# 14.31/14.310 Lecture 3

A couple of housekeeping notes:

Keep an eye out for the problem set.

Lecture notes for lectures 1 and 2 now posted.

Today: random variables, distributions, joint distributions

Tuesday: examples, histograms, kernel density plots, data sources and techniques for gathering (e.g., webscraping)

Wednesday: independence of random variables, conditional distributions

Recitation: more on R, ??

# Probability---random variables

Often times, there is some numerical characteristic of the sample space that we're interested in (e.g., the sum on the faces of two dice, the number of 3-pt FGs Steph Curry makes in his next six attempts, the number of vegetarian toppings I get on my randomly-selected pizza). There is an important and useful mathematical construct we exploit to analyze that numerical characteristic, the random variable.

A random variable is a real-valued function whose domain is the sample space.

# Probability---random variables

probability

$$2^S \xrightarrow{P} [0,1]$$

set of all subsets
of the sample space

random variable

$$S \xrightarrow{X} \mathbb{R}$$

$$P(X \in A) = P\{s : X(s) \in A\}$$

The probability induces the distribution of X.

# Probability---random variables

All of the probability examples we've seen would give rise to a type of random variable called discrete, one that can take on only a finite or countably infinite number of values.

We can also consider a natural generalization (or different flavor) of this construct, a random variable that can take on any value in some interval, bounded or unbounded, of the real line. This is called a continuous random variable.

We will mostly deal with continuous random variables, partly because many discrete random variables can be adequately approximated with a continuous random variable.

# Probability---random variables

For discrete random variables, you can often start with a verbal description, calculate probabilities for each value of the random variable, and then write down a function or draw a graph describing those probabilities for different values of the random variable. This is called a <u>probability function</u> (PF).

We've already done the first part of this exercise for two types of discrete random variables, hypergeometric and binomial. Let's complete the exercise.

# Probability

**EXTRA TOPPINGS**

Caramelized Onions, Pickled Banana Peppers, Mushrooms, Green Olives: **$1.50 | $3**
Arugula, Sopressata, Sausage, Bacon, Chicken *: **$2.50 | $4**
2 Farm Eggs *: **$3.5**
Marinated White Anchovies *: **$5/8**

Hypergeometric (pizza topping) random variable:

Let X be the number of vegetarian toppings I get on my pizza if I draw the Area Four toppings randomly (without replacement).

We can calculate the probability that X = 0, 1, 2, and so forth, up to the maximum of 6 or n, whichever is smaller, using the formula from last time:

$$\frac{\binom{6}{x}\binom{5}{n-x}}{\binom{11}{n}}$$

**EXTRA TOPPINGS**

Caramelized Onions, Pickled Banana Peppers, Mushrooms, Green Olives: **$1.50 | $3**
Arugula, Sopressata, Sausage, Bacon, Chicken *: **$2.50 | $4**
2 Farm Eggs *: **$3.5**
Marinated White Anchovies *: **$5/8**

# Probability

Hypergeometric (pizza topping) random variable:

Let X be the number of vegetarian toppings I get on my pizza if I draw the Area Four toppings randomly (without replacement).

We can calculate the probability that X = 0, 1, 2, and so forth, up to the maximum of 6 or n, whichever is smaller, using the formula from last time:

$$\frac{\binom{6}{x}\binom{5}{n-x}}{\binom{11}{n}}$$

Note that 0! is defined as 1.

Note that I changed up the notation a bit.

**EXTRA TOPPINGS**

Caramelized Onions, Pickled Banana Peppers, Mushrooms, Green Olives: **$1.50 | $3**
Arugula, Sopressata, Sausage, Bacon, Chicken *: **$2.50 | $4**
2 Farm Eggs *: **$3.5**
Marinated White Anchovies *: **$5/8**

# Probability

Note, though, that the probabilities will be a function of n, the number of toppings.

For concreteness, let's choose $n = 3$.

$P(X=0) = 6/99$

$P(X=1) = 36/99$

$P(X=2) = 45/99$

$P(X=3) = 12/99$

# Probability

**EXTRA TOPPINGS**

Caramelized Onions, Pickled Banana Peppers, Mushrooms, Green Olives: **$1.50 | $3**

Arugula, Sopressata, Sausage, Bacon, Chicken *: **$2.50 | $4**

2 Farm Eggs *: **$3.5**

Marinated White Anchovies *: **$5/8**

Note, though, that the probabilities will be a function of n, the number of toppings.

