POL 345: Notes on Random Variables

Tolya Levshin

December 5, 2015

1 Random Variables

Any random variable consits of two components: a range of possible values over which it is defined (for example, age brackets or gender categories), and a probability structure over that range (how probable is each value in that range?). Suppose we want to model the biological sex of individuals in a certain population. Let S be our random variable for sex, and let S_i denote individual i's sex. Suppose, further, that an individual can be either male or female. Then, we can calculate the probability that individual i is, for example, female by computing $p(S_i = \text{female})$. Notice that $p(S_i = \text{female}) + p(S_i = \text{male}) = 1$, as every individual, at least in our specification, has to belong to one of the two sexes. This reflects a more general rule: the values a random variable can assume form a partition of the sample space, which is to say that they are (1) **disjoint** (i.e., if a random variable should realize itself to one value, this will tell us that it did not realize itself to any of the others) and (2) **mutually exhaustive** of all possible outcomes — as in our example, where every individual belongs to no more than one sex and no individual belongs to neither sex.

Random variables come in two varieties:

1. **Discrete**: the random variable assumes a finite number of distinct values — or, to put the point more accurately, at most countably infinitely many distinct values. Examples include any variables with strictly delineated categories: for instance, age defined in terms of non-overlapping age brackets or education defined in terms of highest completed degree.

2. **Continuous**: the random variable is defined over an interval of the real line, which is to say that it can assume uncountably many values. Here, for any two distinct values, however close they may be, we can always find a third value between them. Examples include length, mass, or time measured on a continuous scale.

2 Distributions

It is common to use statistics to model empirical processes in order to estimate the probability of observing particular outcomes. Every modelling exercise begins with a specification of the relevant random variables and the distributional assumptions governining their probability structure. Suppose, for example, that you are asked to estimate the probability of observing 6 heads after tossing a fair coin 10 times. How would you approach this problem? You should begin by identifying the essential features of the empirical process that you are asked to examine. Reformulate the question in more general terms: 10 trials are observed, each with a binary outcome, and exactly 6 successes and 4 failures occur, where the probabilities of success and failure remain constant throughout. The appropriate distribution for this job is the **binomial distribution**, which is used to model the number of successes observed in a series of (independent) binary trials. (The next section will consider the mathematical aspects of this problem in more depth.) It's important to notice that the same distributional assumption can be used to model the entire gamut of empirical situations in which a series of events are observed, each capable of a binary outcome, and a certain number of outcomes of interest are observed — for example, the number of marriages that end in divorce, the number of students who pass a test, or the number of patients who test positive for a particular desease.

(1): the discrete case

The probability structure of a discrete random variable is given by its **probability mass function** (PMF). The PMF tells us how much mass a variable assigns to each of its possible values — or, in other words, how probable each value is. This function has an intuitive interpretation: for each value of a random variable, it returns the probability of observing exactly that value. Suppose that a random variable X is distributed binomial. Then, its PMF, f_X, is given by the following formula:

$$f_X(k) = p(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \tag{1}$$

where n is the total number of binary trials, k is the number of successes, p is the probability of success in a given trial, and $1-p$ is the probability of failure in a trial. More generally, every random variable can be defined in terms of its distribution (X belongs to the binomial, as opposed to, say, the Gaussian, family) and the parameters needed to specify that distribution (in the case of the binomial, we require the total number of trials, n, and the probability of success in a trial, p).

