Random Variables

PSE 2.1-2.3

It will often be convenient to have a numerical representation of random outcomes. This is the idea of a random variable — in fact, throughout the rest of the semester, the "fundamental object" that we'll be working with are random variables.

A formal definition of a **random variable** is that it is a function from the sample space S to the real numbers $\mathbb{R}$. We'll typically denote random variables by capital letters such as X and Y. Based on the definition above, for some $s \in S$, we can write $X = X(s)$ (i.e., the random variable is a function of an element of the sample space) though we will typically drop the explicit dependence on s.

Sometimes the random outcomes are already numeric (e.g., the outcome of rolling a die); other times, we will make some (typically simple) conversion from outcomes to numeric values. For example, for flipping a coin, $S = \{H, T\}$, which is not numeric, but it might be natural to define a random variable as:

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

in other words, we define X to be equal to 1 if a heads is flipped and X to be equal to 0 if a tails is flipped. Let's do a non-trivial example next.

Example: Let X = the number of heads in 3 coin flips. Here $S = \{HHH, HHT, HTH, ...\}$ and $X = X(s)$, so, for example, $X(\{HHH\}) = 3$ (i.e., there are three heads flipped when you flip HHH).

We can also define probabilities for random variables. Let us do this formally for now (though we'll typically simplify the notation once we get used to this): $\mathrm{P}(X = x) = \mathrm{P}(\{s \in S : X(s) = x\})$. In words: this is the probability of any outcome in the sample space occurring such that $X(s)$ takes the value x (as a side-comment: we will typically use lowercase letters like x to denote particular values that a random variable could take).

This all may seem unduly technical and perhaps somewhat tedious, but I think it is helpful to briefly relate this to the previous example of flipping three coins. In that case, $\mathrm{P}(X = 3) = \mathrm{P}(\{HHH\}) = 1/8$ (as all combinations of flips are equally likely and there are 8 possible combinations). But $\mathrm{P}(X = 2) = \mathrm{P}(\{HHT, HTH, THH\}) = 3/8$ (as these are all combinations of flips that result in exactly two heads).

Next, we'll start to study the properties of random variables.

The **support** of a random variable X is the set of possible values that it can take. We will use the notation $\mathcal{X}$ for the support of the random variable X.

It will be helpful to distinguish between **discrete random variables** and **continuous random variables**. A set $\mathcal{X}$ is said to be discrete if it has a finite or countably infinite number of elements.

A formal definition of a discrete random variable is one such that there exists a discrete set $\mathcal{X}$ such that $\mathrm{P}(X \in \mathcal{X}) = 1$. For example, the support of a roll of a die is discrete, $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.

The distribution of a random variable

PSE 2.3, 2.7-2.8

Random variables have distributions. What this means is that, although random variables are (well...) random, that isn't as strong as saying that all possible outcomes are equally likely. For example, in our example of flipping three coins, although X (the number of heads) is random, not all values of X are equally likely. Intuitively, you can think of the distribution of a random variable as containing the information about how likely different outcomes are. There are two main ways to fully summarize the distribution of a random variable. We will start with the **cumulative distribution function** (cdf) and then move to the **probability density function** and **probability mass function**.

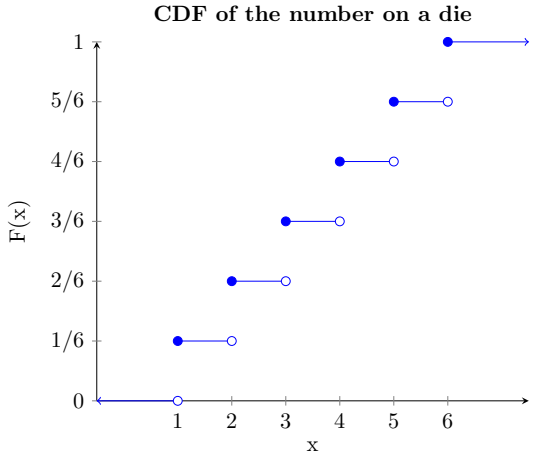
Definition. The cumulative distribution function of X is the function $\mathrm{F}_X(x) := \mathrm{P}(X \leq x)$ (where ":=" indicates "is defined as").

Note that we could more formally define $\mathrm{F}_X(x) = \mathrm{P}(\{s \in S : X(s) \leq x\})$, but we are moving towards the simpler and more common notation used above. Next, notice that $\mathrm{F}_X(x)$ is a function from $\mathbb{R}$ to the unit interval $[0, 1]$, that is $\mathrm{F}_X(x)$ is defined for all $x \in \mathbb{R}$ and $\mathrm{F}_X(x)$ is always between zero and one. The following properties can be proven to hold for any random variable X:

- $F_X(x)$ is a weakly increasing function; that is $F_X(x') \geq F_X(x)$ if $x' > x$
- $\lim_{x \downarrow -\infty} \mathrm{F}_X(x) = 0$
- $\lim_{x \uparrow \infty} \mathrm{F}_X(x) = 1$
- $\mathrm{F}_X(x)$ is right-continuous, i.e. $\mathrm{F}_X(x) = \lim_{\epsilon \downarrow 0} \mathrm{F}_X(x + \epsilon)$

As a side-comment, when the context is clear, we often denote a cdf as $\mathrm{F}(x)$ rather than $F_X(x)$ (the subscript indicates which particular random variable's cdf we are writing). However, when we have multiple random variables like X and Y, we may need the notation $F_X(x)$ and $F_Y(y)$ to be clear about which variable we are referring to.

