Basics of Probability

Alex Houtz*

July 13, 2023

Probability uses the properties of sets and functions that we have talked about so far. As we build the basics of probability, keep in mind what we have reviewed so far.

1 Building the Probability Function

Sample Spaces

In the sets lecture, we defined the universal set. In probability, we call the universal set the **sample space**, denoted by Ω. The sample space is the set of all possible outcomes, ω. We define a subset of outcomes as an **event**.

Before we move forward, recall the Sigma Algebra, $\mathcal{F}$. A Sigma Algebra is a set of subsets of Ω such that:

- $\emptyset \in \mathcal{F}$
- If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
- If $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$

Probability Measure

A **probability measure** is a function $P : \mathcal{F} \to [0, 1]$ that assigns to each event $A \in \mathcal{F}$ a real number $P(A)$, also known as a **probability**, and satisfies three axioms:

(1) $P(A) \geq 0 \qquad \forall\, A \in \mathcal{F}$

(2) $P(\Omega) = 1$

*Math Camp Instructor | University of Notre Dame

(3) If $A_1, A_2, ...$ are pairwise disjoint, then $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$

The first axiom simply states that probabilities can never be negative. The second axiom states that the probability of all the possible outcomes is 1. The last axiom says that the sum of probabilities of each part must equal the probability of the whole, as long as each part is disjoint from all the others.

The **probability space** is a triplet made up of Ω, $\mathcal{F}$, and P. We need Ω, as it lists the possible outcomes. We need P as it defines the probability number we are used to. Why do we need $\mathcal{F}$? Because $\mathcal{F}$ groups outcomes together, thus providing a richer domain than Ω.

The probability measure has 7 key properties:

(1) $P(A^c) = 1 - P(A)$

(2) $P(\emptyset) = 0$

(3) $P(A) \leq 1$

(4) If $A \subset B$, then $P(A) \leq P(B)$

(5) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

(6) $P(A \cup B) \leq P(A) + P(B)$

(7) $P(A \cap B) \geq P(A) + P(B) - 1$

Let's prove the first four properties:

Claim: $P(A^c) = 1 - P(A)$

Proof. By definition, A and A^c are disjoint and $A \cup A^c = \Omega$. We know that $P(\Omega) = 1$ by axiom 2. So:

$$\begin{aligned} P(\Omega) &= 1 \\ P(A \cup A^c) &= 1 \\ P(A) + P(A^c) &= 1 \\ P(A^c) &= 1 - P(A) \end{aligned}$$

This proves the claim to be true. ■

Claim: $P(\emptyset) = 0$

Proof. By axiom 2, $P(\Omega) = 1$. So:

$$\begin{aligned} P(\Omega \cup \emptyset) &= 1 \\ P(\Omega) + P(\emptyset) &= 1 \\ 1 + P(\emptyset) &= 1 \\ P(\emptyset) &= 0 \end{aligned}$$

This proves the claim to be true. ■

Claim: $P(A) \leq 1$

Proof. By axiom 1, $P(A^c) \geq 0$. We also know from property (1) that $P(A^c) = 1 - P(A)$. So:

$$\begin{aligned} P(A^c) &\geq 0 \\ 1 - P(A) &\geq 0 \\ 1 &\geq P(A) \end{aligned}$$

This proves the claim to be true. ■

Claim: If $A \subset B$, then $P(A) \leq P(B)$

Proof. Because $A \subset B$, $A = A \cap B$. Then:

$$\begin{aligned} P(A) &= P(A \cap B) \\ P(A) &= P(B) - P(B \cap A^c) \\ P(A) + P(B \cap A^c) &= P(B) \end{aligned}$$

Case 1: Assume $P(B \cap A^c) = 0$. Then:

$$P(B) = P(A) \qquad \Longrightarrow \qquad P(B) \geq P(A)$$

Case 2: Assume $P(B \cap A^c) > 0$. Then:

$$P(B) = P(A) + \underbrace{P(B \cap A^c)}_{>0} \qquad \Longrightarrow \qquad P(B) \geq P(A)$$

