CLARK UNIVERSITY 1887

Introduction to Random Variables
Math 217 Probability and Statistics

Prof. D. Joyce, Fall 2014

A *random variable* is nothing more than a variable defined on a sample space Ω as either an element of Ω or a function on Ω. We usually denote random variables with letters from then end of the alphabet like X, Y, and Z.

It might refer to an element in Ω. For example, if you flip a coin, the sample space is $\Omega = \{H, T\}$. A random variable X would have one of the two values H or T.

It might refer to a function on Ω. If you toss two dice, and the outcome of the first die is indicated by the random variable X on $\Omega = \{1, 2, 3, 4, 5, 6\}$, and the outcome of the second die is indicated by the random variable $Y \in \Omega$, then the sum of the two dice is a random variable $Z = X + Y$. This random variable Z is actually a function $\Omega^2 \to \mathbf{R}$ since for each ordered pair $(X, Y) \in \Omega \times \Omega$ it gives a number from 2 through 12.

Since random variables on Ω can be considered as functions, by declaring a random variable to be a function on a sample space we've covered both elements and functions.

Real-valued random variables. The random variables we'll consider are all real-valued random variables, so they're functions $\Omega \to \mathbf{R}$. So when we say "let X be a random variable" that means formally a function $X : \Omega \to \mathbf{R}$.

We've looked at lots of random variables. For example, when you toss two dice, their sum, which is an integer in the range from 2 to 12, is a random variable. In a Bernoulli process, there are several interesting random variables including X_n, the number of successes in n trials, and T, the number of trials until the first success.

We can use the notation of random variables to describe events in the original sample space. Let X_n and T be the random variables just mentioned in a Bernoulli process. Then $X_n > 3$ is the event that we get more than 3 successes among n trials, and $P(X_n > 3)$ is the probability of that event. Also, $5 \le T \le 8$ is the event that the first success occurs no sooner than the 5^{th} toss and lo later than the 8^{th} toss. We can even use algebra on random variables like we do ordinary variables: the expression $|T - 30| < 5$ says the first success occurs within 5 trials of the 30^{th} trial.

Real random variables induce probability measures on R. When we have a real random variable $X : \Omega \to \mathbf{R}$ on a sample space Ω, we can use it to define a probability measure on $\mathbf{R}$. For a subset $E \subseteq \mathbf{R}$, we can define its probability as $P(X \in E)$. When $\mathbf{R}$ has a probability measure on it like that, we can do things that we can't do for abstract sample spaces like Ω. We can do arithmetic and calculus on $\mathbf{R}$. We'll do that. First we'll will look at discrete random variables where we can do arithmetic and even take infinite sums. Later on, we'll look at continuous random variables, and for those we'll need differential and integral calculus.

Probability mass functions. A *discrete random variable* is one that takes on at most a countable number of values. Every random variable on a discrete sample space is a discrete random variable.

The *probability mass function* $f_X(x)$, also denoted $p_X(x)$, for a discrete random variable X is defined by

$$f_X(x) = P(X{=}x)$$

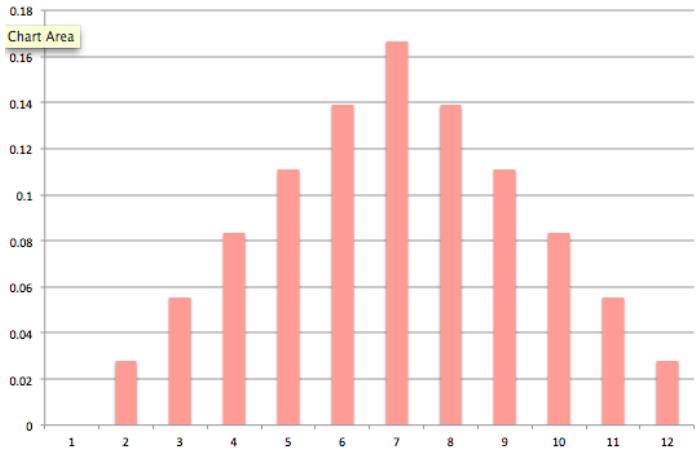
When there's only one random variable under consideration, the subscript X is left off and the probability mass function is denoted simply $f(X)$.