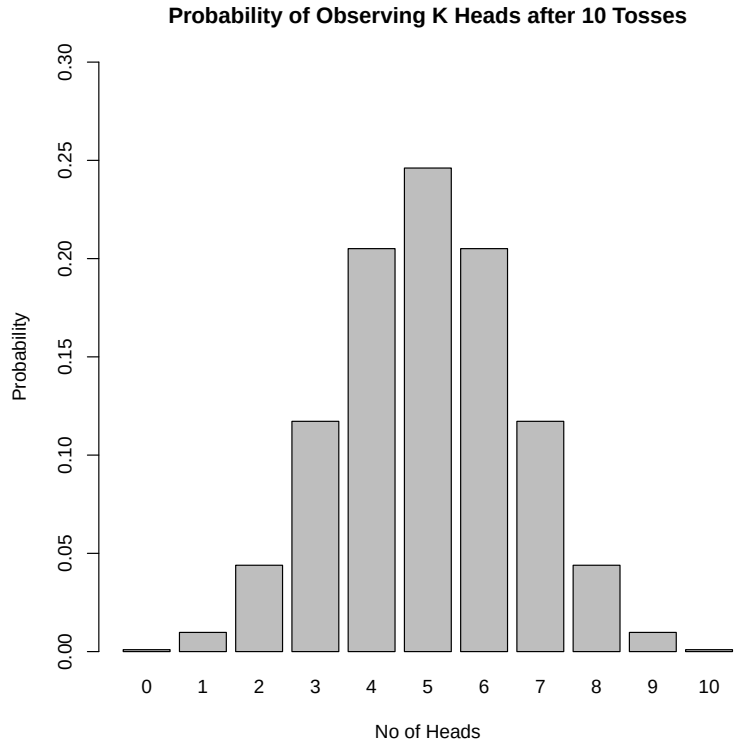
Let's return to the earlier example of observing 6 heads after 10 coin tosses. Let H model the number of heads we observe. Since the coin is fair, we can say that H is distributed binomial with parameters $n = 10$ and $p = 0.5$. The probability of observing exactly 6 heads, $p(H = 6)$, is then given by the PMF of H evaluated at 6:

$$p(H = 6) = f_H(6) = \binom{10}{6} 0.5^6 \times 0.5^{10-6} = 0.21$$

Now, consider a slightly more difficult problem: how many heads are you *most likely* to observe after 10 tosses? To solve this problem, we simply need to compute $p(H = k)$ for every k in the range of possible values over which H is defined. Now, if a coin is tossed 10 times, the number of heads you observe can be anywhere from 0 to 10. (This reflects the more general principle that, for a binomial random variable, $k \in [0, n]$).

It is straightforward to calculate these probabilities in **R** and visualize them on a graph:

```
> ## Define the binomial parameters
> no.trials <- 10
> prob.heads <- 0.5
> possible.no.heads <- 0:10
>
> ## Create a container to record the probability
> ## of observing exactly k heads
> probs <- rep(NA, 11)
>
> for (k in 0:10) {
+   probs[k+1] <- choose(no.trials,k) * (prob.heads^k) *
+     ((1-prob.heads)^(no.trials-k))
+   ## Because R assigns the first element of a vector
+   ## the index of 1, we have to index probs by (k+1)
+ }
>
> names(probs) <- possible.no.heads
>
> ## Plot the probabilities
> barplot(probs, ylim = c(0,0.3),
+         xlab = "No of Heads", ylab = "Probability",
+         main = "Probability of Observing K Heads after 10 Tosses")
```



The above plot visualizes the PMF of H. The height of each bar corresponds to the probability of observing that many heads. We can see that, for a fair coin, the most probable outcome is exactly 5 heads out of a series of 10 tosses:

```
> probs[which.max(probs)]

     5
0.2461
```

But suppose now that, instead of calculating the probability of observing exactly k heads, you are instead asked to calculate the probability of observing *at most k heads*, *more than k heads*, or, for example, *between (k-2) and (k+3) heads*, for some k. How would you approach this problem? Begin by reformulating the condition in mathematical terms. The probability of observing at most k heads, for example, refers to the cumulative probability of observing either 0 heads, or 1 head, or 2 heads, and so forth on to k heads. (Why? Because each of the values of our random variable H are disjoint, as we discussed above; and the probability of a series of disjoint events

is just the sum of the probabilities of the individual events.) So,

$$p(\text{at most } k \text{ heads}) = p(H = 0) + \cdots + p(H = k) = f_H(0) + \cdots + f_H(k)$$

Likewise,

$$p(\text{more than } k \text{ heads}) = p(H = k + 1) + \cdots + p(H = n) = f_H(k + 1) + \cdots + f_H(n)$$

Thus, the probability of a cumulative event can be calculated by simply adding up the probabilities of each of the constituent events, as given by the PMF evaluated at each of those values. It is crucial, in this regard, **to accurately translate the verbal problem you're given into the corresponding mathematical formulation**. For example, the condition *at least k heads* requires us to begin summation from the PMF evaluated at k and onward through to n, but the condition *more than k heads* requires us to begin by calculating the PMF at $k + 1$ and move onward through to n.