So in either case, $P(A) \leq P(B)$, proving the claim. ■

2 Calculating Probabilities

We first start with equally-likely outcomes

Theorem 1 (Principle of Equally-Likely Outcomes)**.** *If an experiment has N symmetric outcomes $a_1, \ldots, a_N$, then $P(a_i) = \frac{1}{N}$*

Example 1. Suppose we flip a fair coin. The sample space here is $\Omega = \{H, T\}$. There are two outcomes. As such, according the theorem:

$$P(H) = P(T) = \frac{1}{2}$$

We need to be careful applying this principle though. We need to make sure that we have accurately accounted the sample space.

Example 2. Suppose we flip a fair coin twice and order does not matter. The sample space here is $\Omega = \{HH, TT, TH\}$. There are three outcomes. As such, according the theorem $P(HH) = 1/3$. But what if order matters? Then $\Omega = \{HH, TT, HT, TH\}$ and $P(HH) = 1/4$.

Clearly, both $P(HH) = 1/3$ and $P(HH) = 1/4$ cannot be correct. To resolve this problem, we need to learn about independence. Before we can define independence, we need to look at conditional probability.

Conditional Probability

Say we want to find the probability that a person earns \$100,000 a year given the level of that person's education. We are now **conditioning** the probability on the level of education. To generalize, consider two events A and B. Assume that we can observe B and that it has already occurred. Therefore, we know that A only occurs in the

intersection of the two events. But because we know B has happened, we want to re-scale by the probability of B occurring. In math:

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

$P(A|B)$ is the **conditional** probability. We call $P(A)$ the **unconditional** probability.

Example 3 (Hansen 1.7)**.** Suppose we roll a fair dice. Let $A = \{1, 2, 3, 4\}$ and $B = \{4, 5, 6\}$. We want to find $P(A|B)$.

We first need $P(A \cap B)$. The only element A and B share is 4. So $A \cap B = \{4\}$. By the principle of equally likely outcomes: $P(A \cap B) = 1/6$.

We then need $P(B)$. B consists of half of the sample space. Therefore: $P(B) = 1/2$. Using the formula for conditional probability:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{1/6}{1/2} \\ &= 1/3 \end{aligned}$$

Independence

Events are **independent** if the outcome of one event does not impact the probability of the other event. Mathematically:

(a) $P(A \cap B) = P(A)P(B)$

(b) $P(A|B) = P(A)$, if $P(B) > 0$

(c) $P(B|A) = P(B)$, if $P(A) > 0$

Let's return to our example of flipping a coin twice. Each coin toss is independent. As such, in the first flip, we can either get H or T. In the second flip, we can get H or T. Combining all the outcomes gives the proper sample space: $\Omega = \{HH, TT, TH, HT\}$.

So:

$$\begin{aligned} P(\text{one } H \text{ and one } T) &= P(HT) + P(TH) \\ &= \frac{1}{4} + \frac{1}{4} \\ &= \frac{1}{2} \end{aligned}$$

Not 1/3 as applying the principle of equally-likely outcomes would say given the incorrect Ω.

Law of Total Probability

Recall the partitioning theorem:

Theorem 2 (Partitioning Theorem)**.** *If* $\{B_1, B_2, \dots\}$ *is a partition of* Ω*, then for any set* A*:*

$$A = \bigcup_{i=1}^{\infty} A \cap B_i$$

and the sets $(A \cap B_i)$ *are mutually disjoint.*

Apply the probability function to both sides:

$$P(A) = P\left(\bigcup_{i=1}^{\infty} A \cap B_i\right)$$

Using the third axiom of probability, as $A \cap B_i$ are all pairwise disjoint, gives:

$$P(A) = \sum_{i=1}^{\infty} P(A \cap B_i)$$

Now we can substitute in conditional probability:

$$P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i)$$

Theorem 3 (Law of Total Probability)**.** *If* $\{B_1, B_2, \dots\}$ *is a partition of* Ω *and* $P(B_i) > 0$*,* $\forall\, i$*, then:*

$$P(A) = \sum_{i=1}^{\infty} P(A|B_i)P(B_i)$$

Example 4 (Hansen 1.9)**.** Suppose we roll a fair dice. Let $A = \{1, 3, 5\}$ and $B_i = \{i\}$. If we apply the law of total probability:

$$\begin{aligned} P(\{1,3,5\}) &= 1 \times \frac{1}{6} + 0 + 1 \times \frac{1}{6} + 0 + 1 \times \frac{1}{6} + 0 \\ &= \frac{1}{2} \end{aligned}$$

