathbf{X}_1^2] = 50$ and $\mathbb{E}[\mathbf{X}_2^2] = 2$, so $\mathbb{E}[\mathbf{X}^2] = 45.2$. Then, $\mathrm{Var}[\mathbf{X}] = 45.2^2 - 4.6^2 = 24.04$.

Bonus: law of total variance

We know that $\mathrm{Var}[\mathbf{X}]$ is bigger than $w_1\mathrm{Var}[\mathbf{X}_1] + w_2\mathrm{Var}[\mathbf{X}_2]$. How much bigger is it? The law of total variance says:

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

which is hard to read and make sense of. What it means is that there are two sources of variance:

1. **The average variance of $\mathbf{X}_1$ and $\mathbf{X}_2$.** That's what $w_1\mathrm{Var}[\mathbf{X}_1] + w_2\mathrm{Var}[\mathbf{X}_2]$ is measuring, but it's not the only effect.

   In the computer example, $0.9 \cdot 25 + 0.1 \cdot 1 = 22.6$.

2. **The variance in the average of $\mathbf{X}_1$ and $\mathbf{X}_2$.**

   In the computer example, $\mathbb{E}[\mathbf{X}_1] = 5$ and $\mathbb{E}[\mathbf{X}_2] = 1$. A random variable that's 5 w.p. 0.9 and 1 w.p. 0.1 has variance 1.44.

Buying a car and Murphy's law

Here is a slightly different problem with a mixture of two distributions.

You buy a car, and:

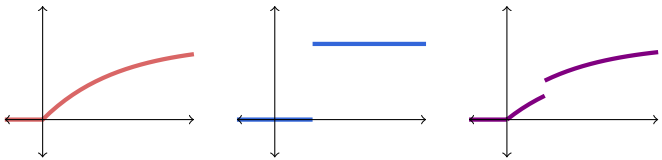
- $\frac{9}{10}$ of the time, it's a great car that lasts *Exponential*$(\frac{1}{10})$ years. (Average 10.)
- $\frac{1}{10}$ of the time, it breaks down after **exactly one year**, right after the warranty expires.

If $\mathbf{X}$ is the lifetime of the car, then $\mathbf{X}$ is a mixture of $\mathbf{X}_1 \sim$ *Exponential*$(\frac{1}{10})$ and $\mathbf{X}_2 = 1$ with weights $w_1 = \frac{9}{10}$ and $w_2 = \frac{1}{10}$.

What does the distribution of $\mathbf{X}$ look like? What kind of random variable is $\mathbf{X}$: discrete, continuous, or... what?

Finding the distribution

Let's try to find the CDF, because all random variables have a CDF.



The CDF of *Exponential*$(\frac{1}{10})$ is $1 - e^{-t/10}$ for $t \geq 0$ (and 0 for $t < 0$).

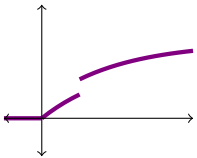
The CDF of the constant 1 is 1 for $t \geq 1$ and 0 for $t < 1$.

The CDF of the mixture is

$$F_{\mathbf{X}}(t) = \begin{cases} 0 & t < 0 \\ \frac{9}{10} - \frac{9}{10}e^{-t/10} & 0 < t \leq 1 \\ 1 - \frac{9}{10}e^{-t/10} & t \geq 1 \end{cases}$$

Discrete or continuous?

The graph of the CDF of the car's lifetime looks like this:



- This is not a continuous distribution: the CDF is not continuous.

  In particular, there's no PDF: we can't get these probabilities by integrating something.

- This is not a discrete distribution either: the range is $[0, \infty)$.

We call distributions like this **mixed distributions**.

Mixed distributions

How can we deal with mixed distributions?

1. Everything we've done that depends on the CDF still works.
2. We can split the random variable into a discrete part and a continuous part.

   In our example, the CDF is **almost** the integral of a PDF: there is just a jump at $t = 1$ we can't handle this way.
3. We can use the delta function: a fictional, nonexistent function that lets us pretend that a PDF exists.

Splitting up the random variable

Our CDF is

$$F_{\mathbf{X}}(t) = \begin{cases} 0 & t < 0 \\ \frac{9}{10} - \frac{9}{10}e^{-t/10} & 0 < t \leq 1 \\ 1 - \frac{9}{10}e^{-t/10} & t \geq 1 \end{cases}$$

Is the derivative $F'_{\mathbf{X}}(t) = \frac{9}{100}e^{-t/10}$ (for $t \geq 0$, and 0 for $t < 0$) a PDF?