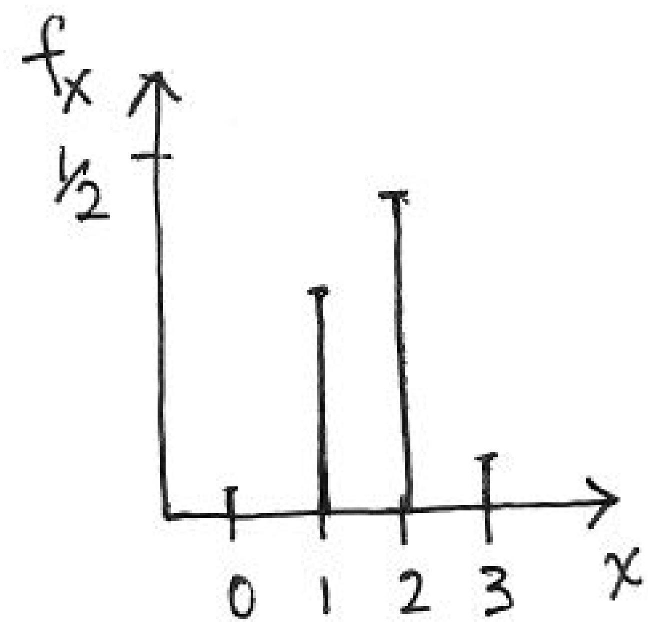
For concreteness, let's choose n = 3.

$P(X=0) = 6/99$

$P(X=1) = 36/99$

$P(X=2) = 45/99$

$P(X=3) = 12/99$



# Probability

EXTRA TOPPINGS

Caramelized Onions, Pickled Banana Peppers, Mushrooms, Green Olives: **$1.50 | $3**

Arugula, Sopressata, Sausage, Bacon, Chicken *: **$2.50 | $4**

2 Farm Eggs *: **$3.5**

Marinated White Anchovies *: **$5/8**

Note, though, that the probabilities will be a function of n, the number of toppings.

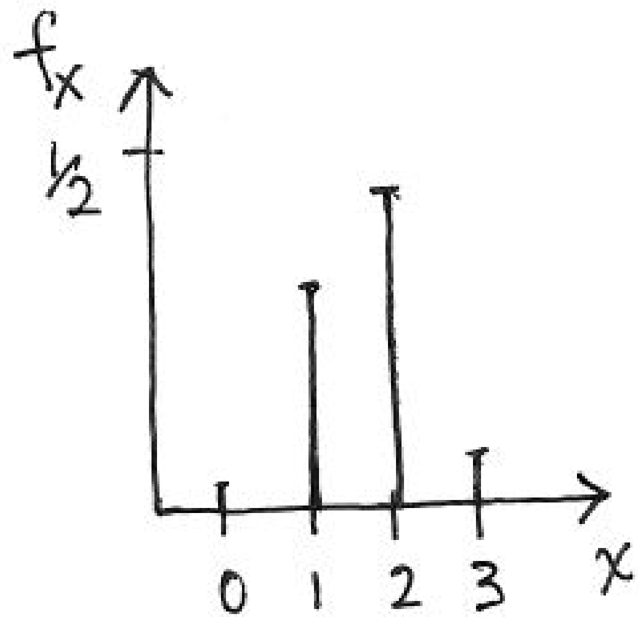
For concreteness, let's choose $n = 3$.

$P(X=0) = 6/99$

$P(X=1) = 36/99$

$P(X=2) = 45/99$

$P(X=3) = 12/99$




By convention, when we graph PFs, we put vertical lines beneath each point on the graph. Also, each point is known as a "point mass."

# Probability


More generally, we say that X has a "hypergeometric distribution with parameters N, K, & n," denoted $X \sim H(N,K,n)$. Its PF is

$$f_X(x) = \frac{\binom{K}{x}\binom{N-k}{n-x}}{\binom{N}{n}} \qquad x = \max(0, n+K-N), \ldots, \min(n, K)$$

The hypergeometric distribution describes the number of "successes" in n trials where you're sampling without replacement from a sample of size N whose initial probability of success was K/N.

# Probability

Binomial (3pt FG) random variable:

Let X be the number of 3pt shots that Steph Curry makes in the next six shots he takes.

