Random Variables Pt. 1

A Few Basics

Elliot Pickens

May 25, 2021

1 Intro

Random variables are essential to any study of probability and statistics, but they can often be a tricky subject to handle. They are not exactly the most intuitive concept (due in part to their notoriously vague name), but that confusion can be easily explained away if we start from the basics. In this post (and the next few that'll come after it) I'm going to spend some time laying out a few key concepts related to random variables that'll hopefully help anyone who's reading this start off their journey into statistics strong.

2 Random Variables (The Real Basics)

2.1 What is a Random Variable?

Let's begin with the definition of a random variable. A random variable is a function that outputs real values and is defined on a sample space. The key to this definition is the inclusion of the sample space. Being a real valued function alone would be nothing interesting, but with random variables we are trying to quantify "random" events occurring within the confines of our sample space S. As an example let's consider a coin flip. X will represent the number of heads we get in a given sequence of flips. For a single flip X can equal 1 or 0, as our sample space $S = \{H, T\}$. As the number of flips increases the size of our sample space increase as 2^n (since each flip be either H or T), while X can take on the value of any integer between 0 and n (reflecting the more limited number of heads we can get in any sequence). As we will see moving forward, this basic definition can lead to a lot of *fun*.

2.2 Distributions of Random Variables

Using our newly defined random variables we can create distributions. For a random variable X, its distribution is set of probabilities $P(X \in C)$ where C is the composition of all sets of real numbers for which $\{X \in C\}$ is an event. That is to say the distribution of a random variable X is the set of probabilities associated with the values that X can take on. These distributions can either be continuous or discrete. If X can only take on a finite number of values (or an infinite sequence of values) then it is called a finite distribution and will have a distribution that is similarly discrete. Alternatively, if X can take on more than a finite number of values (i.e. every value on some interval) then it is defined as a continuous random variable.

2.2.1 Probability Functions

For a discrete random variable X the probability function (or probability mass function) is defined for all $x \in \mathbb{R}$ as

$$f(x) = P(X = x) \tag{1}$$

We should also note the *support* of X is the closure of the set $\{x | f(x) > 0\}$.

Then since $f(x) = 0$ if an event x is not a possible value of X we can get the result that $\sum_{i=1}^{\infty} f(x_i) = 1$ (assuming the sequence $x_1, x_2, ...$ contains all possible values of X). Furthermore, we can get the probability of any subset of the real line C by calculating $P(x \in C) = \sum_{x_i \in C} f(x_i)$.

2.2.2 A Few Important (Discrete) Distributions

No need to spend much time on these, but given how important these distributions are they deserve a small spot in this write-up.

Definition 2.1. A **Bernoulli** Distribution/Random Variable is a random variable that can only take on the values 0 and 1, and will take on the value of 1 with probability p. For such a random variable Z we can say that $P(Z = 1) = p$, and that Z is a Bernoulli random variable with parameter p

Definition 2.2. A discrete **Uniform** Distribution/Random Variable is a random variable that takes on the values of all integers $a, ..., b$ (where $a \leq b$) with equal probability. Therefore if we have a uniform distribution on a set of n the random variable will take on the value of any integer in the set with probability $1/n$.

Definition 2.3. A **Binomial** Distribution/Random Variable is a distribution defined by the probability function $f(x) = \binom{n}{k} p^x (1-p)^{n-x}$ where $\{x \in \mathbb{Z} | 0 \leq x \leq n\}$. n is the parameter for the number of trials, and p is the parameter that defines the probability of a "positive" outcome for a given trial.

2.3 Continuous Random Variables and Distributions

2.3.1 Probability Density Functions

We can define continuous random variables as random variables with a valid probability density function. Unlike the probability functions we saw in section 2.2.1, PDFs are over the entire real line and their integral can be evaluated on any interval (open or closed). By computing such an integral we can find the probability that our *continuous* random variable will take on a value on the interval in question.

More formally we can say that a random variable X is continuous if there exists a function f that is defined on the entire real line, is always non-negative, and whose integral over any interval evaluates to the probability of X embodying a value on that interval. This means that for any interval (a, b) (or $[a, b]$)

$$P(a \leq X \leq b) = \int_a^b f(x)dx \tag{2}$$

f in this case is the probability density function (or PDF). Just like probability functions we define the *support* of a PDF as the closure of the set $\{x | f(x) > 0\}$. And just as with probability functions the total probability must always evaluate to 1 as

$$\int_{-\infty}^{\infty} f(x)dx = 1 \tag{3}$$

One important thing to remember about PDFs is that although $f(x) \geq 0 \,\forall\, x$, $P(X = a) = 0 \,\forall\, a$. This follows from the fact that integrals evaluate to 0 for single points, which can be proven using a straightforward epsilon squeezing proof approach.