- It **almost** works: we have $\Pr[a \leq \mathbf{X} \leq b] = \int_a^b \frac{9}{100}e^{-t/10}\,dt$ for any interval $[a, b]$ **not including 1**.
- For intervals containing 1, our probability will be too low by $\frac{1}{10}$, which is exactly $\Pr[\mathbf{X} = 1]$.

We can always write down a pseudo-PDF for the continuous part of the behavior, but we must handle values with positive probability separately.

The delta function

Definition. The delta function $\delta(t)$ is an object we define to have the property

$$\int_{-\infty}^{\infty} g(t)\delta(t-t_0)\,dt = g(t_0).$$

Intuition: $\delta(t)$ is 0 for $t \neq 0$, and "infinitely large" for $t = 0$ in a way that "integrates to 1".

For our random variable $\mathbf{X}$, if we define

$$f_{\mathbf{X}}(t) = \begin{cases} \frac{9}{100}e^{-t/10} + \frac{1}{10}\delta(t-1) & t \geq 0 \\ 0 & t < 0 \end{cases}$$

then $\Pr[a \leq \mathbf{X} \leq b]$ is **always** the integral of $f_{\mathbf{X}}(t)$ over $[a, b]$.

Using the delta function

What is the expected value of our random variable $\mathbf{X}$ with "PDF" $f_{\mathbf{X}}(t) = \frac{9}{100}e^{-t/10} + \frac{1}{10}\delta(t-1)$?

We can write the integral

$$\mathbb{E}[\mathbf{X}] = \int_{-\infty}^{\infty} t f_{\mathbf{X}}(t)\, dt = \int_{0}^{\infty} t\left(\tfrac{9}{100}e^{-t/10} + \tfrac{1}{10}\delta(t-1)\right) dt.$$

This is

$$\mathbb{E}[\mathbf{X}] = \int_{0}^{\infty} \tfrac{9}{100}te^{-t/10}\, dt + \int_{0}^{\infty} \tfrac{t}{10}\delta(t-1)\, dt = 9 + \frac{1}{10}.$$

"PDF" of a discrete random variable

Let $\mathbf{X} \sim \textit{Binomial}(3, \frac{1}{3})$. What is the PDF of $\mathbf{X}$?

That's a stupid question. $\mathbf{X}$ is a discrete random variable. It has a PMF given by

$$P_{\mathbf{X}}(0) = \frac{8}{27} \quad P_{\mathbf{X}}(1) = \frac{4}{9} \quad P_{\mathbf{X}}(2) = \frac{2}{9} \quad P_{\mathbf{X}}(3) = \frac{1}{27}.$$

But we can represent the same thing by a "PDF":

$$f_{\mathbf{X}}(t) = \frac{8}{27}\delta(t) + \frac{4}{9}\delta(t-1) + \frac{2}{9}\delta(t-2) + \frac{1}{27}\delta(t-3).$$

Integrating $f_{\mathbf{X}}(t)$ on $[a, b]$ picks up $\frac{8}{27}$ if $0 \in [a, b]$, plus $\frac{4}{9}$ if $1 \in [a, b]$, and so on. This is exactly what $\Pr[a \leq \mathbf{X} \leq b]$ should do.

General recipe for the delta function

In general, whenever a CDF has jumps at values $t_1, t_2, \ldots, t_k$, we can write down a "PDF" with the delta function:

$$f_{\mathbf{X}}(t) = F'_{\mathbf{X}}(t) + \sum_{i=1}^{k} \Pr[\mathbf{X} = t_i] \cdot \delta(t - t_i).$$

Any rules that involve integrating PDFs still apply when we work with this $f_{\mathbf{X}}(t)$, as long as we follow the rules for integrating the delta function.

This is more of a book-keeping tool than a magic formula.

Any time you integrate anything involving $f_{\mathbf{X}}(t)$, you'll have to deal with the δ terms separately.

Null hypothesis significance testing

There are many statistical tests that attempt to rule out some **null hypothesis** about the world/some data.

All of them have the same structure: they give a ***p*-value** between 0 and 1 as their answer.

Informally, the p-value is the probability that we'd get "data like this" if the null hypothesis is true.

- If the p-value is very low, then this suggests that the null hypothesis is false: it is not very good at explaining the data! We **reject the null hypothesis**.
- If the p-value is high, then the null hypothesis is one plausible explanation for the data. We **fail to reject the null hypothesis**.