We can calculate the probability that X = 0, 1, 2, and so forth, up to the maximum of 6 using the formula described last time:

$$\binom{6}{x} .44^{x} .56^{6-x}$$

# Probability

Plugging in, we get,

$P(X=0) = .03$

$P(X=1) = .15$

$P(X=2) = .29$

$P(X=3) = .30$

$P(X=4) = .18$

$P(X=5) = .06$

$P(X=6) = .01$

# Probability

Plugging in, we get,

$P(X=0) = .03$

$P(X=1) = .15$

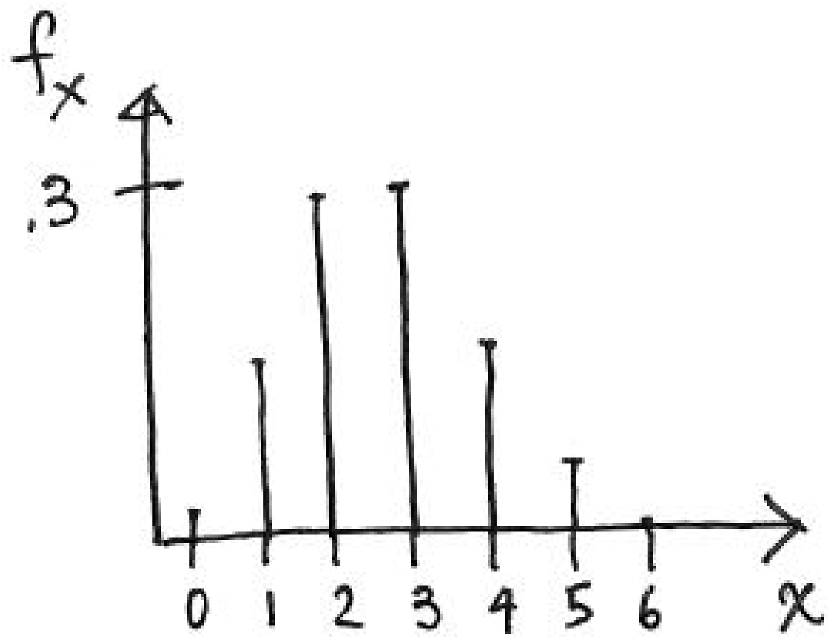
$P(X=2) = .29$

$P(X=3) = .30$

$P(X=4) = .18$

$P(X=5) = .06$

$P(X=6) = .01$



# Probability

More generally, we say that X has a "binomial distribution with parameters n & p," denoted $X \sim B(n,p)$. Its PF is

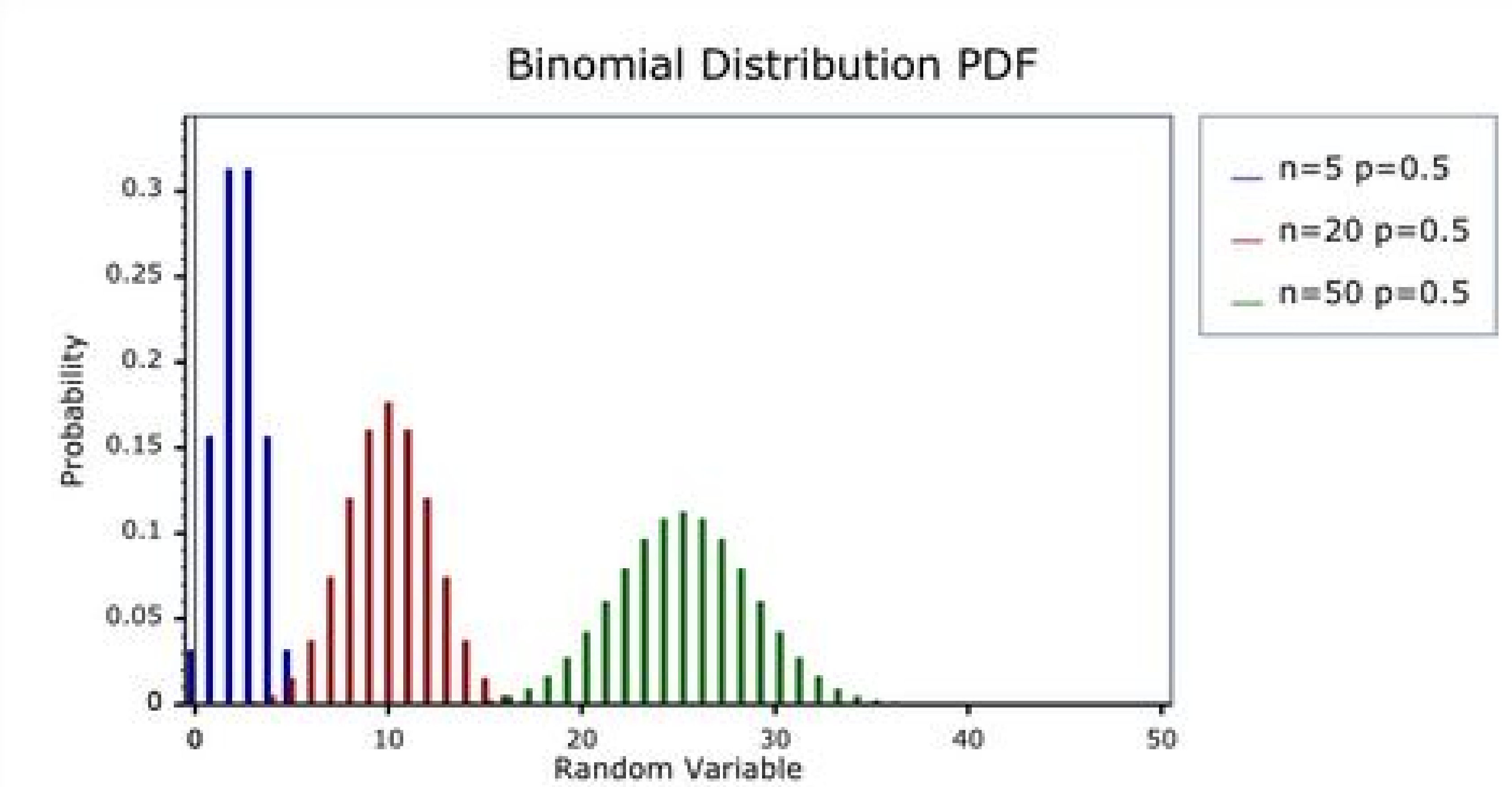
$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad x = 0, 1, \ldots, n$$