Let's return to our example of the fair coin and calculate the probability of observing no more than 6 heads. Tis is straightfoward: we simply need to add up the PMF evaluated at 0, 1, 2, 3, 4, 5, and 6:

```
> sum(probs[1:7])

[1] 0.8281

> ## Why do we index from 1 through 7, rather than 0
> ## through to 6? Remember that R assigns the first
> ## element of any vector the index of 1, and so
> ## the probs vector is indexed from 1 through 11,
> ## even though the binomial variable H is defined
> ## over the range 0 through 10
```

We can visualize this calculation as the process of adding up the bars on the above graph for the values of 0, 1, 2, 3, 4, 5, and 6, and then measuring the total height.

This key intuition — to wit, that, for discrete random variables, the probability of observing a range of values is given by the sum of the PMF evaluated at the corresponding values, one value at a time — is formalized in the concept of the **cumulative distribution function** (CDF). For a discrete random variable X, defined over a range of values from a to b, with a PDF f_X, we formally define its CDF, F_X thusly:

$$F_X(k) = p(X \leq k) = \sum_{i=a}^{k} f_X(i) \tag{2}$$

So, the CDF of a random variable, evaluated at a point, returns the cumulative probability of observing that variable realize a value *at most that great*.

It's good practice to familiarize yourself with how to manipulate a random variable's CDF to calculate a desired cumulative probability. Remember that **the CDF always returns the cumulative probability of observing the random variable realize itself up to a certain value** — that is, the cumulative probability from the lowest possible to the specified value. Therefore, the probability of observing a discrete random variable X yield a value of at most k is simply its CDF evaluated at k: $p(X \leq k) = F_X(k)$. But the probability of observing a value smaller than k is $p(X < k) = F_X(k-1) = F_X(k) - f_X(k)$, as we need to exclude the probability of observing k from the cumulative calculation.

Now, returning to our fair coin example, suppose that we should like to calculate the probability of observing *at least 4 heads*. Note that this is **NOT** given by $F_H(4)$. Why? Because the CDF returns the probability of observing a value no greater than a specified trheshold. So $F_H(4)$ would tell us the probability of observing *at most 4 heads* rather than *at least 4 heads*. A brute-force solution would be to simply add up H's PMF evaluated at each of the relevant points:

$$p(H \geq 4) = \sum_{k=4}^{10} p(H = k) = \sum_{k=4}^{10} f_H(k)$$

We can easily calculate this value in **R**:

```
> sum(probs[5:11])

[1] 0.8281
```

To refomulate our solution in terms of the CDF, we must first realize that the event of observing *at least 4 heads* is the complement (negation) of the event of observing *at most three heads*. The total probability of these two events is exactly 1, and we can therefore formulate the former in terms of the latter:

$$\begin{aligned} p(H \geq 4) + p(H < 4) &= 1 \\ p(H \geq 4) &= 1 - p(H < 4) \end{aligned}$$

Now, remember that, for discrete variables, $p(X < k) \neq p(X \leq k) = F_X(k)$. If our discrete values move in increments of 1, we can simplify thusly:

$$\begin{aligned} p(H \geq 4) &= 1 - \underbrace{p(H < 4)}_{= p(H \leq 3)} = 1 - \underbrace{p(H \leq 3)}_{= F_H(3)} \\ p(H \geq 4) &= 1 - F_H(3) \end{aligned}$$

Calculating this solution in **R**, we obtain the same answer as above:

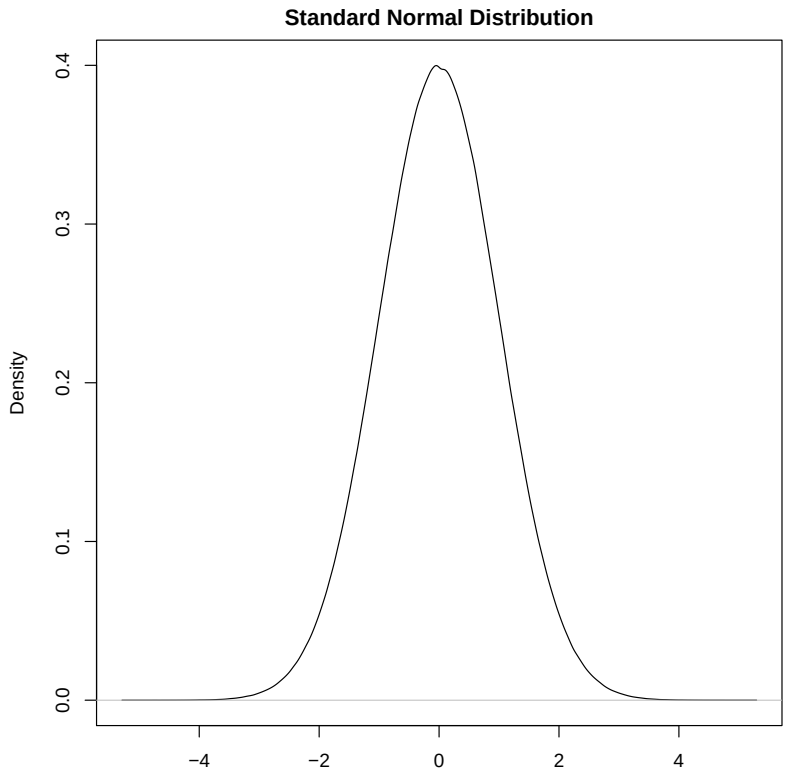
```
> 1-sum(probs[1:4])

[1] 0.8281
```

(2): the continuous case

Continuous variables are a bit trickier to work with, and this note will only summarize those aspects of their mathematics which you require for the quiz. (Strictly speaking, only the application at the end of this section will be necessary for the quiz, but, to make sense of that application, it is necessary to review the conceptual foundations of working with continuous random variables.) The probability structure of a continuous variable Y is given by its **probability density function** (PDF), f_Y. This change in language has an important implication for how you should (and should not) interpret a PDF. $f_Y(k)$ returns the density that Y places on point k. (Remember the formal definition of density: it is, at a given value, **the proportion of observations exhbiting that value per unit of measure**.) Why did we move away from probability to density in the case of continuous variables? Putting the point somewhat crudely, this is because continuous variables have so many possible values — in fact, they have uncountably many values — that the probability of observing any one value, in the continuous case, is exactly 0. It is, therefore, more informative to consider the density of a continuous variables, which, evaluated at a point, can be positive. Although you should not conflate density and probability, a continuous variable's density at a point can still be intepreted as a measure of how probable the neighbourhood immediately surrounding that point is.

One of the most popular continuous distributions is the Gaussian (Normal) distribution, parametrized by its mean, commonly denoted μ, and standard deviation, commonly denoted σ. The range of values that a Normal random variable can assume is given by the whole of the real line, $\mathbb{R}$. (What does this mean? The Normal is defined over the entire family of numbers with the exception of complex numbers — that is, numbers that contain an imaginary component i such that $i^2 = -1$.) The praticular case of a Normal distribution with $\mu = 0$ and $\sigma = 1$ is known as the **standard Normal** distribution, and a graph of its PDF is given below:



The case of the standard Normal is worth examining a bit more closely. Any Normal distribution is symmetrical around its mean and concentrates most of its density within one standard deviation of the mean. Normal distributions can differ in terms of their scale, however, and the scale of the standard Normal should be familiar to you at this point: it is the *z-score* scale. We will return to this point in a moment.

Although continuous random variables assign null probability to individual values in their range, we can still calculate the cumulative probability of observing the variable realize itself to any of a possible interval of values. The only difference from the discrete case is that, instead of summing over the variable's PMF evaluated at each of the points in the interval, we will now integrate the variable's PDF over that interval. This points to the more general relationship between density and probability in the continuous case: the cumulative probability of a certain interval of values is given by the corressponding area under the curve of that variable's PDF.

To illustrate, suppose that X is a continuous variable defined over $[a, b]$. Suppose that we wish to calculate $p(X < c)$, for some $c \in [a, b]$. **Notice that, because continuous random variables assign null probability to any single value in their range,** $p(X < c) = p(X \leq c)$**, which, as you recall, is not true for**

discrete random variables. So,

$$p(X < c) = p(X \leq c) = \int_a^c f_X(x)dx$$

Likewise, the probability of observing X realize a value of at least c is given by:

$$p(X > c) = p(X \geq c) = \int_c^b f_X(x)dx,$$

where, in the case of a continuous random variable defined over the whole of the real line, such as a Normal variable, you should substitute $-\infty$ for a and ∞ for b.