Bayes' Rule

Recall the formula for conditional probability:

$$\begin{array}{ccccc} P(A|B) = \dfrac{P(A \cap B)}{P(B)} & & \text{or} & & P(B|A) = \dfrac{P(A \cap B)}{P(A)} \\ P(A \cap B) = P(A|B)P(B) & & & & P(A \cap B) = P(B|A)P(A) \\ \hookrightarrow & & P(A|B)P(B) = P(B|A)P(A) & & \hookleftarrow \\ & & P(A|B) = \dfrac{P(B|A)P(A)}{P(B)} & & \end{array}$$

Rewriting using the law of total probability:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$$

By using induction, we can broaden this derivation to obtain Bayes' Rule:

Theorem 4 (Bayes' Rule)**.** *Consider the probability space* $(\Omega, \mathcal{F}, P)$ *and a collection of mutually disjoint events* A_j *for* $j = 1, \dots, n$*. Then:*

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}$$

for every $A \in \mathcal{F}$ *such that* $P(A) > 0$*.*

Example 5. Suppose a rare genetic disorder affects 0.1% of the population. The test developed is 99% accurate, with a false positive rate of 2%. What is the probability

that a patient has this disorder if they test positive?

Let A be the event that the patient has the disorder and B be the event that the patient tests positive. We want to find $P(A|B)$. The problem gives us:

$$\begin{aligned} P(A) &= 0.1\% \\ P(B|A) &= 99\% \\ P(B|A^c) &= 2\% \end{aligned}$$

Using Bayes' Rule:

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \\ &= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.02 \cdot 0.999} \\ &= 4.7\% \end{aligned}$$

3 Combinatorics

No review of probability is complete without counting, permutations, and combinations.

Counting

Theorem 5 (Counting Rule)**.** *Let $\{A_1, A_2, A_3, \ldots, A_4\}$ be a set of disjoint events, with the j^{th} event having n_j elements. Then the number of possible sets that can be created by taking one element from each set is:*

$$n_1 \cdot n_2 \cdot n_3 \ldots n_j$$

Example 6. Suppose you have 6 hats, 3 watches, and 10 pairs of shoes. How many different outfits can you create?

Using the counting rule: $6 \cdot 3 \cdot 10 = 180$

Permutation

But what if we wanted to figure out how many ways we can order a set of elements? We call this the number of **permutations**. The number of permutations of a set of N objects is $N!$.

Example 7. Suppose you have 10 books to line up on a bookshelf. How many different line-ups can you create?

Here, $N = 10$, so the number of permutations is $10! = 3628800$

Sometimes we want to select a subset K from the set of N objects. We calculate the number of permutations here by:

$$\frac{N!}{(N-K)!}$$

So now suppose that you only want to line-up 4 of the 10 books. Then:

$$\begin{aligned}\frac{N!}{(N-K)!} &= \frac{10!}{(10-4)!} \\ &= \frac{10!}{6!} \\ &= 5040\end{aligned}$$

Combination

Now, what if we do not care about the order of the elements selected? Then we want the number of **combinations**. The number of combinations of a group of N objects taken K at a time is:

$$\binom{N}{K} = \frac{N!}{K!(N-K)!}$$

And now suppose that you only want to choose any group of 4 books. Then:

$$\begin{aligned}\binom{10}{4} &= \frac{10!}{4!(10-4)!} \\ &= \frac{10!}{4!(6!)} \\ &= 10 \cdot 7 \cdot 3 \\ &= 210\end{aligned}$$

Binomial Theorem

We can use combinatorics to calculate binomial expressions using the Binomial Theorem:

Theorem 6 (Binomial Theorem)**.** *For any integer $N \geq 0$:*

$$(a+b)^N = \sum_{K=0}^{\infty} \binom{N}{K} a^K b^{N-K}$$

Example 8. Suppose that $N = 3$. Calculate $(x+y)^N$.