The binomial distribution describes the number of "successes" in n trials where the trials are independent and the probability of success in each is p.

# Probability

The binomial distribution is useful and comes up pretty often.

When p = .5, the binomial PF is symmetric.



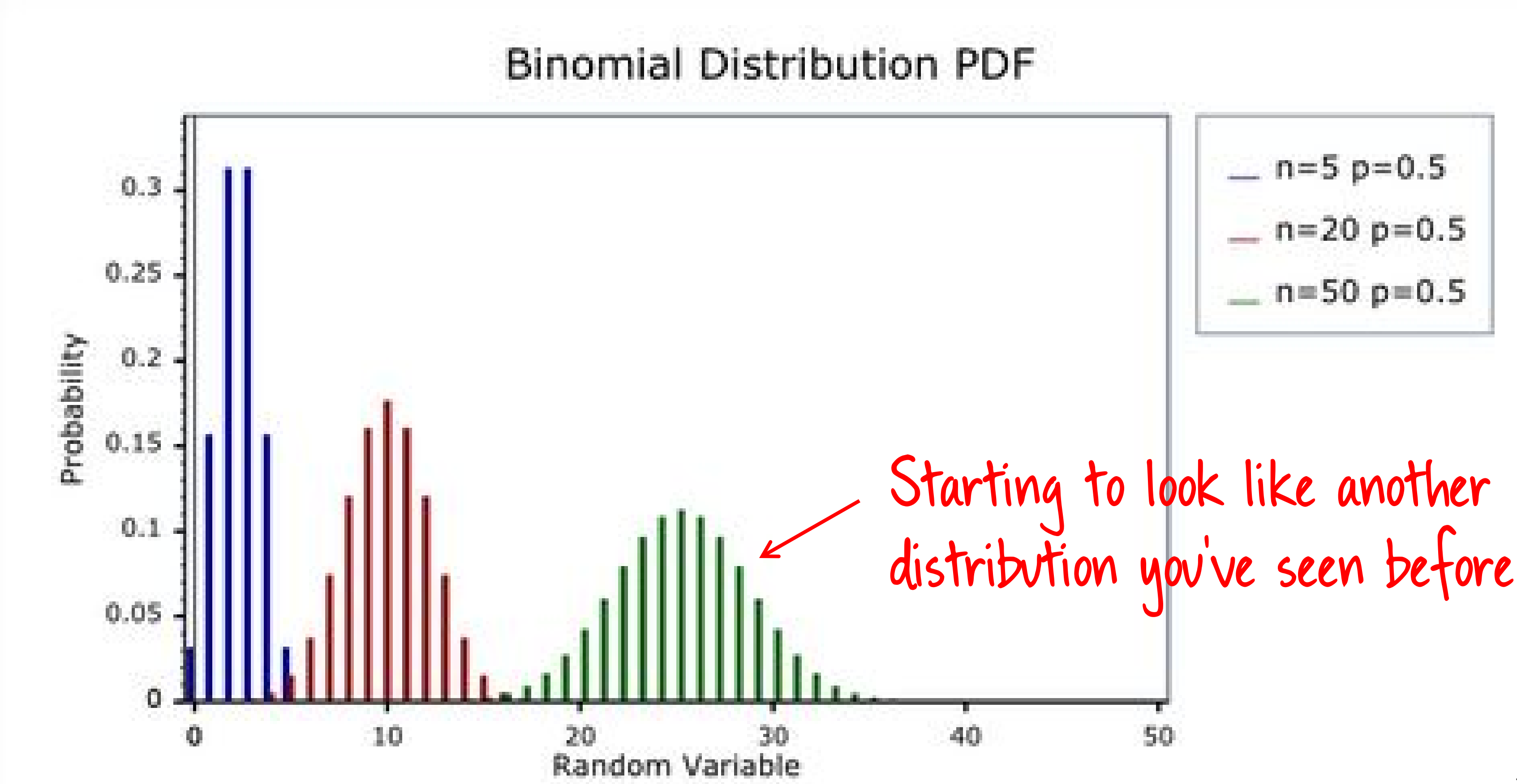



© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

# Probability

The binomial distribution is useful and comes up pretty often.

When p = .5, the binomial PF is symmetric.




Starting to look like another distribution you've seen before



© Source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see https://ocw.mit.edu/help/faq-fair-use/