We define the CDF of a continuous random variable X, F_X, which returns the cumulative probability of observing X realize itself to a most a certain value, similarly to the discrete case, substituting integration for summation:

$$F_X(c) = p(X < c) = p(X \leq c) = \int_a^c f_X(x)dx$$

To calculate $p(X > c)$ in terms of the CDF of X, it is easy to proceed by analogy to the discrete case. Because $p(X > c) + p(X < c) = 1$, we have:

$$p(X > c) = 1 - p(X < c) = 1 - F_X(c)$$

Application: Comparing Relative Probabilities

As you can see, calculating the CDFs of continuous random variables can be tricky, and we will not require you to conduct any such calculations on the quiz. However, understanding how CDFs work can prove useful for more general problems. Suppose, for example, that you are given two Normal random variables:

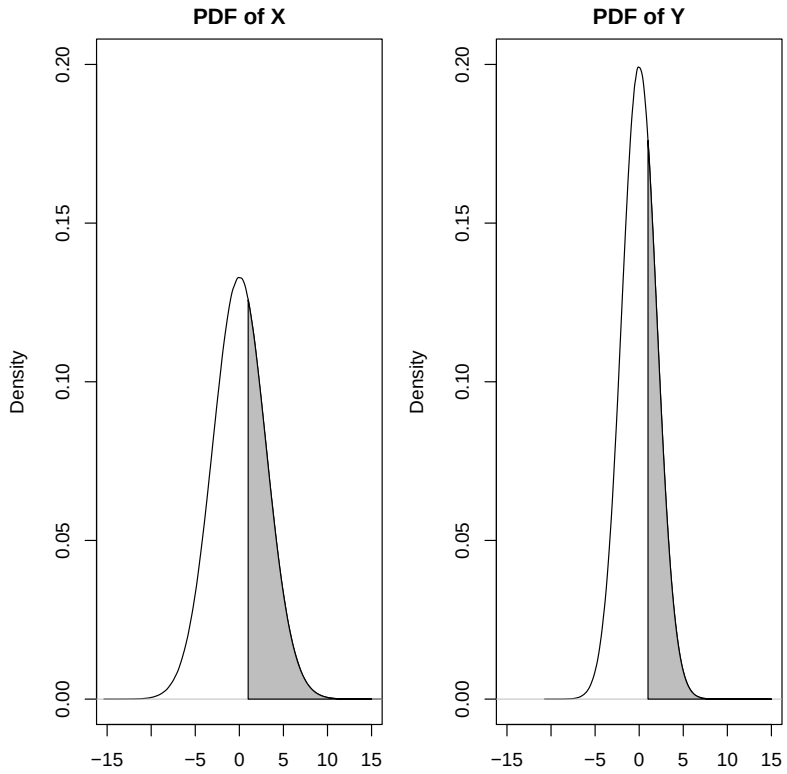
$$X \sim \mathcal{N}(0, 9)$$
$$Y \sim \mathcal{N}(0, 4),$$

where the sign $\sim$ means "distributed", and the notation $Z \sim \mathcal{N}(\mu, \sigma^2)$ means that Z has a Normal distribution with mean μ and standard deviation σ (variance σ^2). Suppose you are asked: which of the following two events is more probable, $X > 1$ or $Y > 1$? One reasonable place to begin is by writing out these probabilities in terms of the corresponding CDFs:

$$p(X > 1) = 1 - p(X < 1) = 1 - F_X(1)$$
$$p(Y > 1) = 1 - p(Y < 1) = 1 - F_Y(1)$$

Thus, to figure out whether $p(X > 1) > p(Y > 1)$, we need to know whether $1 - F_X(1) > 1 - F_Y(1)$ or, equivalently, whether $F_X(1) < F_Y(1)$. Notice that, to solve this problem, you do not need to calculate the CDFs! This is because X and Y belong come from the same family of distributions — the Normal family — and differ only in the parameters of their distributions. If we could put both X and Y on the same scale, it would then be possible to say which CDF generates a higher cumulative probability simply by comparing the relative values of $X = 1$ and $Y = 1$ on that baseline scale.

To contextualize this problem — although you will not be able to do this on the quiz — it's useful to begin by visualizing the two probabilities. We can, first, graph the two densities. Next, we can visualize $p(X > 1)$ and $p(Y > 1)$ as the shaded areas under the two curves, from $X = 1$ to ∞ and from $Y = 1$ to ∞, respectively:



It is difficult to compare the two areas visually because the standard deviations of X and Y are different. But we can put both variables on the same scale to simplify the problem. It is useful to consider the z-scale for the task. Since X and Y are both Normal random variables, putting them on the z-scale will transform them into standard Normal random variables. To perform the transformation, we simply need