Here is an example, it is the cdf of a six-sided die:



A few things to notice: first, the open/closed dots at e.g. $x = 1$ indicate the F(1) is equal to 1/6, and not 0 (although it is equal to 0 for x arbitrarily close but to the left of 1). We see from this graph why cdfs are right-continuous but not necessarily left-continuous. Second, at each point in its support $\{1, 2, 3, 4, 5, 6\}$, the cdf for the die jumps by $P(X = x)$, or 1/6. "Jumps" are typical features of cdfs of discrete random variables (though, in other cases, (e.g., a person's years of education), the jumps would be unlikely to be the same size). When X is a discrete random variable, its cdf ends up looking like a staircase: flat everywhere except at each x in its support, where it "jumps" up by an amount $\mathrm{P}(X = x)$.

The cdf provides all information about the distribution of a random variable. And, in particular, from the cdf, we can derive anything we'll need to know about a single random variable. An example along this line is that, given that we know the cdf of a random variable X, we can recover the probability that X would take a value in a particular range as in the following proposition.

Proposition: For any numbers a and b such that $b \geq a$, $\mathrm{P}(a < X \leq b) = \mathrm{F}(b) - F(a)$, (note: the distinction between $<$ and $\leq$ can matter especially when X is discrete). *Proof:* First, notice that

$$\mathrm{P}(a < X \leq b) = 1 - \mathrm{P}(X \leq a \text{ or } X > b)$$

Next, because the sets $\{x \in \mathbb{R} : x \leq a\}$ and $\{x \in \mathbb{R} : x > b\}$ are disjoint, we have that $\mathrm{P}(X \leq a \text{ or } X > b) = \mathrm{P}(X \leq a) + \mathrm{P}(X > b)$. Thus:

$$\mathrm{P}(a < X \leq b) = 1 - \Big(\mathrm{P}(X \leq a) + \mathrm{P}(X > b)\Big) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a) = \mathrm{F}(b) - \mathrm{F}(a)$$

where the second equality used that $\mathrm{P}(X \leq b) = 1 - \mathrm{P}(X > b)$. This completes the proof.

Earlier, we briefly defined discrete random variables. Let's formally define discrete and continuous random variables now in terms of cdfs now.

Definition. A random variable X is **continuous** if $F_X(x)$ is a continuous function of x. X is discrete if $F_X(x)$ is a step-function of x.

It is worth briefly mentioning that not all random variables are either exactly continuous or discrete; and this is especially true in economics. One example is a person's earnings (for example, a non-trivial fraction of people have exactly 0 earnings though is arguably otherwise continuous).

Probability mass and density functions

Next, we'll consider alternative ways to fully describe the distribution of a random variable; these will be more convenient in some cases and can sometimes be more intuitive. For this part, we'll provide separate treatments for discrete and continuous random variables.

Case 1: Discrete random variables and the probability mass function

Definition. The **probability mass function** (pmf) of a discrete random variable X is the function $\pi(x) = \mathrm{P}(X = x)$

This is a very natural way to describe the distribution of a discrete random variable. Like the cdf, it provides a complete summary of the distribution of a random variable. And, in particular, given the cdf we can recover the pmf, and vice versa.

Let's do that now. First, a little additional notation. For a discrete random variable, we can express the pmf alternatively as a sequence, rather than a function. Label the points in the support of X as $\{x_1, x_2, x_3, \dots\}$, in increasing order so that $x_1 < x_2 < x_3 \dots$. Let x_j denote the j^{th} value in this sequence. For any j, let $\pi_j = \pi(x_j) = \mathrm{P}(X = x_j)$.

- *Obtaining the pmf from the cdf:* For a given support point x_j: $\pi_j = F(x_j) - F(x_{j-1})$, and $\pi(x) = 0$ otherwise. In words, this is the height of the "jump" as you move from x_{j-1} to x_j.
- *Obtaining the cdf from the pmf:* $F(x) = \sum_{j:x_j \leq x} \pi_j$. In words, this just adds up the pmfs for all values less than or equal to x.

A useful thing to note about pmfs (and one that follows from the previous expression), is that, since $\lim_{x\to\infty} F(x) = 1$, we must have that $\sum_j \pi_j = 1$; that is, probability mass functions sum to one when the sum is taken across all support points j.

Case 2: Continuous random variables and the probability density function

The pmf above is, in my view, the way that is natural for people to think about the distribution of a random variables. And, although it works great for discrete random variables, defining a similar notion for continuous random variables is trickier. The reason for this is that, if X really is continuous, then $\mathrm{P}(X = x) = 0$ for any particular value x; that is, the probability that X takes

exactly the value x is equal to 0. To deal with this complication, we'll instead define the **probability density function**:

Definition. The **probability density function** (pdf) of a continuous random variable X is given by $f_X(x) = \frac{d}{dx}\mathrm{F}_X(x)$.

As a side-comment, we are using the slightly stronger requirement for a continuous random variable here that its cdf is differentiable (recall that for a function to be differentiable, it must be continuous; thus, the cdf of a continuous random variable must be continuous, lacking any jumps like those that characterize the cdf of a discrete random variable, but it does not necessarily have to be differentiable). However, we'll ignore this complication here.