For example, let X be the random variable which gives the sum of two tossed fair dice. Its probability mass function has these values

x	2	3	4	5	6	6	8	9	10	11	12
$p(x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

Probability mass functions are usually graphed

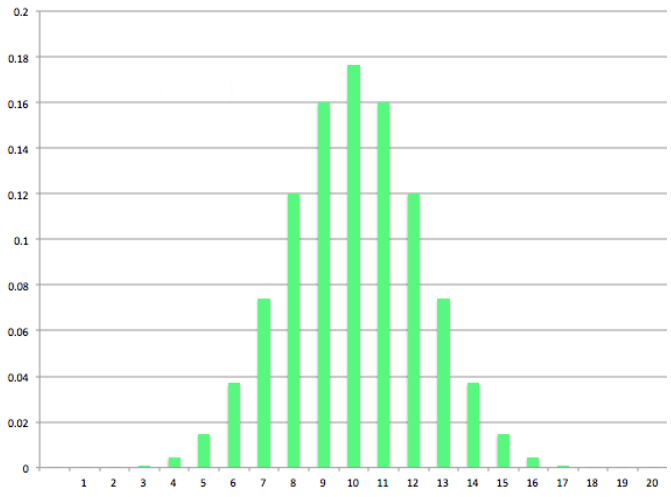
as histograms, that is, as bar charts. Here's the one for the dice.



For another example, consider the random variable X_n, the number of successes in n trials of a Bernoulli process with probability p of success. It has a binomial distribution. We've already computed the probability mass function, although we didn't call it that, and we found that

$$f(x) = \binom{n}{x} p^x q^{n-x}.$$

Here's a histogram for it in the case that $n = 20$ and $p = \frac{1}{2}$.

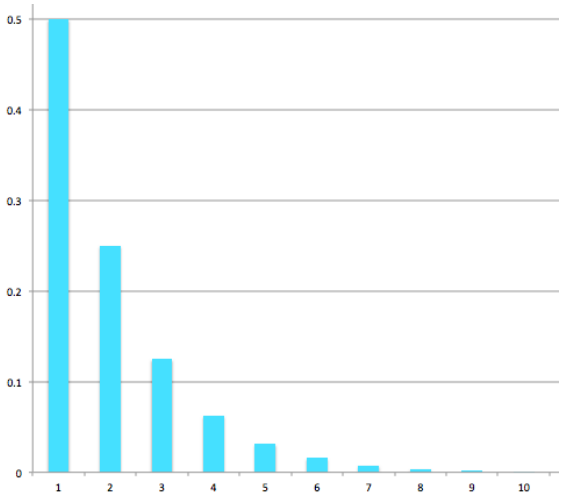


This shape that we're getting is approximating what we'll call a *normal distribution*. Jacob Bernoulli studied it and showed the first version of the Central Limit Theorem, that as $n \to \infty$, a Bernoulli distribution actually does approach a normal distribution.

One last example of a probability density function. Let T be the number of trials until the first success in a Bernoulli process with $p = \frac{1}{2}$. That has a geometric distribution, and we found that

$$f(t) = pq^{t-1} = \frac{1}{2^t}.$$

The first part of its histogram is shown below. It continues off to the right, but the bars are too short to see.



Probability mass functions aren't used for continuous probabilities. Instead, something called a probability density function is used. We'll look at them later.

Cumulative distribution functions. Let X be any real-valued random variable. It determines a function, called the *cumulative distribution function* abbreviated *c.d.f.* or, more simply, the *distribution function* $F_X(x)$ defined by

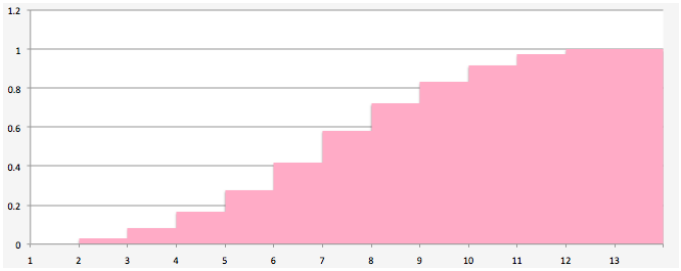
$$F_X(x) = P(X \leq x).$$

When there's only one random variable under consideration, the subscript X is left off and the c.d.f. is denoted $F(x)$. In the expression $X \leq x$, the symbol X denotes the random variable, and the symbol x is a number.

For a discrete random variable, you can find the c.d.f. $F(x)$ by adding, that is, accumulating, the values of probability mass function $p(x)$ for values less than or equal to x. Thus, the c.d.f. for the sum of two tossed dice has these values

x	2	3	4	5	6	7	8	9	10	11	12
$F(x)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{6}{36}$	$\frac{10}{36}$	$\frac{15}{36}$	$\frac{21}{36}$	$\frac{26}{36}$	$\frac{30}{36}$	$\frac{33}{36}$	$\frac{35}{36}$	$\frac{36}{36}$

The cumulative distribution function for a discrete random variable is a step function with values between 0 and 1, starting off at 0 on the left and stepping up to 1 on the right. Cumulative distribution functions usually aren't shown as graphs, but here's the one for the sum of two dice.



Math 217 Home Page at `http://math.clarku.edu/~djoyce/ma217/`