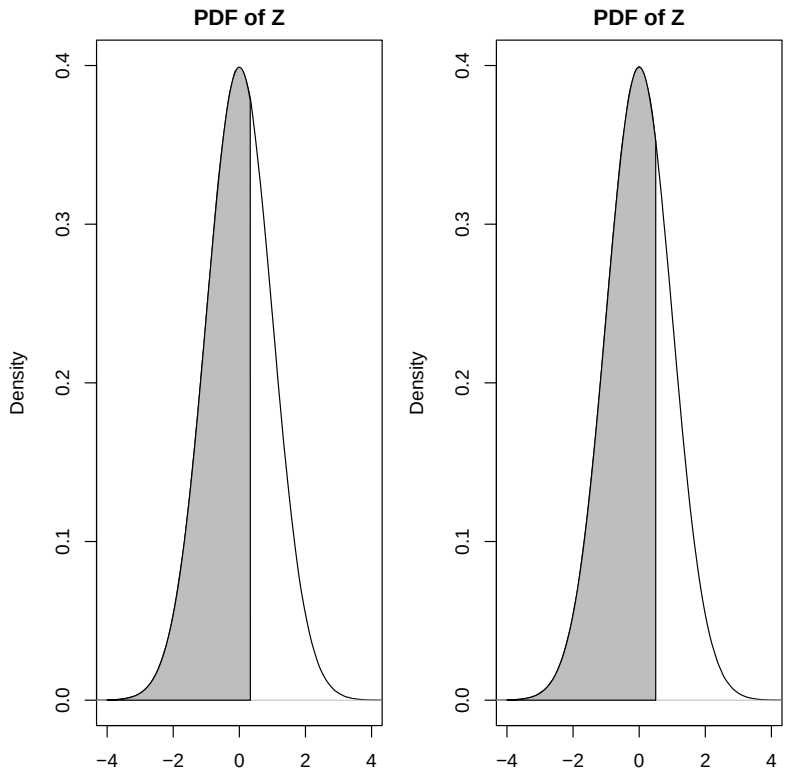
to calculate, separately for X and Y, the z-scores corresponding to each value of X and Y. Specifically, in order to compare $p(X > 1)$ and $p(Y > 1)$, we need to calculate the z-scores corresponding to $X = 1$ and $Y = 1$:

$$z_{x=1} = \frac{1 - \mu_X}{\sigma_X} = \frac{1 - 0}{3} = \frac{1}{3}$$
$$z_{Y=1} = \frac{1 - \mu_Y}{\sigma_Y} = \frac{1 - 0}{2} = \frac{1}{2}$$

Thus, calculating whether $p(X > 1) > p(Y > 1)$ is equivalent to calculating the following expressions:

$$\begin{aligned} p(X > 1) > p(Y > 1) &\Leftrightarrow 1 - F_X(1) > 1 - F_Y(1) \\ &\Leftrightarrow F_X(1) < F_Y(1) \\ &\Leftrightarrow F_Z(Z_{X=1}) < F_Z(Z_{Y=1}) \\ &\Leftrightarrow F_Z(\frac{1}{3}) < F_Z(\frac{1}{2}) \end{aligned}$$

Recall that the CDF calculates the cumulative probability of a random variable up to a specified value. If we compare the values of a CDF for a Normal random variable at two distinct points, we know that the greater of the two points — that is, the one father to the right on the x-axis — will also correspond to the higher value on the CDF, because it allows the CDF to accumulate more probability than a lower value. This general relationship is illustrated by the following two graphs. Both plot the standard Normal density. The plot on the left visualizes $F_Z(\frac{1}{3}) = p(Z < \frac{1}{3})$, and the plot on the right visualizes $F_Z(\frac{1}{2}) = p(Z < \frac{1}{2})$. Observe how the mere fact that $\frac{1}{2} > \frac{1}{3}$, and, by implication, that the CDF of Z evaluated at $\frac{1}{2}$ can accumulate more probability than the same CDF evaluated at $\frac{1}{3}$, is enough to deduce that $p(Z < \frac{1}{2}) > p(Z < \frac{1}{3})$, quite without having to perform any calculations:



Thus, since $\frac{1}{3} < \frac{1}{2}$, we know that $F_Z(\frac{1}{3}) < F_Y(\frac{1}{2})$ and that, from what we demonstrated above, $p(X > 1) > p(Y > 1)$.