Like the cdf (both continuous and discrete random variables) and the pmf (for discrete random variables), the pdf fully describes the distribution of a random variable. The above definition of the pdf describes how to convert from a cdf to a pdf for a continuous random variable. The definition also suggests how to recover the cdf from the pdf. In particular,

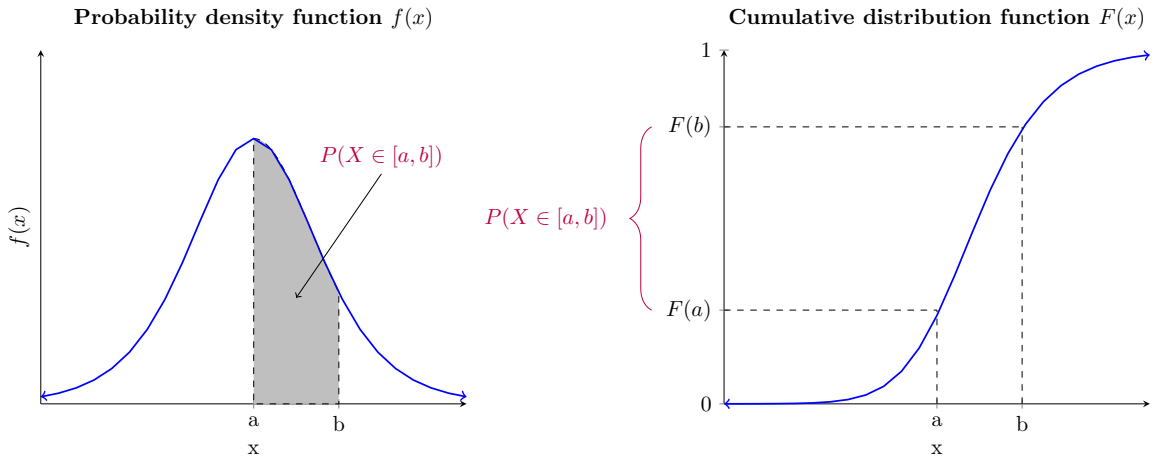
$$\mathrm{F}_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$$

In fact, this is sometimes taken as the definition of the pdf. The two statements are equivalent due to the fundamental theorem of calculus (recall: the fundamental theorem of calculus is what connects the operations of differentiation and integration). Moreover, it is worth mentioning the similarities between converting between pmfs and cdfs; notice that, essentially, we are just replacing integrals and summations, and deriviates and finite differences.

It immediately follows that the probability that X lies in any interval $[a, b]$ can be obtained by integrating over the density function:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$$

Intuitively, this gives us the area under the curve $f(x)$ between points a and b, as depicted below. Note that $\int_a^b f(x)\,dx = \mathrm{F}(b) - \mathrm{F}(a)$, because the cdf is the anti-derivative of the pdf.



In the figure, the left panel depicts an example of the pdf $f(x)$ of a random variable X. The probability that $a \leq X \leq b$ is given by the area under the $f(x)$ curve between $x = a$ and $x = b$. $\mathrm{P}(a \leq X \leq b)$ is also equal to $\mathrm{F}(b) - \mathrm{F}(a)$, the difference in the cdf of X evaluated at $x = b$ and at $x = a$, as depicted in the right panel.

The above result is also useful for interpreting the "shape" of a plot of the pdf. Recall that, for a continuous random variable, $\mathrm{P}(X = x) = 0$ for a continuous random variable; rather $f(x)$ can be interpreted as telling us the probability that X is close to x, in the following sense. Consider a point x and some small $\epsilon > 0$. Recall the definition of $f(x)$ as the derivative of $F(x)$:

$$f(x) = \frac{d}{dx}\mathrm{F}(x) = \lim_{\epsilon \to 0} \frac{\mathrm{F}(x+\epsilon) - \mathrm{F}(x)}{\epsilon} = \lim_{\epsilon \to 0} \frac{\mathrm{P}(x \leq X \leq x+\epsilon)}{\epsilon}$$

Thus $f(x)$ is limit of the ratio of the probability that X lies in a small interval that begins at x, and the width ϵ of that interval. Thus, the pdf will tend to be larger for values of x that are "relatively more common" and the pdf will be smaller for values of x that are "relatively less common". This intuition is similar to that of the pmf for discrete random variables.

Additional Properties of pmfs:

1. $\int_{-\infty}^{\infty} f(x)dx = 1$. In words: pdfs integrate to 1. This holds because $\int_{-\infty}^{\infty} f(x)dx = \mathrm{F}(\infty) - \mathrm{F}(-\infty) = 1 - 0 = 1$. This property is analogous to pmfs summing up to 1.

2. $f(x)$ is *positive* everywhere; that is $f(x) \geq 0$ for all x. This holds because $\mathrm{F}_X(x)$ is increasing and $f_X(x)$ is its derivative. This property is analogous to pmfs being positive.

Expected value

PSE 2.5, 2.13, 2.16-2.17

While cdfs, pdfs, and pmfs fully describe the distribution of a random variable, they can be complicated to report (especially in cases that we'll be interested in soon where there are multiple random variables). An alternative summary of the distribution of a random variable is its expectation. The **expectation** of a random variable is its average value and is denoted by $\mathbb{E}[X]$. The expected value is a measure of the central tendency of a random variable (other measures of central tendency are the median and mode, but we will stick to the expectations mainly). Unlike cdfs, pdfs, and pmfs, the expected value is not a fully summary of the distribution of a random variable, but the tradeoff here is that it is a single number (and hence easier to report / comprehend). The expectation is also arguably the most important single summary measure of a distribution (that is, if you could only have one number to summarize the distribution of a random variable, most people would choose to know its expectation (i.e., mean)).

To motivate how $\mathbb{E}[X]$ will be defined, think of task of computing the average of a list of numbers. For example, the average of the numbers 1, 2, 2, and 4 is $(1 + 2 + 2 + 4)/4 = 2$. Notice that the number 2 occurred twice in the series, so we added 2 to the sum two times. We could thus have

written the averaging calculation as $\frac{1}{4}(1 \cdot 1 + 2 \cdot 2 + 4 \cdot 1)$, where each number is multiplied by the number of times it occurs in the list. The general formula could be written

$$\text{average of a list of numbers} = \sum_j (j^{th} \text{ distinct number }) \cdot \underbrace{\frac{\#\text{ times } j^{th} \text{ distinct number occurs in the list}}{\text{length of the list}}}_{w_j}$$

where notice that "weight" w_j on the j^{th} distinct number sums to one over all j, i.e. $\sum_j w_j = 1$.