# Probability---random variables

More formally, the probability function (PF) of X, where X is a discrete random variable, is the function $f_X$ such that for any real number x, $f_X(x) = P(X=x)$.

The probability function has properties induced by our earlier definition of a probability. In particular,

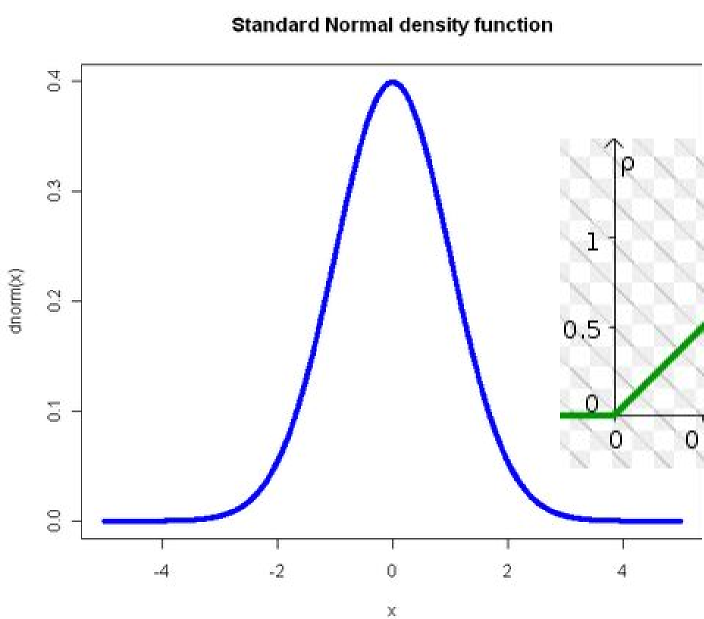
$0 <= f_X(x_i) <= 1$

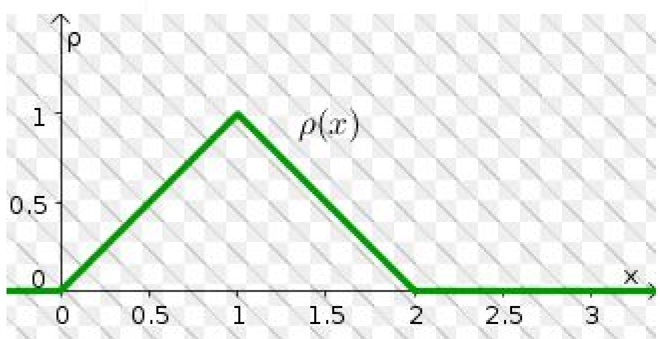
$\Sigma_i f_X(x_i) = 1$

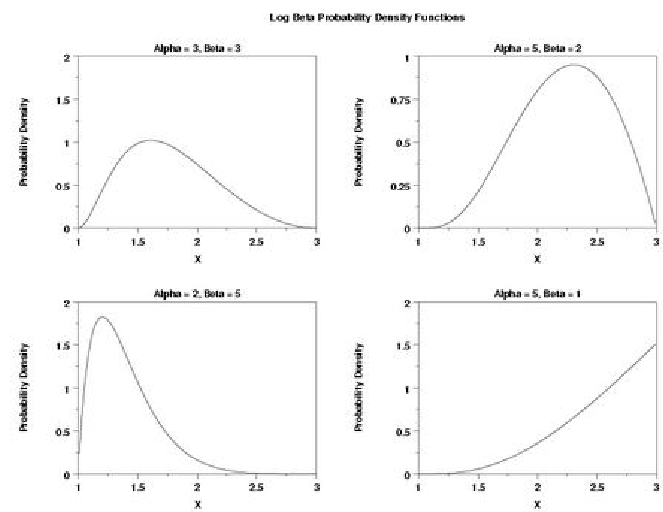
$P(A) = P(X \subset A) = \Sigma_A f_X(x_i)$

# Probability--random variables

For continuous random variables, we rarely start with a verbal description that we use to calculate probabilities. Instead, we are typically given a function, called a density, that describes the probability that the random variable is in various regions. You've seen these.