3 Expectation and Variance

We define the **expectation** of a random variable as the weighted mean of all of its possible values, each weighted either by its probability, in the discrete case, or by its density, in the continuous case:

$$\mathbb{E}(X) \quad = \quad \begin{cases} \sum_x xf(x), & \text{if the variable is discrete} \\ \int_x xf(x)dx, & \text{if the variable is continuous} \end{cases} \tag{3}$$

We define the variance of a random variable, X, as the average squared deviation from its expectation:

$$\mathbb{V}(X) \quad = \quad \mathbb{E}\big[(X - \mathbb{E}(X))^2\big]$$

For example, suppose that X is distributed binomial, with the total number of trials given by n and the probability of success in a trial by p. Then, $\mathbb{E}(X) = np$ and

$\mathbb{V}(X) = np(1-p)$. To illustrate, in our example of the fair coin tossed 10 times, the expected (i.e., average) number of observing heads over 10 tosses is given by $np = 10 \times 0.5 = 5$, with variance of $np(1-p) = 10 \times 0.5 \times 0.5 = 2.5$. This means that, if we repeatedly tossed a fair coin 10 times, then, on average across these repeated sequences of 10 tosses, the average number of times that heads would come up is 5, and the variance from one sequence to another is 2.5. For a Normal random variable Y, distributed $Y \sim \mathcal{N}(\mu, \sigma^2)$, we have $\mathbb{E}(Y) = \mu$ and $\mathbb{V}(Y) = \sigma^2$.

Although you're not expected to remember the above formulae, it is important that you memorize the following rules for working with the expectation and variance functions. Suppose that X and Y are two random variables, which may or may not be independent, and a and b are some constants. Then, the following results hold:

1. $\mathbb{E}(a) = a$ and $\mathbb{V}(a) = 0$
2. $\mathbb{E}(aX) = a\mathbb{E}(X)$ and $\mathbb{V}(aX) = a^2\mathbb{V}(X)$
3. $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$ and $\mathbb{V}(aX + b) = a^2\mathbb{V}(X)$
4. $\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$

Suppose that you are further given that X and Y are independent. Then, the following additional result holds:

1. $\mathbb{V}(aX + bY) = \mathbb{V}(aX) + \mathbb{V}(bY) = a^2\mathbb{V}(X) + b^2\mathbb{V}(Y)$

Application: Linear Combinations of Independent Normal Variables

One important application of the aforementioned rules is to the class of problems in which you need to calculate the distribution of a random variable which is defined as a combination of other random variables. We only expect you to know one straightfoward case of this general problem: linear combinations of independent normal variables. Suppose, for example, that you're given two independent random variables, X and Y, with the following distributional assumptions:

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$
$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Suppose, next, that you're given some $Z = aX + bY + c$, for some constants $a, b,$ and c. How is Z distributed? There are two steps to solving this problem. The **first step** is

to determine the distributional family to which Z belongs. There is a result, which we state here for you without proof, that **a linear combination of independent normally-dsitributed random variables is itself normal**. Since Z is a linear combination of X and Y, two indepndent and normally-distributed random variables, you can therefore infer that Z is itself normal: $Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$, for some unknown μ_Z and σ_Z^2.

In the **second step**, we calculate these two quantities. Remember that $\mathbb{E}(Z) = \mu_Z$ and $\mathbb{V}(Z) = \sigma_Z^2$. Simply plug in the formula for Z that you're given, and apply the above rules:

$$\begin{aligned}
\mu_Z &= \mathbb{E}(Z) \\
&= \mathbb{E}(aX + bY + c) \\
&= \mathbb{E}(aX) + \mathbb{E}(bY) + \mathbb{E}(c) \\
&= a\mathbb{E}(X) + b\mathbb{E}(Y) + c \\
&= a\mu_X + b\mu_Y + c
\end{aligned}$$

Further taking advantage of the fact that X and Y are independent, you can calculate the variance of Z:

$$\begin{aligned}
\sigma_Z^2 &= \mathbb{V}(Z) \\
&= \mathbb{V}(aX + bY + c) \\
&= \mathbb{V}(aX) + \mathbb{V}(bY) + \mathbb{V}(c) \\
&= a^2\mathbb{V}(X) + b^2\mathbb{V}(Y) \\
&= a^2\sigma_X^2 + b^2\sigma_Y^2
\end{aligned}$$

Thus, the solution is that $Z \sim \mathcal{N}(a\mu_X + b\mu_Y + c, a^2\sigma_X^2 + b^2\sigma_Y^2)$.