The definition of $\mathbb{E}[X]$ for a discrete random variable is exactly analogous to this formula, where we average over the values x that X can take, and use as "weights" the probabilities $\mathrm{P}(X = x)$:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x \cdot \pi(x)$$

where $\pi(x) = \mathrm{P}(X = x)$ is the pmf and recall that the $\pi(x)$ sum to one.

When X is continuous, the expectation is defined by replacing the sum with an integral and the pmf with the pdf; that is:

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

When working with expectations of random variables, we have to be somewhat careful to make sure that the expectation actually *exists*, i.e., that it is finite. An expectation, $\mathbb{E}[X]$, is said to exist if $\mathbb{E}|X| < \infty$ (where $|\cdot|$ is absolute value). For the most part, it is reasonable to think that expectations should exist, though there are some cases where an expectation may not exist; these would arise in cases where the pdf does not "go to zero" fast enough (which amounts to extremely large values of X being common enough that the mean may not be finite). A classic example is when X follows a Cauchy distribution, which has a pdf rather like a normal distribution but with "fatter" tails. Another example with of a discrete random variable that does not have a finite expectation is provided in PSE 2.6.

Example (Rolling a die): Note that, in this case,

$$f_X(x) = \begin{cases} \frac{1}{6} & \text{if } x = 1, 2, 3, 4, 5, 6 \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$\mathbb{E}[X] = \sum_{j=1}^{6} j f_X(j) = \sum_{j=1}^{6} j\Big(\frac{1}{6}\Big) = 3.5$$

Example (Exponential random variable): Suppose X follows an exponential distribution with parameter λ, written $X \sim \exp(\lambda)$; in other words, X is continuously distributed with $f_X(x) = \frac{1}{\lambda}\exp(-x/\lambda)$ for $0 \leq x < \infty$ and $\lambda > 0$ (you can verify that this is a valid pdf, but we'll just

calculate the expectation here). Then,

$$\begin{aligned}\mathbb{E}[X] &= \int_{\mathcal{X}} x f_X(x)\, dx \\ &= \int_0^\infty x \frac{1}{\lambda} \exp(-x/\lambda)\, dx\end{aligned}$$

We will use integration by parts here; i.e., $\int u\, dv = uv - \int v\, du$. In our case, we will set $u = x$ and $dv = \frac{1}{\lambda}\exp(-x/\lambda)\, dx$ and, thus, $du = dx$ and $v = -\exp(-x/\lambda)$. Then,

$$\begin{aligned}\mathbb{E}[X] &= -x\exp(-x/\lambda)\Big|_0^\infty + \int_0^\infty \exp(-x/\lambda)\, dx \\ &= \lim_{x\to\infty} -\frac{x}{\exp(x/\lambda)} + 0 - \lambda\exp(-x/\lambda)\Big|_0^\infty \\ &= \underbrace{\lim_{x\to\infty} \frac{-1}{\frac{1}{\lambda}\exp(x/\lambda)}}_{=0} - \lambda\cdot 0 + \lambda = \lambda\end{aligned}$$

where, the third equality uses L'Hopital's rule. Thus, when X follows an exponential distribution, its mean is equal to λ.

Here are two practice questions:

- Consider a Bernoulli random variable X which is one that takes a value 1 with probability p and 0 with probability $1 - p$. Find $\mathbb{E}[X]$.
- Consider a uniform$[0, 1]$ random variable, that is a continuous random variable with density $f(x) = x$ for all $0 \le x \le 1$, and $f(x) = 0$ everywhere else. Find $\mathbb{E}[X]$.

A key property of the expectation operator that is very useful is that it is *linear*. It's actually "linear" in a few distinct senses, but for now we'll mention one version of linearity of expectations: for two constants a and b, $\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$. This holds for both discrete and continuous random variables (and even more general random variables), but we'll provide the proof for the case where X is continuous — it basically follows from properties of integrals.

Proof:

$$\begin{aligned}
\mathbb{E}[a+bX] &= \int_{\mathcal{X}} (a+bx) f_X(x)\, dx \\
&= \int_{\mathcal{X}} (a+bx) f_X(x)\, dx \\
&= \int_{\mathcal{X}} a f_X(x)\, dx + \int_{\mathcal{X}} bx f_X(x)\, dx \\
&= a \underbrace{\int_{\mathcal{X}} f_X(x)\, dx}_{=1} + b \underbrace{\int_{\mathcal{X}} x f_X(x)\, dx}_{=\mathbb{E}[X]} \\
&= a + b\mathbb{E}[X]
\end{aligned}$$

This is a very useful property and essentially says that you can move constants outside of expectations.

Besides the mean, probably the next most useful feature of the distribution of a random variable is its variance. The **variance** of a random variable is a measure of its "spread", and it is defined as $\text{var}(X) := \mathbb{E}[(X - \mathbb{E}[X])^2]$. You can think of this as the average "distance" between X and its expectation; in particular, taking the difference between two numbers and squaring it is one of the most common measures of distance between them and we will use this frequently.

Practice: Use the linearity of the expectation operator to prove the following (very useful) alternative expression for the variance: $\text{var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

The variance of a random variable shows up naturally in many expressions in statistics. However, it is somewhat difficult to interpret as it is "squared units" — that is, if X is a person's income in dollars, then its variance is in dollars squared, which is likely to be unfamiliar and hard to interpret. Therefore, it is fairly common to instead report the **standard deviation** of a random variable which is defined as $\text{s.d.}(X) = \sqrt{\text{var}(X)}$.