To do so, simply plug $N = 3$ into the binomial theorem formula:

$$\begin{aligned}(x+y)^3 &= \sum_{K=0}^{3} \binom{3}{K} x^K y^{3-K} \\ &= \binom{3}{0} x^0 y^3 + \binom{3}{1} x^1 y^2 + \binom{3}{2} x^2 y^1 + \binom{3}{3} x^3 y^0 \\ &= y^3 + 3y^2 x + 3yx^2 + x^3\end{aligned}$$

Poker Hands

The classic application for combinatorics is calculating the probability of receiving a certain hand in poker. For our purposes, let's consider the game five card draw. We want to know what the probability is of drawing a four-of-a-kind on the first deal.

Let's break this down. There are 13 values in a standard card deck. We need to choose 1 of those values. Each value appears 4 times in the deck. We need to choose all 4 of those cards. Lastly, we need to choose 1 card to fill our hand. Putting this

mathematically:

$$\binom{13}{1}\binom{4}{4}\binom{48}{1} = 13 \cdot 1 \cdot 48$$

Now, this is just the number of ways we can get four-of-a-kind. To find the probability, we need to divide by the total number of hands. The total number of hands is any combination of 5 cards from the 52 card deck. Then:

$$\begin{aligned} P(\text{four-of-a-kind}) &= \frac{13 \cdot 1 \cdot 48}{\binom{52}{5}} \\ &= \frac{624}{2,598,960} \\ &\approx 0.0\% \end{aligned}$$

Counting Table

Importantly, we have assumed no replacement in the above formulae. We could certainly calculate the number of permutations or combinations with replacement. To do so, we would have 4 formulae, so I have included the following table from Hansen.

Table 1.1: Number of possible arrangments of size K from N items

	Without Replacement	With Replacement
Ordered	$\frac{N!}{(N-K)!}$	N^K
Unordered	$\binom{N}{K}$	$\binom{N+K-1}{K}$

Figure 1: From Hansen 1.12

4 Random Variables

A **random variable**, X, is a function from Ω to the real numbers:

$$X : \Omega \to \mathbb{R}$$

A random variable assigns a real number, $X(\omega)$, to each outcome $\omega \in \Omega$. Note that by this definition, X is neither random nor a variable. It is a function.

Example 9. Suppose $\Omega = \{H, T\}$. Define:

$$X = \begin{cases} 1 & \text{if } H \\ 0 & \text{if } T \end{cases}$$

We can now take $P(X = 1)$ instead of $P(H)$.

Every random variable has a **cumulative distribution function** (CDF). The CDF is a function $F_x : \mathbb{R} \to [0, 1]$ defined by:

$$F_x(x) = P(X \leq x), \ \forall x \in \mathbb{R}$$

All CDFs have four properties:

(1) $F(X)$ is non-decreasing.

(2) $\lim_{x \to -\infty} F(x) = 0$

(3) $\lim_{x \to \infty} F(x) = 1$

(4) $F(x)$ is right-continuous such that $\lim_{x \downarrow x_0} F(x) = F(x_0)$

Properties 1 and 2 come from the first axiom of probability. Property 3 comes from axiom two. Property 4 ensures that the CDF is continuous as we increase x. Figure 2 shows the CDFs of the χ^2 and normal distributions.