Example

 Sir, the possibility of successfully navigating an asteroid field is approximately 3,720 to 1!

 Never tell me the odds!

(Later. . .)

 Sir, I reject the null hypothesis that we **don't** have plot armor with a p-value of approximately 0.0002688!

Guarantees about p-values

We collect some data, run a statistical analysis, and get a p-value $\mathbf{P}$.

- A statistical test **must always** make the following promise: for any "significance level" $\alpha \in [0, 1]$, $\Pr[\mathbf{P} < \alpha]$ is at most α, assuming the null hypothesis is true.

  (Typical values of α are 0.05 or 0.01, depending on field.)

- Ideally, if a statistical test is **powerful**, then there are plausible alternative hypotheses under which $\Pr[\mathbf{P} < \alpha]$ is much higher.

- Typical framework: sort possible outcomes of the experiment by how "unexpected" they are.

  For an outcome s, let $\mathbf{P}(s)$ be the probability of getting an outcome at least as "unexpected" as s.

"Type I" and "Type II" errors

A **Type I** or **Type II** error is terrible terminology that I've never learned to remember.

- NHST makes guarantees about the **false rejection** rate: the probability you'll reject the null, when it's actually true.

  If you pick a significance level of $\alpha = 0.05$, then the false rejection rate is at most 0.05.

- There is also the **false acceptance** rate: the probability you'll fail to reject the null, when it's false.

  This has to do with the power of the test, and the general framework of NHST makes no promises about it...

  ...and **also** depends on what alternate hypotheses are plausible.

A stupid example

This example just exists to make the point that power matters.

Here is a valid statistical test. We collect some data, then (ignoring the data) let $\mathbf{P} \sim \textit{Uniform}(0, 1)$.

- It **is true** that $\Pr[\mathbf{P} < \alpha]$ is at most (actually, exactly) α.

  If we run this test on many hypotheses, and reject the ones for which $\mathbf{P} < 0.05$, we will falsely reject only about 5% of our hypotheses.

- This statistical test **has no power**. Under the null hypothesis, it is unlikely that $\mathbf{P} < 0.001\ldots$ but it is equally unlikely that $\mathbf{P} < 0.001$ under any other hypothesis!

Is the coin fair?

I hand you a coin, and you want to know if it is fair or not. (Null hypothesis: the coin is fair.) Here are three tests we can run.

1. You flip the coin 100 times; say it lands heads k times.

   Taking $\mathbf{X} \sim \textit{Binomial}(100, \frac{1}{2})$ for comparison, set $\mathbf{P} = \Pr[\mathbf{X} \leq k]$.

2. You flip the coin 100 times, say it lands heads k times.

   Taking $\mathbf{X} \sim \textit{Binomial}(100, \frac{1}{2})$, set $\mathbf{P} = \Pr[|\mathbf{X} - 50| \geq |k - 50|]$.

3. You flip the coin until it lands heads 100 times; say it takes a total of n coin flips.

   Taking $\mathbf{X} \sim \textit{Pascal}(100, \frac{1}{2})$ for comparison, set $\mathbf{P} = \Pr[\mathbf{X} \geq n]$.

One-tailed vs. two-tailed tests

Let's compare tests 1 and 2. We flip the coin 100 times and it lands heads k times; we compare k to the distribution $\mathbf{X} \sim \textit{Binomial}(100, \frac{1}{2})$.

$$\mathbf{P}_1 = \Pr[\mathbf{X} \leq k] \qquad \mathbf{P}_2 = \Pr[|\mathbf{X} - 50| \geq |k - 50|].$$

1. Case 1: $k = 40$.

   In this case, $\mathbf{P}_1 = \sum_{i=0}^{40} \binom{100}{i} (\frac{1}{2})^{100} \approx 0.028$, and $\mathbf{P}_2 \approx 0.056$.

   Test 1 has more power to reject the null hypothesis here.

2. Case 2: $k = 66$.

   In this case, $\mathbf{P}_1 \approx 0.9995$, and $\mathbf{P}_2 \approx 0.0018$.

   Test 2 has more power to reject the null hypothesis here.

Same result, different experiments

Suppose you flip a coin 100 times, and it lands heads 38 times.

1. You say "I was doing test 1: I flipped the coin 100 times". You compare to *Binomial*$(100, \frac{1}{2})$, and get a p-value of $\mathbf{P}_1 \approx 0.0105$.
2. You say "I was doing test 3: I flipped the coin until it landed heads 38 times". You compare to *Pascal*$(38, \frac{1}{2})$, and get a p-value of $\mathbf{P}_3 \approx 0.0077$.