A useful property of variance is the following

$$\text{var}(a+bX) = b^2 \text{var}(X)$$

This property says that the variance does not change due to additive shifts, and that constants can come out of the variance but ought to be squared first.

Proof:

$$\begin{aligned}
\text{var}(a+bX) &= \mathbb{E}\Big[\big((a+bX)-\mathbb{E}[a+bX]\big)^2\Big] \\
&= \mathbb{E}\Big[\big(b(X-\mathbb{E}[X])\big)^2\Big] \\
&= \mathbb{E}[b^2(X-\mathbb{E}[X])^2] \\
&= b^2\mathbb{E}[(X-\mathbb{E}[X])^2] \\
&= b^2\text{var}(X)
\end{aligned}$$

where the first equality comes from the definition of variance, the second equality cancels a and rearranges, the third equality squares the entire inside term, the fourth equality pulls the constant b^2 outside of the expectation, and the last equality holds from the definition of variance.

In some cases, expectations of other functions of X could be of interest. The **mth moment** of X is defined as $\mathbb{E}[X^m]$. The **mth central moment** of X is defined as $\mathbb{E}[(X-\mathbb{E}[X])^m]$.

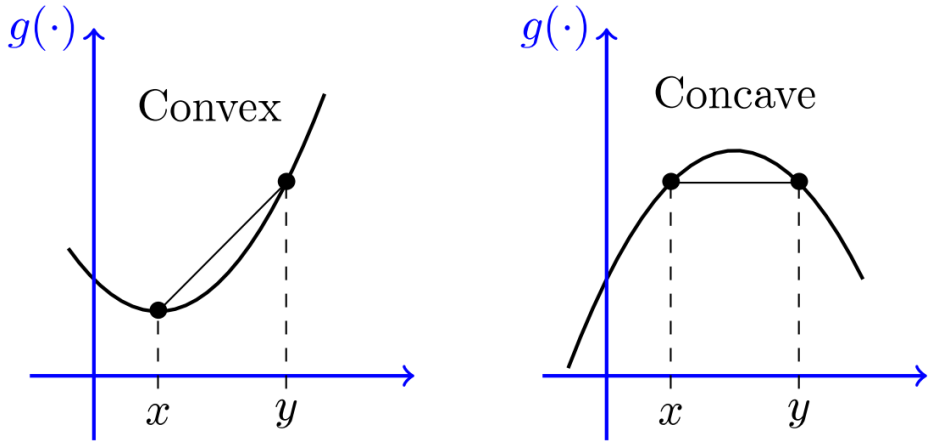
Some useful inequalities

PSE 2.18-2.19 and 7.4

Next, I want to briefly mention Jensen's inequality, which is useful for showing a number of results that will be useful for us later in the semester. Recall that a function g is said to be **convex** if for any $\lambda \in [0,1]$ and for all x and y, $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$. Similarly, a function g is said to be **concave** if $g(\lambda x + (1-\lambda)y) \geq \lambda g(x) + (1-\lambda)g(y)$. Intuitively, convex functions are "cup" shaped $\cup$; examples include $g(x) = x^2$ and $g(x) = \exp(x)$. On the other hand, concave functions are "cap" shaped $\cap$; examples include $g(x) = x^{1/2}$ and $g(x) = \log(x)$.

Jensen's Inequality. For any random variable X, if $g(x)$ is a convex function, then $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$. If $g(x)$ is a concave function, then $g(\mathbb{E}[X]) \geq \mathbb{E}[g(X)]$.

There is a proof of Jensen's inequality in Theorem 2.9 in the textbook. It is not too complicated, but I am going to skip the proof and instead just explain the intuition for the case where g is convex. $g(\mathbb{E}[X])$ amounts to evaluating the function at the average value of X (a point "in the middle" of the possible values of X); while $\mathbb{E}[g(X)]$ amounts to computing $g(x)$ at all values of X and then averaging. Because the function is convex, evaluating $g(x)$ at a middle point tends to be smaller than averaging the function (with weights given by the value of the pdf) across all possible values of X.



One useful implication of Jensen's inequality is that $|\mathbb{E}[X]| \leq \mathbb{E}|X|$. This holds because $|x|$ is a convex function and then by applying Jensen's inequality. The textbook calls this the **expectation inequality**.

Another useful implication of Jensen's inequality is **Lyapunov's inequality**. This says that, for any random variable X and any $0 < r \leq p$, $\left(\mathbb{E}|X|^r\right)^{1/r} \leq \left(\mathbb{E}|X|^p\right)^{1/p}$. Just to be clear on the notation: for example, $\mathbb{E}|X|^r := \mathbb{E}[|X|^r]$ rather than $\mathbb{E}[|X|]^r$. (and, if you want, you can add some extra brackets to make the notation more clear)

Proof: Take $g(x) = x^{p/r}$ which is a convex function for $x > 0$ because $p \geq r$. Let $Y = |X|^r$; then, Jensens' inequality implies that

$$g(\mathbb{E}[Y]) \leq \mathbb{E}[g(Y)] \iff \left(\mathbb{E}|X|^r\right)^{p/r} \leq \mathbb{E}|X|^{rp/r} = \mathbb{E}|X|^p$$

Raising both sides of the inequality on the right hand side to the $1/p$ power completes the proof.

An implication of Lyapunov's inequality is that, if we know that a higher order moment exists (say, $\mathbb{E}[X^2]$), it implies that lower order moments also exist (say, $\mathbb{E}[X]$). To see this, notice that, for $0 < r \leq p$, $\mathbb{E}|X|^r = \left(\left(\mathbb{E}|X|^r\right)^{1/r}\right)^r \leq \left(\left(\mathbb{E}|X|^p\right)^{1/p}\right)^r = \left(\mathbb{E}|X|^p\right)^{r/p} \leq \mathbb{E}|X|^p \leq \infty$ where the first inequality comes from Lyapunov's inequality, the second inequality holds because $r/p \leq 1$, and the last inequality holds when $\mathbb{E}[X^p]$ exists.