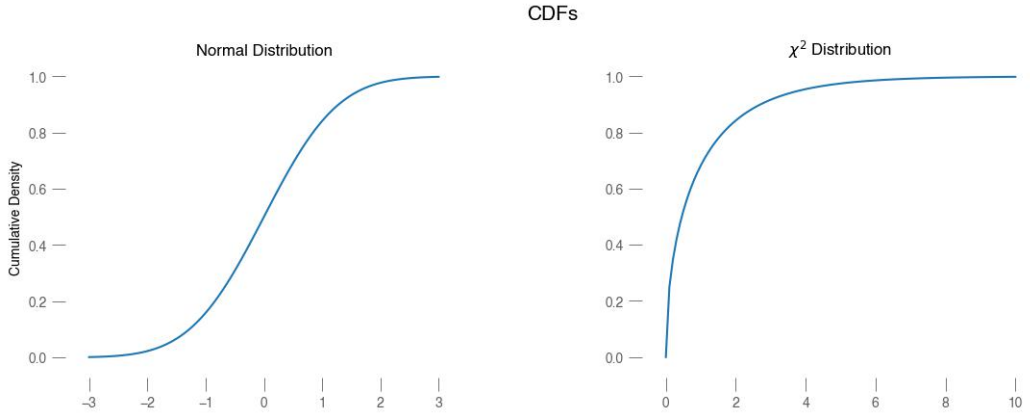

Figure 2: The left panel displays the CDF of a normal distribution. The right panel displays the CDF of a χ^2 distribution.

4.1 Types of Random Variables

There are three types of random variables: discrete, continuous, and combination. **Discrete** random variables have a number of outcomes that is finite or countably infinite. **Continuous** random variables have a number of outcomes that is uncountably infinite. **Combination** random varaibles are a mix between continuous and discrete. These do not have pmfs or pdfs.

Discrete Random Variables

An **atom** of random variable X is any real number x such that $P(X = x) > 0$. X is discrete iff X has a countable number of atoms x_i such that:

$$\sum_{x \in X} P(X = x) = 1$$

The **probability mass function** (pmf) of X is:

$$\pi(x) = f_x(x) = \begin{cases} P(X = x_i) & \text{if } x = x_i \\ 0 & \text{otherwise} \end{cases}$$

Example 10 (Bernoulli Distribution)**.** Random variable X has a **Bernoulli** distribu-

tion if:

$$f_x(x) = \begin{cases} 1-p & \text{if } x = 0 \\ p & \text{if } x = 1 \end{cases}$$

The CDF is then:

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1-p & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

Example 11 (Binomial Distribution)**.** Let $X_i \sim \text{Bernoulli}(p)$ and each X_i be *i.i.d.* Then let $Y = X_1 + X_2 + \cdots + X_n$. Y has a **binomial** distribution with a pmf of:

$$f(y|n,p) = \begin{cases} \binom{n}{y} p^y (1-p)^{n-y} & \text{if } y = 0, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Example 12 (Poisson Distribution)**.** Random variable Z has a **Poisson** distribution if the pmf is:

$$f(z|\lambda) = \frac{\lambda^z}{z!} e^{-\lambda}$$

Figure 3 shows the pmf and CDF of a Poisson distribution:

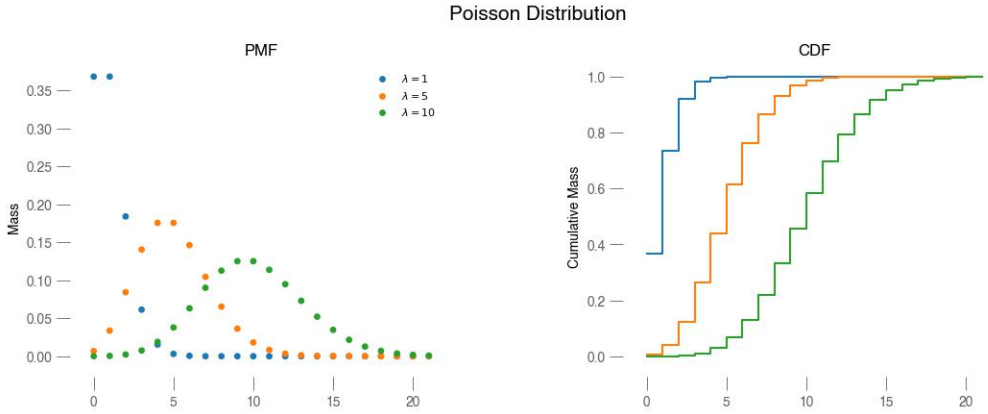

Figure 3: The left panel displays the pmf of a Poisson distribution, while the right panel displays the CDF of a Poisson distribution.