Moral 1: The p-value depends on the test. Dishonest researchers can boost their p-values by picking the test after collecting data.

Moral 2: NHST is weird and should be treated with caution beyond the promises it specifically makes.

See also: `https://xkcd.com/882/`

NHST and z-scores

Here is a fourth test to see if a coin is fair. As before, we flip the coin 100 times, say it lands heads k times.

1. Compute the value $z = \frac{k-50}{5}$.
2. Taking $\mathbf{Z} \sim \textit{Normal}(0, 1)$ for comparison, set $\mathbf{P}$ to be either $\Pr[\mathbf{Z} \leq z]$ (one-tailed) or $\Pr[|\mathbf{Z}| \geq |z|]$ (two-tailed).

Why does this make sense? Well, if $\mathbf{X} \sim \textit{Binomial}(100, \frac{1}{2})$, then $\mathbb{E}[\mathbf{X}] = 50$, and $\mathrm{SD}[\mathbf{X}] = 5$, so $\mathbf{Z}$ is a good approximation of $\frac{\mathbf{X}-50}{5}$.

Normal distributions are a good approximation for many things. Statisticians often report NHST results as a z-score which would be $\textit{Normal}(0, 1)$ if the null hypothesis is true.

Useful references: $\Pr[|\mathbf{Z}| \geq 1.96] \approx 0.05$ and $\Pr[|\mathbf{Z}| \geq 2.58] \approx 0.01$.

The Bayesian approach to coin tosses

Suppose that we do n trials of an experiment that can **succeed** or **fail**. We get k successes and $n - k$ failures.

What is the probability that the next trial will succeed?

We could model each trial as a *Bernoulli*(p) random variable, but the parameter p is unknown. So we can imagine a two-stage model:

1. Choose a probability $\mathbf{P} \sim$ *Uniform*$(0, 1)$.
2. Independently choose $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \sim$ *Bernoulli*$(\mathbf{P})$.

   (Or choose their sum $\mathbf{X} \sim$ *Binomial*$(n, \mathbf{P})$.)

Bayes' rule, again

We are given:

$$f_{\mathbf{P}}(t) = \begin{cases} 1 & 0 \le t \le 1 \\ 0 & \text{otherwise} \end{cases} \qquad P_{\mathbf{X}|\mathbf{P}}(k,t) = \binom{n}{k} t^k (1-t)^{n-k}$$

Bayes' rule for a joint continuous-discrete distribution says:

$$f_{\mathbf{P}|\mathbf{X}}(t,k) = \frac{P_{\mathbf{X}|\mathbf{P}}(k,t) f_{\mathbf{P}}(t)}{P_{\mathbf{X}}(k)}.$$

How do we find $P_{\mathbf{X}}(k)$?

1. Reasonable (and correct) guess: $P_{\mathbf{X}}(k) = \frac{1}{n+1}$ for $k = 0, 1, \ldots, n$.
2. General approach: use the condition $\int_0^1 f_{\mathbf{P}|\mathbf{X}}(t,k)\, dt = 1$.

Finding the normalizing constant

Plugging things in, we get that

$$f_{\mathbf{P}|\mathbf{X}}(t,k) \propto t^k(1-t)^{n-k}$$

for $0 \leq t \leq 1$, with a constant depending on k (and on n).

To find the constant:

- When $k = n$, $\int_0^1 t^n \, dt = \frac{1}{n+1}$, so $f_{\mathbf{P}|\mathbf{X}}(t,n) = (n+1)t^n$.
- When $k = 0$, $\int_0^1 (1-t)^n \, dt = \frac{1}{n+1}$, so $f_{\mathbf{P}|\mathbf{X}}(t,0) = (n+1)(1-t)^n$.
- In general, integration by parts says that

$$\int_0^1 t^k(1-t)^{n-k} \, dt = \frac{k}{n-k+1} \int_0^1 t^{k-1}(1-t)^{n-k+1} \, dt$$

which eventually gets us to the $k = 0$ case...

The Beta distribution

A random variable $\mathbf{P}$ has the *Beta*(a, b) distribution if

$$f_{\mathbf{P}}(t) = \begin{cases} \frac{(a+b+1)!}{a!b!} t^a (1-t)^b & 0 \le t \le 1 \\ 0 & \text{otherwise.} \end{cases}$$

The coefficient $\frac{(a+b+1)!}{a!b!}$ is the constant we want.