The next inequalities are Markov's inequality and Chebyshev's inequality. These will be useful in proving the law of large numbers later in the semester.

Markov's Inequality. Suppose X is a non-negative random variable, then for any $a > 0$, $\mathrm{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}$.

Proof:

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f_X(x)\,dx \\
&= \int_0^a x f_X(x)\,dx + \int_a^\infty x f_X(x)\,dx \\
&\geq \int_a^\infty x f_X(x)\,dx \\
&\geq a \int_a^\infty f_X(x)\,dx \\
&= a(1 - F_X(a))
\end{aligned}$$

where the first equality holds because X is non-negative, the second equality holds by splitting the integral into two parts, the next inequality holds because (by construction) both integrals are positive from the previous line, the next inequality holds because the integration region is from a to ∞ and $x \geq a$ in that region, and the last equality holds by the previously discussed connection between pdfs and cdfs. Recalling that $\mathrm{P}(X > a) = 1 - \mathrm{F}_X(a)$ and rearranging terms completes the proof.

Chebyshev's Inequality. For any random variable X and any $\delta > 0$, $\mathrm{P}(|X - \mathbb{E}[X]| \geq \delta) \leq \frac{\mathrm{var}(Y)}{\delta^2}$.

Proof: Let $Y = (X - \mathbb{E}[X])^2$. Then, Markov's inequality implies that

$$\mathrm{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}$$

Taking $a = \delta^2$, we have that

$$\begin{aligned}
&\mathrm{P}\big((X - \mathbb{E}[X])^2 \geq \delta^2\big) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\delta^2} \\
\iff\ &\mathrm{P}(|X - \mathbb{E}[X]| \geq \delta) \leq \frac{\mathrm{var}(X)}{\delta^2}
\end{aligned}$$

where the last line uses the definition of variance.

Moment generating functions

PSE 2.23, 2.25

To conclude this section, let's mention one more way to fully describe the distribution of a random variable, the **moment generating function** (mgf). The mgf of a random variable X is defined as $M_X(t) = \mathbb{E}[\exp(tX)]$.

Properties of moment generating functions

1. If two random variables have the same mgf for all t in an interval around 0, then they follow the same distribution

2. The k^{th} moment of X is equal to the k^{th} derivative (w.r.t. t) of its mgf evaluated at $t = 0$, that is, $\mathbb{E}[X^k] = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0}$.

In order to see the second property, notice that

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} \mathbb{E}[\exp(tX)] \\
&= \frac{d}{dt} \int \exp(tx) f_X(x)\, dx \\
&= \int \frac{d}{dt} \exp(tx) f_X(x)\, dx \\
&= \int x \exp(tx) f_X(x)\, dx
\end{aligned}$$

If you evaluate, this at $t = 0$, you get $\int x f_X(x)\, dx = \mathbb{E}[X]$. Similarly, notice that

$$\begin{aligned}
\frac{d^2}{dt^2} M_X(t) &= \frac{d}{dt}\Big(\frac{d}{dt} M_X(t)\Big) \\
&= \int x \frac{d}{dt} \exp(tx) f_X(x)\, dx \\
&= \int x^2 \exp(tx) f_X(x)\, dx
\end{aligned}$$

Again, evaluating this at $t = 0$, you get $\int x^2 f_X(x) = \mathbb{E}[X^2]$. You can keep going along these lines for higher order derivatives of the mgf.

mgf's may not exist for for all random variables, so it is common to instead consider the **characteristic function** $\mathbb{E}[\exp(itX)]$ where $i = \sqrt{-1}$. The characteristics function always exists, but for simplicity, (whenever mgfs come up) we'll stick to cases where it exists this semester.

Example: income and education in the U.S.

To conclude this set of notes, let's run through some of the above topics using an actual example where the random variables are income (a continuous random variable) and education (a discrete random variable) of people in the United States. For this section, I'll use data from the U.S. Census Bureau from 2019; in practice, this is actually a sample (though its a fairly large one) rather than the actual population, but we'll pretend that its the full population for now.

```
# load useful packages
library(haven)
library(dplyr)
library(ggplot2)
```

```
# load data
load("us_data.RData")
```

```
# plot pmf of education
ggplot(data=us_data, aes(x=educ, y=..density..)) +
    geom_histogram(binwidth=1) +
    xlab("Years of Education") +
    ylab("pmf") +
    theme_bw()
```