# Probability---random variables

The density, or probability density function (PDF) is the continuous analog to the discrete PF in many ways. We'll talk about how they're similar and different, but first a definition. Well, in fact, we define continuous random variables in terms of this function.

A random variable $X$ is continuous if there exists a non-negative function $f_X$ such that for any interval $A \subset R$,

$$P(X \subset A) = \int_A f_X(x)dx$$

# Probability---random variables

The density, or probability density function (PDF) is the continuous analog to the discrete PF in many ways. We'll talk about how they're similar and different, but first a definition. Well, in fact, we define continuous random variables in terms of this function.

A random variable X is continuous if there exists a non-negative function $f_X$ such that for any interval $A \subset R$,

$$P(X \subset A) = \int_A f_X(x)dx$$

This is the PDF

# Probability---random variables

Just like the PF, the PDF has properties induced by our earlier definition of a probability. In particular,

$0 <= f_X(x)$

$\int f_X(x) = 1$

$P(A) = P(a <= X <= b) = \int_A f_X(x)dx$

Note the value of a PDF at a particular x does not have the same interpretation as a probability. In fact, $P(X=x) = 0$ for any x if X is continuous.

# Probability---random variables

Just like the PF, the PDF has properties induced by our earlier definition of a probability. In particular,

$0 <= f_X(x)$ It can be greater than 1

$\int f_X(x) = 1$

$P(A) = P(a <= X <= b) = \int_A f_X(x)dx$

Note the value of a PDF at a particular x does not have the same interpretation as a probability. In fact, $P(X=x) = 0$ for any x if X is continuous.

# Probability---random variables

Just like the PF, the PDF has properties induced by our earlier definition of a probability. In particular,

$0 <= f_X(x)$

$\int f_X(x) = 1$ Integrates to 1 instead of summing to 1.

$P(A) = P(a <= X <= b) = \int_A f_X(x)dx$

Note the value of a PDF at a particular x does not have the same interpretation as a probability. In fact, $P(X=x) = 0$ for any x if X is continuous.

# Probability---random variables

Just like the PF, the PDF has properties induced by our earlier definition of a probability. In particular,

$0 <= f_X(x)$

$\int f_X(x) = 1$

$P(A) = P(a <= X <= b) = \int_A f_X(x)dx$

Again, an integral

Note the value of a PDF at a particular x does not have the same interpretation as a probability. In fact, $P(X=x) = 0$ for any x if X is continuous.

# Probability---random variables

## A word about terminology and notation

There are a lot of different types of functions to keep track of (even something called a "random variable" is actually a function), and more to come. I have tried to use what I think is the most standard and accepted terminology and notation for these. I will also try to be consistent. Do, please, understand that notation and terminology vary quite a bit by source, and sometimes we will use less formal terms, like saying "distribution of a random variable" for either a PF or a PDF.

# Probability---example

Let a, b be real numbers such that $a < b$.

Let $S = \{x : a <= x <= b\}$.

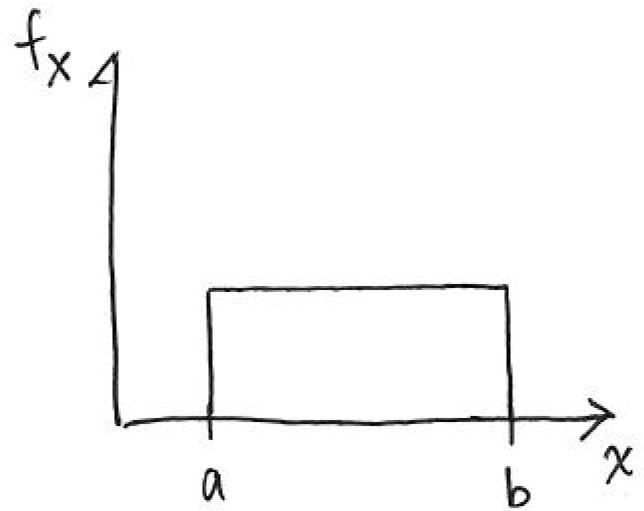
Suppose X is defined in such a way that the probability of X belonging to any subinterval of S is proportional to the length of the subinterval. Then,

$$f_X(x) = \begin{cases} 1/(b-a) & a <= x <= b \\ 0 & \text{otherwise} \end{cases}$$

We call this random variable X "uniform with parameters a & b," denoted $X \sim U[a,b]$.