What we've learned: if $\mathbf{P} \sim$ *Uniform*$(0, 1)$ and for each $t \in [0, 1]$, $\mathbf{X} \mid \mathbf{P} = t \sim$ *Binomial*(n, t), then for each $k \in \{0, 1, 2, \ldots, n\}$,

$$\mathbf{P} \mid \mathbf{X} = k \sim \textit{Beta}(k, n-k).$$

As a special case, *Beta*$(0, 0)$ is the same as *Uniform*$(0, 1)$, which is very reasonable.

Is the coin fair?

Coin lands heads with probability $\mathbf{P}$. In n flips, it lands heads $\mathbf{X}$ times.

We now know that if our **prior distribution** was $\mathbf{P} \sim \mathit{Uniform}(0, 1)$, then our **posterior distribution** is $\mathbf{P} \mid \mathbf{X} = k \sim \mathit{Beta}(k, n - k)$.

What does this mean? What's the probability that the coin is fair?

It's zero: $\mathit{Beta}(k, n - k)$ is a continuous distribution. Uhh...

- Mathematically, this is not unexpected. Our prior has $\Pr[\mathbf{P} = \frac{1}{2}] = 0$. This can't be changed by any amount of evidence.
- Philosophically, how can you ever be sure that $\mathbf{P}$ is **exactly** $\frac{1}{2}$, no matter how much evidence you get?
- Practically, we can ask $\Pr[0.49 \leq \mathbf{P} \leq 0.51 \mid \mathbf{X} = k]$. (Depending on the tolerance we're willing to accept.)

Making predictions

Suppose k of n trials were successes and we do another trial $\mathbf{X}_{n+1}$. What is the probability of success (that $\mathbf{X}_{n+1} = 1$)?

By the law of total probability, $\Pr[\mathbf{X}_{n+1} = 1 \mid \mathbf{X} = k]$ is

$$\int_0^1 \Pr[\mathbf{X}_{n+1} = 1 \mid \mathbf{P} = t] f_{\mathbf{P}|\mathbf{X}=k}(t)\, dt$$

where $f_{\mathbf{P}|\mathbf{X}=k}$ is the PDF of *Beta*$(k, n-k)$, and $\Pr[\mathbf{X}_{n+1} = 1 \mid \mathbf{P} = t]$ is just t.

We get

$$\int_0^1 t \cdot \frac{(n+1)!}{k!(n-k)!} t^k (1-t)^{n-k}\, dt \ldots$$

Simplifying the integral

Evaluating this integral is hard, but there's a trick. We want to know:

$$\int_0^1 \frac{(n+1)!}{k!(n-k)!} t^{k+1}(1-t)^{n-k}\,dt.$$

If the coefficient were instead $\frac{(n+2)!}{(k+1)!(n-k)!}$, this would be the integral of the *Beta*$(k+1, n-k)$ distribution from 0 to 1, so we'd get 1. So let's rewrite the integral as

$$\frac{k+1}{n+2}\int_0^1 \frac{(n+2)!}{(k+1)!(n-k)!} t^{k+1}(1-t)^{n-k}\,dt$$

which simplifies to a prediction of $\frac{k+1}{n+2}$ that the next trial will succeed!

Sanity checking

Rule of succession. If we flip a "completely mysterious" coin n times, and it lands heads k times, the probability it lands heads on the next flip is $\frac{k+1}{n+2}$.

Let's check if this makes sense.

- If $k = \frac{n}{2}$, the rule gives a probability of $\frac{n/2+1}{n+2} = \frac{1}{2}$.
- If $k = 0$, we still get a probability of $\frac{1}{n+2}$.

  (That's an advantage over a prediction like $\frac{k}{n}$, which will think that an outcome is impossible if it's never happened before.)
- In real life, if we flip a coin once and it lands heads, we don't make a prediction of $\frac{2}{3}$.

  That's because real coins aren't "completely mysterious"!

Bayesian vs. Frequentist approach

How does this method compare to the null hypothesis approach?

Disadvantages:

- Picking a prior distribution for $\mathbf{P}$ seems arbitrary.
- The calculations can get much more complicated.

Advantages:

- The probability we get depends only on the data, not on the structure of the experiment.
- That's because we are actually computing the probability we want: $\Pr[\text{fair coin} \mid \text{data}]$, not the reverse probability $\Pr[\text{data} \mid \text{fair coin}]$.

There is much more to the story!