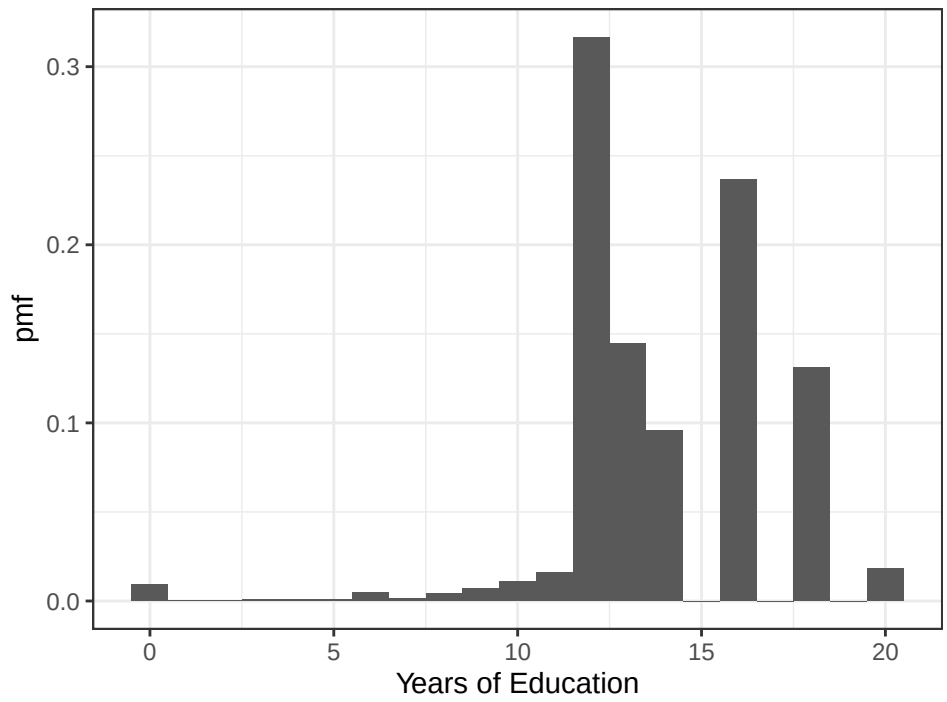


Figure 1: pmf of U.S. education

There are some things that are perhaps worth pointing out here. The most common amount of education in the U.S. appears to be exactly 12 years — corresponding to graduating from high school; about 32% of the population has that level of education. The next most common number of years of education is 16 — corresponding to graduating from college; about 24% of individuals have this level of education. Other relatively common values of education are 13 years (14% of individuals) and 18 (13% of individuals). About 1% of individuals report 0 years of education. It's not clear to me whether or not that is actually true or reflects some individuals mis-reporting their education. Next, we'll make a plot of the cdf of education.

```
ggplot(data=us_data, aes(x=educ)) +
    stat_ecdf() +
    xlab("Years of education") +
    ylab("cdf") +
    theme_bw()
```

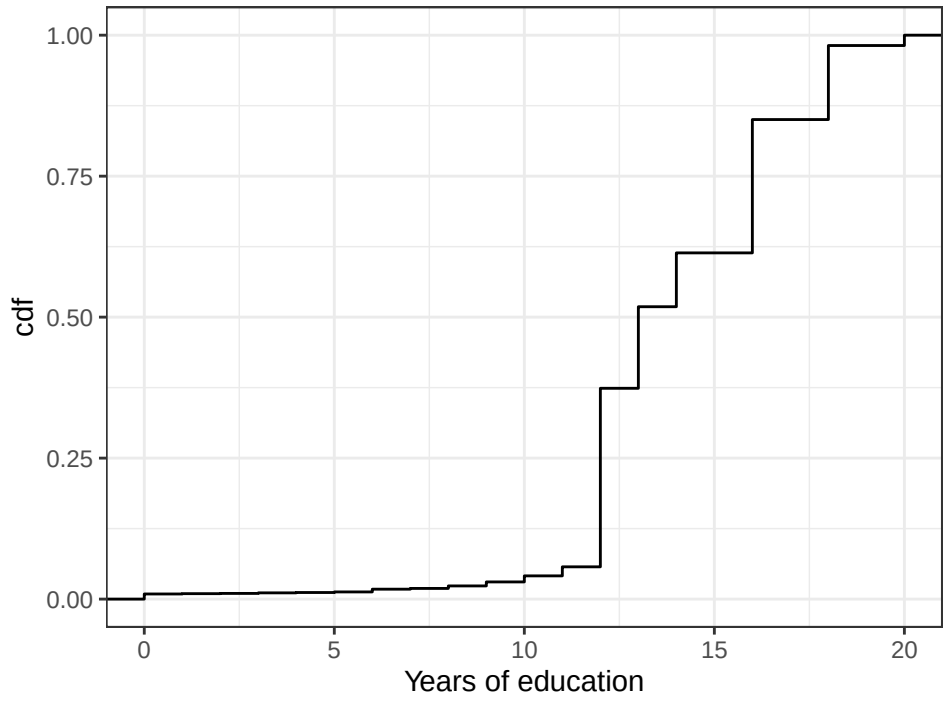


Figure 2: cdf of U.S. educ

You can see that the cdf is increasing in the years of education. And there are big "jumps" in the cdf at values of years of education that are common such as 12 and 16.

Next, let's switch to yearly income. The cdf is plotted in the next figure.

```
ggplot(data=us_data, aes(x=incwage)) +
    stat_ecdf() +
    xlim(c(0,200000)) +
    xlab("Wage Income") +
    ylab("cdf") +
    theme_bw()
```

From the figure, we can see that about 24% of working individuals in the U.S. earn $20,000 or less per year, 61% of working individuals earn $50,000 or less, and 88% earn $100,000 or less. Next, we'll plot the pdf.