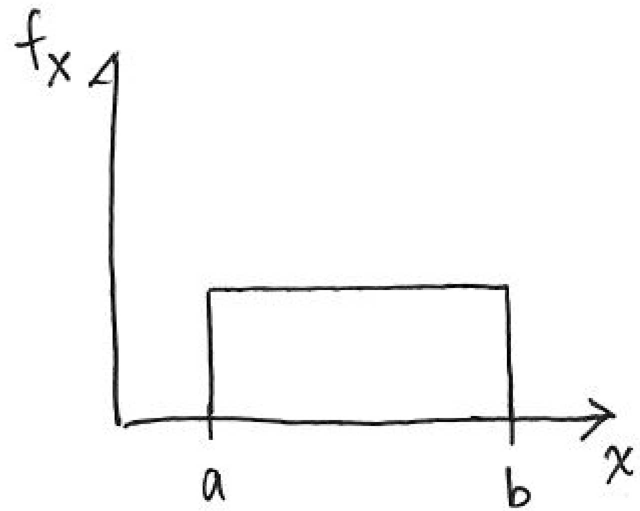
# Probability---example

Here's a picture:



# Probability---example

Here's a picture:




If you want to compute the probability of a U[a,b] random variable being in some interval [c,d] in [a,b], you can just integrate 1/(b-a) over that region. Or, since the PDF is flat, you can just use (d-c)/(b-a).

# Probability---random variables

Sometimes it's handy to be able to express probabilities related to a random variable in an alternative form. Doubly handy is the fact that this alternative form has the same definition regardless of whether the random variable is discrete or continuous.

The cumulative distribution function (CDF) $F_X$ of a random variable X is defined for each x as

$F_X(x) = P(X <= x)$.

# Probability---random variables

Properties of probability imply certain things about CDFs:

$0 <= F_X(x) <= 1$

$F_X(x)$ is non-decreasing in x

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

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

$F_X(x)$ is right continuous

Note that CDFs are continuous everywhere for continuous random variables and have jumps for discrete random variables.

# Probability---random variables

A PF/PDF and a CDF for a particular random variable contain exactly the same information about its distribution, just in a different form. It stands to reason, then, that given the PF/PDF, one could recover the CDF and vice versa. This is true. Here's how to do it with continuous random variables:

$$F_X(x) = P(X <= x) = \int_{-\infty}^{x} f_X(x)dx$$

$$F'X(x) = \frac{dF(x)}{dx} = f_X(x)$$

# Probability--random variables

A PF/PDF and a CDF for a particular random variable contain exactly the same information about its distribution, just in a different form. It stands to reason, then, that given the PF/PDF, one could recover the CDF and vice versa. This is true. Here's how to do it with continuous random variables:

$$F_X(x) = P(X <= x) = \int_{-\infty}^{x} f_X(x)dx$$

$$F'X(x) = \frac{dF(x)}{dx} = f_X(x)$$

This is true provided X continuous and F' exists at all but a finite number of points

# Probability---joint distributions

When we were talking about probability last lecture, I said that, going forward, it was going to be important for us to be able to discuss relationships between stochastic events. We then talked about independence and conditional probability. We will talk about analogous concepts in the context of random variables, but first we must a <u>joint distribution</u>. (In the case where only two random variables are involved, we call them <u>bivariate distributions</u>.)

# Probability---joint distributions

Why?

Might be interested in the relationship and joint behavior of two or more random variables.

--rainfall and crop growth

--length of the regular checkout line and the length of the express checkout line

--the number of veg toppings and the number of non-veg toppings on my pizza

--dollar/euro exchange rate and the stock price of an exporting firm

# Probability---joint distributions

If X and Y are continuous random variables defined on the same sample space S, then the joint probability density function of X&Y, $f_{XY}(x,y)$, is the surface such that for any region A of the xy-plane,

$$P((X,Y) \subset A) = \iint_A f_{XY}(x,y)\,dx\,dy$$

Like before, properties of probability imply certain properties of the joint PDF, such as it must integrate to 1 over the xy-plane, and any individual point or one-dimensional curve has probability zero.

# Probability---joint distributions

(The analogous joint PF exists for discrete random variables,

$$f_{XY}(x,y) = P(X=x \text{ and } Y=y),$$

but I won't say anything else about it right now.
Sometimes I will just give definitions in terms of continuous random variables.)

# Probability---example

Suppose after hours of writing lecture notes, I develop a splitting headache. I rummage around in my drawer and find one tablet of naproxen and one of acetaminophen. I take both. Let X be the effective period of naproxen. Let Y be the effective period of acetaminophen. Suppose

$$f_{XY}(x,y) = \lambda^2 \exp\{-\lambda(x+y)\} \quad \text{for } x,y >= 0$$

What is the probability that my headache comes back within three hours?