Example 13 (Discrete Uniform Distribution)**.** X has a **discrete uniform distribution** if it has pmf:

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

where n denotes the number of possible outcomes. Figure 4 displays the pmf and CDF:

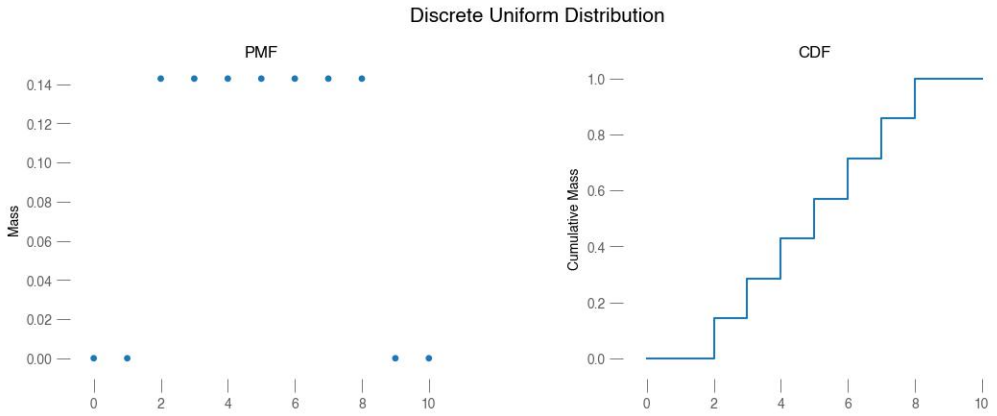

Figure 4: The left panel displays the pmf of a discrete uniform distribution, while the right panel displays the CDF.

Continuous Random Variables

A random variable $X \sim F(x)$ is continuous if its distribution function, $F(x)$, is continuous. A function, $f : \mathcal{D} \to \mathbb{R}$ is **continuous** at point $x_0 \in \mathcal{D}$ if for every $\varepsilon > 0$, there exists a $\delta > 0$ such that $\forall x \in \mathcal{D}$:

$$|x - x_0| < \delta \Rightarrow |f(x) - f(x_0)| < \varepsilon$$

To prove continuity, we use epsilon-delta proofs.

Example 14. Suppose $f(x) = \sqrt{x}$. Prove that $f : [0, \infty) \to [0, \infty)$ is continuous.

Before starting these proofs we want to solve for a δ first. We assume the conclusions

is true:

$$\begin{aligned}
\left|\sqrt{x}-\sqrt{x_0}\right| &< \varepsilon \\
\left|\sqrt{x}-\sqrt{x_0}\right|\left|\sqrt{x}+\sqrt{x_0}\right| &< \varepsilon\left|\sqrt{x}-\sqrt{x_0}\right| \\
|x-x_0| &< \varepsilon\left|\sqrt{x}+\sqrt{x_0}\right| \\
\delta &< \varepsilon\left|\sqrt{x}+\sqrt{x_0}\right|
\end{aligned}$$

With a potential δ in hand, we can start the proof.

Proof. Let $\varepsilon > 0$ and $|x - x_0| < \delta$. Pick $\delta = \varepsilon\left|\sqrt{x}+\sqrt{x_0}\right|$. Then:

$$\begin{aligned}
|x-x_0| &< \delta \\
|x-x_0| &< \varepsilon\left|\sqrt{x}+\sqrt{x_0}\right| \\
\frac{|x-x_0|}{\left|\sqrt{x}+\sqrt{x_0}\right|} &< \varepsilon \\
\left|\sqrt{x}-\sqrt{x_0}\right| &< \varepsilon \\
|f(x)-f(x_0)| &< \varepsilon
\end{aligned}$$