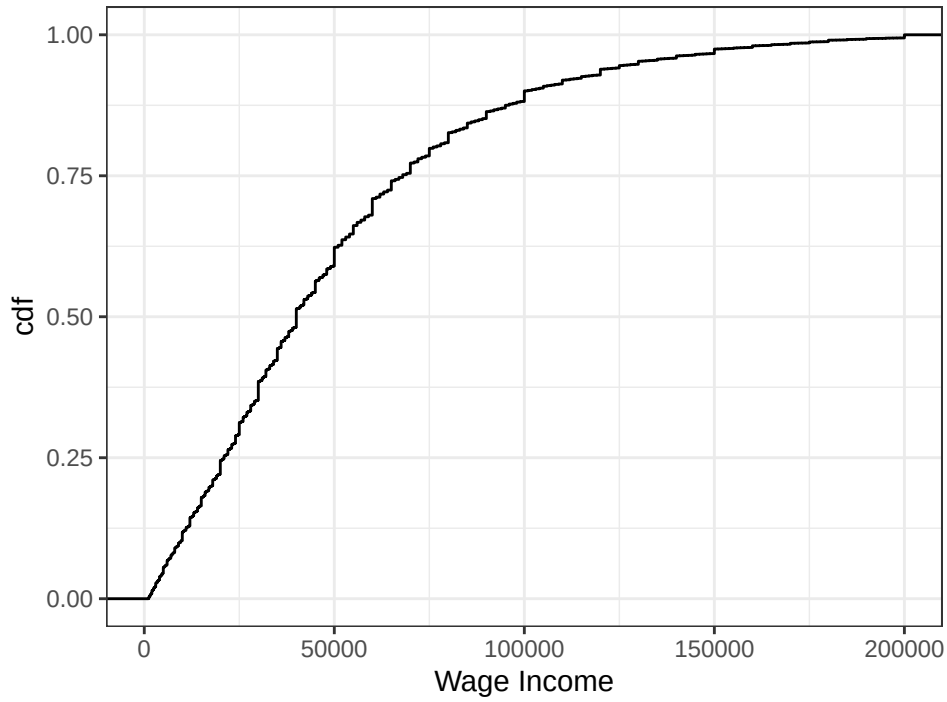


Figure 3: cdf of U.S. wage income

```
dens <- density(us_data$incwage, bw=5000, from=0, to=200000)
plot_df <- data.frame(incwage=dens$x, dens=dens$y)

ggplot(data=plot_df, aes(x=incwage,y=dens)) +
  geom_line() +
  geom_ribbon(data=subset(plot_df, incwage>=50000 & incwage<=100000),
              aes(ymax=dens),
              ymin=0,
              fill="red",
              alpha=.5) +
  geom_ribbon(data=subset(plot_df, incwage>=150000),
              aes(ymax=dens),
              ymin=0,
              fill="green",
              alpha=.5) +
  xlim(c(0,200000)) +
  xlab("Wage Income") +
  ylab("pdf") +
  theme_bw()
```

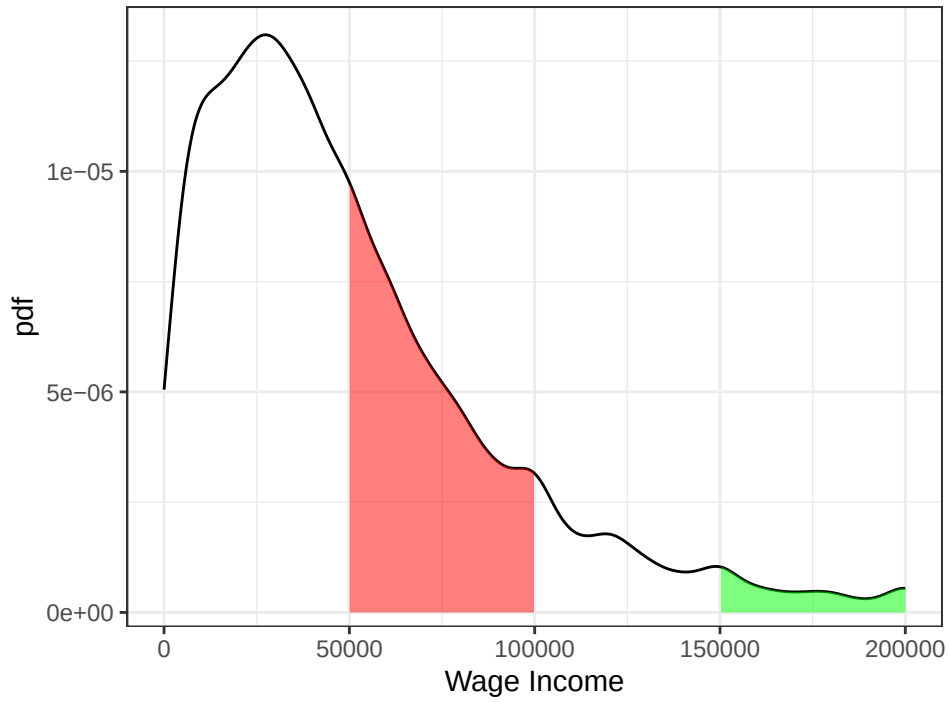


Figure 4: pdf of U.S. wage income

From the figure, we can see that the most common values of yearly income are around $25-30,000 per year. Notice that this corresponds to the steepest part of the cdf from the previous figure. The right tail of the distribution is also long. This means that, while incomes of $150,000+ are not common, there are some individuals who have incomes that high.

Moreover, we can use the properties of pdfs/cdfs above to calculate some specific probabilities. In particular, we can calculating probabilities by calculating integrals (i.e., regions under the curve) / relating the pdf to the cdf. First, the red region above corresponds to the probability of a person's income being between $50,000 and $100,000. This is given by $F(100,000) - F(50000)$. We can compute this in `R` using the `ecdf` function. In particular,

```
incwage_cdf <- ecdf(us_data$incwage)
round(incwage_cdf(100000) - incwage_cdf(50000),3)
```

```
## [1] 0.27
```

The green region in the figure is the probability of a person's income being above $150,000. Using the above properties of cdfs, we can calculate it as $1 - F(150000)$ which is

```
round(1-incwage_cdf(150000), 3)
```

```
## [1] 0.052
```

One might also be interested in calculating the mean, variance, and standard deviation of income in the U.S. We can compute this by

```
round(mean(us_data$incwage),3)
```

```
## [1] 58605.75
```

```
round(var(us_data$incwage),3)
```

```
## [1] 4776264026
```

```
round(sd(us_data$incwage),3)
```

```
## [1] 69110.52
```