# Probability---example

What is the probability that my headache comes back within three hours, i.e., P(X <= 3 and Y <= 3)?

$$= \int_0^3 \int_0^3 \lambda^2 \exp\{-\lambda(x+y)\} dy dx$$

$= \ . \ . \ .$

$$= (1-\exp\{-3\lambda\})^2$$

# Probability---example

What is the probability that my headache comes back within three hours, i.e., P(X <= 3 and Y <= 3)?

$$= \int_0^3 \int_0^3 \lambda^2 \exp\{-\lambda(x+y)\}dydx$$

$$= \ . \ . \ .$$

$$= (1-\exp\{-3\lambda\})^2$$

These dots mean that you can work out the details at home.

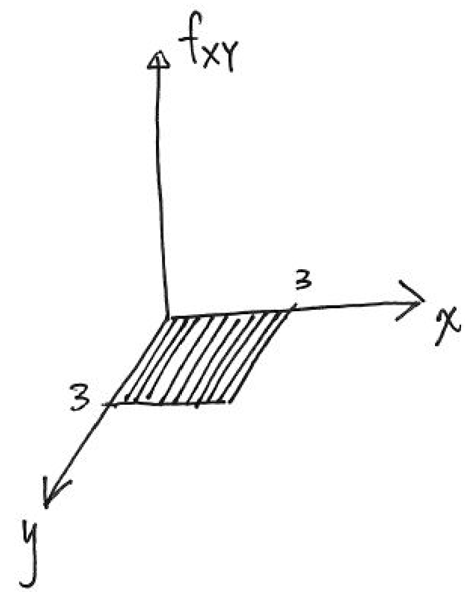
# Probability---example

What is the probability that my headache comes back within three hours, i.e., P(X <= 3 and Y <= 3)?

$$= \int_0^3 \int_0^3 \lambda^2 \exp\{-\lambda(x+y)\} dy dx$$

$$= \ldots$$

$$= (1-\exp\{-3\lambda\})^2$$




This is the region over which we are integrating.

# Probability---example

What if I only took the acetaminophen after the naproxen stopped working?

# Probability---example

What if I only took the acetaminophen after the naproxen stopped working? Now I'm asking about P(X+Y <= 3).

$$= \int_0^3 \left[ \int_0^{3-x} \lambda^2 e^{-\lambda x} e^{-\lambda y} \, dy \right] dx$$

$$= \ . \ . \ .$$

$$= 1 - (1 + 3\lambda) e^{-3\lambda}$$

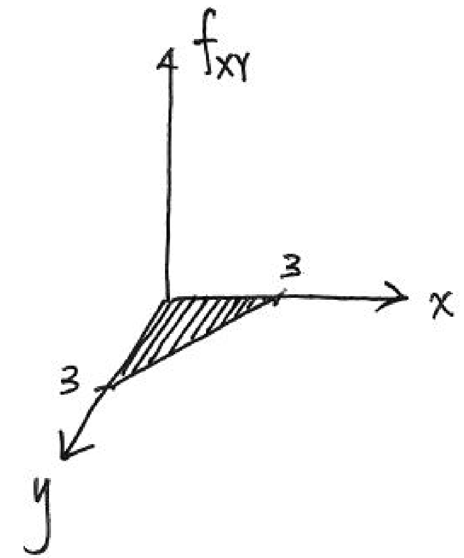
# Probability---example

What if I only took the acetaminophen after the naproxen stopped working? Now I'm asking about P(X+Y <= 3).

$$= \int_0^3 \left[ \int_0^{3-x} \lambda^2 e^{-\lambda x} e^{-\lambda y}\, dy \right] dx$$

$$= \ldots$$

$$= 1-(1+3\lambda)e^{-3\lambda}$$




This is the region over which we are integrating.

# Probability---example

What if I defined a new random variable $Z$ = total effective life of naproxen and acetaminophen taken sequentially = $X + Y$.

What is $F_Z(z)$? We've already computed it! $F_Z(z) = P(Z <= z) = P(X+Y <= z) = 1- (1 + z\lambda)\exp\{- z\lambda\}$, for $z > 0$.

What is $f_Z(z)$? Well, just take the derivative. $f_Z(z) = F'_Z(z) = \lambda^2 z\exp\{- z\lambda\}$, for $z > 0$.

MIT OpenCourseWare
https://ocw.mit.edu/

14.310x Data Analysis for Social Scientists
Spring 2023

For information about citing these materials or our Terms of Use, visit: https://ocw.mit.edu/terms.