This is the definition of continuity, so we have proven the claim. ■

If the distribution $F(x)$ of a continuous random variable X is differentiable, then the **probability density function** (pdf) is:

$$f(x) = \frac{d}{dx}F(x)$$

We can decompose a continuous pdf into two parts, a normalizing constant and a kernel. Let $g(x) \geq 0 \ \forall \ x$ such that $\int_{-\infty}^{\infty} g(x)dx = C$. $g(x)$ is the **kernel**, while C is the **normalizing constant**. To get the pdf:

$$f(x) = \frac{1}{C}g(x)$$

When computing probabilities, we take the difference of the CDF evaluated at the boundary points:

$$P(a < X < b) = F(b) - F(a) = \int_a^b f(x)dx$$

Example 15 (Gamma Distribution)**.** X has a **gamma** distribution if it has pdf:

$$f(x|\alpha,\beta) = \frac{1}{\Gamma(\alpha)}\beta^{\alpha}x^{\alpha-1}e^{-\frac{x}{\beta}}$$

for $\alpha, \beta, x > 0$. We can rewrite the pdf as a combination of the kernel and normalizing constant:

$$C = \Gamma(\alpha)\beta^{-\alpha}$$
$$g(x) = x^{\alpha-1}e^{-\frac{x}{\beta}}$$

The Gamma distribution makes use the of gamma function:

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt$$

The gamma function has a number of special properties. Here are five of them:

(1) For positive integers n: $\Gamma(n) = (n-1)!$

(2) For $x > 1$, $\Gamma(x) = (x-1)\Gamma(x-1)$

(3) $\int_0^{\infty} t^{\alpha-1}e^{-\beta t}dt = \beta^{-\alpha}\Gamma(\alpha)$

(4) $\Gamma(1) = 1$

(5) $\Gamma(1/2) = \sqrt{\pi}$

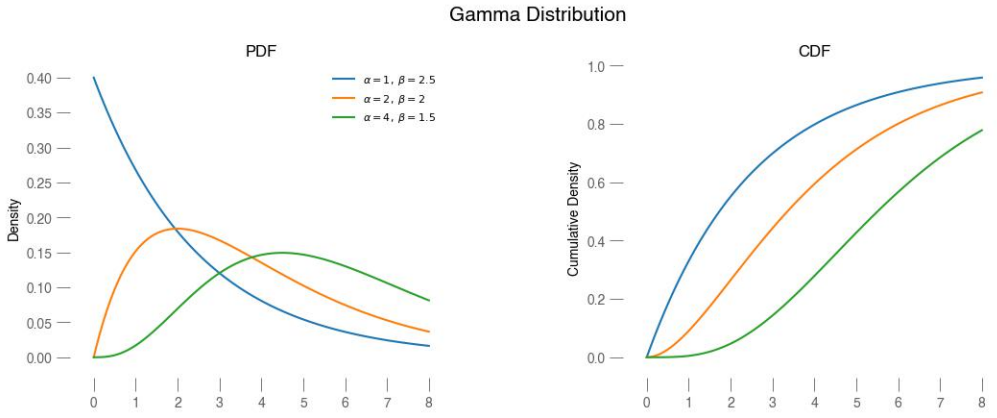

Figure 5: The left panel displays the pdf of a gamma distribution, while the right panel displays the CDF.

Figure 5 displays the pdf and CDF of the gamma distribution for different parameter combinations.

Example 16 (Normal Distribution)**.** X has a **normal** distribution if the pdf is:

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

where μ is the mean of the distribution and σ^2 is the variance. We can again rewrite the pdf as a kernal and normalizing constant:

$$C = \sqrt{2\pi\sigma^2}$$
$$g(x) = e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

Figure 6 displays the pdf and CDF for the normal distribution under different parameterizations:

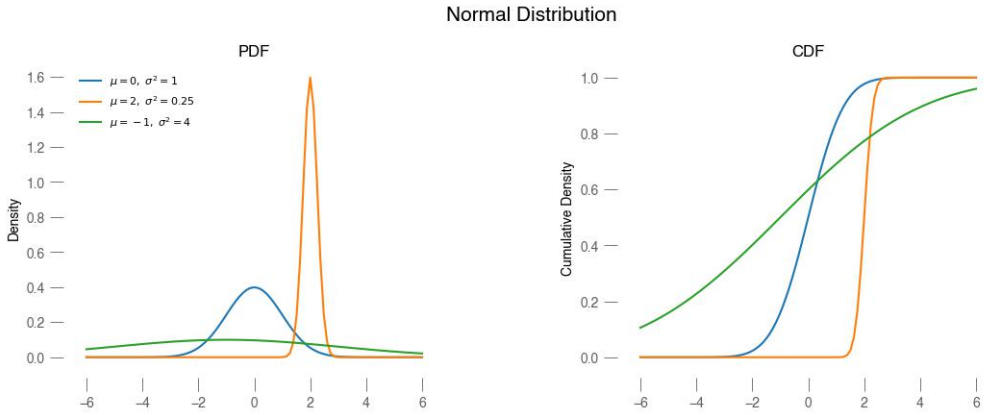

Figure 6: The left panel displays the pdf of a normal distribution, while the right panel displays the CDF.

4.2 Transformations

Sometimes we want to transform one random variable into another. We use a **transformation**, or **change-of-variables**.

Let $X \sim f_x(x)$ with support $\mathcal{X}$ and $Y = h(x)$ be a monotone function $g : \mathcal{X} \to \mathcal{Y}$. Let h^{-1} denote the inverse of h and assume that $\frac{dh^{-1}(y)}{dy}$ is continuously differentiable.

Then the pdf of Y is given by:

$$f_y(y) = \left| \frac{dh^{-1}(y)}{dy} \right| f_x(h^{-1}(y))$$

if $y \in \mathcal{Y}$ and zero otherwise.

Example 17. We want to transform a normal distribution into a log-normal distribution. Let $X \sim N(\mu, \sigma^2)$ and $Y = e^X$. First we solve for h^{-1}:

$$\begin{aligned} Y &= e^x = h(x) \\ h^{-1}(Y) &= \ln(Y) = X \end{aligned}$$

Next, we look at this function for some $y \in \mathcal{Y}$ and take the derivative with respect to y:

$$\begin{aligned} \frac{dh^{-1}(y)}{dy} &= \frac{dx}{dy} \\ &= \frac{1}{y} \end{aligned}$$

Plug these both into the transformation formula:

$$f_y(y) = \left| \frac{1}{y} \right| f_x(\ln(y))$$

We know the pdf of a normal distribution, let's plug that in:

$$f_y(y) = \frac{1}{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(\ln(y)-\mu)^2}$$

4.3 Moments of Random Variables

The **expected value** is the mean of the distribution, defined as:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{x \in \mathcal{X}} x f(x) \\ \mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x) dx \end{aligned}$$

More generally, we can write the expected value as:

$$\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$$

as sometimes we want, for example, $\mathbb{E}[x^2]$.

The **variance** of the spread of a distribution. It is calculated as:

$$Var(X) = \sum_{x \in \mathcal{X}} (x - \mu)^2 f(x)$$
$$Var(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx$$

More simply, we can write the variance as $\mathbb{E}[(X - \mu)^2]$. But maybe we want to take a linear transformation of our random variable.

Theorem 7 (Moments of a Linear Transformation)**.** *Let $Y = a + bX$. Then:*

$$\mathbb{E}[Y] = a + b\mathbb{E}[X]$$
$$Var(Y) = b^2 Var(X)$$

This can be proven using the definition of expected value.

While the definition of variance can be useful in certain situations, we often want a simpler formula for common use:

Theorem 8 (Variance Decomposition)**.** *Let X be a random variable and assume that all expectations exist. Then:*

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

Proof. Assuming that X is a random variable with proper expectations, we start with definition of variance:

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

This proves the claim to be true. ■

However, variance is usually in units that do not make much sense. Suppose we are looking at US income per household. US income is measured in dollars. But that means the variance is measured in dollars squared. To get units that make sense, we take the square root of the variance to get the **standard deviation**:

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)}$$