An indirect consequence of this is that PDFs are not unique. Since individual points hold density and not probability, we can change the value of a finite or even countable number of points without changing the outcome of the integral. Generally speaking this is not particularly important, but occasionally it becomes useful and should be noted for times when discontinuities are present in PDFs.

2.4 Cumulative Distribution Functions

In the past few sections we have seen how PFs and PDFs can be used to find the probability of a random variable taking on a value. It's an essential task when working with random variables, but we don't have to think about it purely in terms of PF/PDFs. Instead we can use a random variable's cumulative distribution function.

The cumulative distribution function or CDF is a function that exists for all random variables (discrete, continuous, and anything in between), and is used to find the probability of a random variable taking on a probability less than some value. We define this function F as

$$F(x) = P(X \leq x) \; where \; -\infty < x < \infty \tag{4}$$

This definition of F is strikingly close to what we wrote out in 1 and 2, but none of these three are the same. F returns the the probability that X will be *less* than or equal to some value x. It may seem like a small detail, but it nets us some very nice results.

These include the property that $F(x)$ cannot decrease as x increases, so $F(a) \leq F(b)$ for all $a \leq b$. This allows us to find the probability of X landing within an interval by calculating $P(a < x \leq b) = F(b) - F(a)$. And when paired with the fact that the limits of F $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$ we can get $P(X > x) = 1 - F(x)$.

The results in the previous paragraph are broadly speaking the most important for general calculations involving CDFs, but there are a few other things I'll briefly mention. Given that CDFs are directly related to their underlying PF/PDFs (which can have discontinuities) not everything that is true of CDFs when approaching from the right is true from the left. To talk about these discrepancies we use the notation $F(x^+) = \lim_{y \to x,\, y > x} y$ to represent a rightward approach, and $F(x^-) = \lim_{y \to x,\, y < x} y$ for leftward approaches.

An example of this is that CDFs are always continuous from the right, but are not necessarily continuous from the left. Thus, $F(x) = F(x^+)$ for all x, but the same cannot be said of $F(x^-)$. This asymmetry can also be used to show that $P(X < x) = F(x^-)$ and $P(X = x) = F(x) - F(x^-)$. The sorts of jumpy behavior that can be seen here is particularly true of discrete CDFs, which tend to have very noticeable discontinuities at each point of non-zero probability.

2.4.1 CDFs in the Continuous Context

The CDF of a continuous random variable F is

$$F(x) = \int_{-\infty}^{x} f(t)dt \tag{5}$$

which leads us to the following identity (so long as f is continuous at x)

$$f(x) = \frac{dF(x)}{d(x)} \tag{6}$$

This is probably not a very surprising result, but it does reinforce some of what we already know about the continuity of CDFs. Since CDFs can only increase or stay the same as $x \to \infty$ and no single point holds any actual probability we get that F is continuous and we can use just a little calculus to pin down its relation to f.

2.5 Quantiles

Let's say we have a random variable that represents the amount an investor could lose in a single day. They would like to know the risk they're holding. How might they do this? One way is to find the level of loss they're [insert tolerable risk]% sure the day's losses will be less than. They could do this is by just trying different values of x until they find the x that satisfies $F(x) = 0.99$. This would work, but it comes with the serious drawback that even with an optimized algorithm it could still be a very slow process of trial and error.

Instead we can use the quantile function $F^{-1}(p)$. The quantile function acts as an inverse of $F(x)$ by returning the smallest x that gives us $F(x) \geq p$. $F^{-1}(p)$ is not, however, always a true inverse of $F(x)$. If F is based on a random variable that is continuous and one-to-one then yes, F^{-1} is a proper inverse, but otherwise it cannot be taken as such. An example of where this may come into play is when two random variables have the same distribution. Both will have the same quantile function if even they are not the same, so long as they follow the same distribution. If X_1 has a continuous, one-to-one distribution and X_2 is the same as X_1 aside from random changes at a finite number of points then both will have the same quantile function F^{-1}, but it will only be a true inverse for X_1.

2.5.1 Medians and Quartiles

The $50th$ percentile is called the median. The $25th$ percentile is the lower quartlie, and the $75th$ percentile is the upper quartile. Each of these can be quite useful when summarizing distributions, but it is important to remember that medians and quartiles normally refer to the smallest value of x that produces the desired p value from F. They can, however, be used as terms for other things like the set of values associated with say $p = 0.5$, so you should be aware that these terms may not be communicating exactly what you expect.

3 Conclusion

This is the first installment in what is to be a several part series on basics of random variables. In this first part we touched on univariate random variables, their distributions, and a few ways we can work with those distributions. Moving forward we'll introduce a little more complexity as we deal with multivariate random variables and functions of random variables.

4 Acknowledgments

These notes were based on *Probability and Statistics (Fourth Edition)* by DeGroot & Schervish.