14.30 Introduction to Statistical Methods in Economics
Spring 2009

14.30 Introduction to Statistical Methods in Economics
Lecture Notes 6

Konrad Menzel

February 24, 2009

1 Examples

Suppose that a random variable is such that on some interval $[a, b]$ on the real axis, the probability of X belonging to some subinterval $[a', b']$ (where $a \le a' \le b' \le b$) is proportional to the length of that subinterval.

Definition 1 *A random variable X is* uniformly *distributed on the interval $[a, b]$, $a < b$, if it has the probability density function*

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

In symbols, we then write

$$X \sim U[a, b]$$

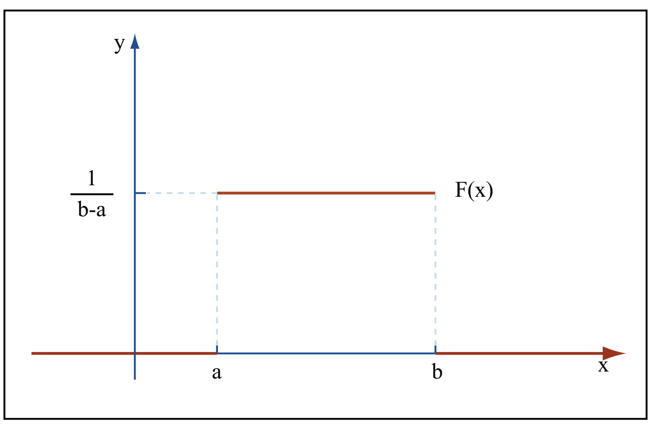

Image by MIT OpenCourseWare.

Figure 1: p.d.f for a Uniform Random Variable, $X \sim [a, b]$

For example, if $X \sim U[0, 10]$, then

$$P(3 < X < 4) = \int_3^4 f(t)dt = \int_3^4 \frac{1}{10}dt = \frac{1}{10}$$

What is $P(3 \le X \le 4)$? Since the probability that $P(X = 3) = 0 = P(X = 4)$, this is the same as $P(3 < X < 4)$.

Example 1 *Suppose X has p.d.f.*

$$f_X(x) = \begin{cases} ax^2 & \text{if } 0 < x < 3 \\ 0 & \text{otherwise} \end{cases}$$

What does a have to be? - since $P(X \in \mathbb{R}) = 1$, the density has to integrate to 1, so a must be such that

$$1 = \int_{-\infty}^{\infty} f_X(t)dt = \int_0^3 at^2 dt = \left[\frac{ax^3}{3}\right]_0^3 = \frac{27}{3}a - 0 = 9a$$

Therefore, $a = \frac{1}{9}$.
What is $P(1 < X < 2)$? - let's calculate the integral

$$P(1 < X < 2) = \int_1^2 \frac{t^2}{9} dt = \frac{2^3}{9 \cdot 3} - \frac{1^3}{9 \cdot 3} = \frac{7}{27}$$

What is $P(1 < X)$?

$$P(1 < X) = \int_1^{\infty} f_X(t)dt = \int_1^3 \frac{t^2}{9} dt = \frac{27 - 1}{27} = \frac{26}{27}$$

1.1 Mixed Random Variables/Distributions

Many kinds of real-world data exhibit point masses at some values mainly for two different reasons:

- some outcomes are restricted to certain values mechanically, so a lot of probability mass tends to cumulate right at the corners of the range of the random variable, e.g. daily rainfall can possibly take any positive real value, but there are many days at which rainfall is zero.
- individuals taking economic decisions may respond to certain institutional rules by positioning themselves right at some kind of kinks or discontinuities, e.g. if we look at incomes reported to Social Security or the Internal Revenue Service, we observe "bunching" of individuals at the top ends of the tax brackets (since for those individuals, a small increase in income would mean a discrete jump in the tax rate).

The corresponding distributions are, strictly speaking, *not continuous*, because even though realizations can be any real-valued numbers, we can't define a probability density function as we did in the previous section, but we'll have to deal with the point masses separately. Some of this is going to come up in your econometrics classes, but we won't spend time on this for now and only look at one example.

Example 2 *The following graph is constructed using data from the Current Population Survey (CPS) for 1979.*[1]
For the graph, the authors chose a subpopulation with very low income so that the fraction of the sample for whom the minimum wage was binding was relatively high. There are some individuals to the left of the 1979 value of the minimum wage presumably corresponding to employment in sectors which are in part exempt from minimum wage laws (e.g. farming, youth employment).

[1] Figure 3b) on p. 1017 in DiNardo, J., N. Fortin and T. Lemieux. "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach." *Econometrica* 64, no. 5 (1996): 1001-1044.

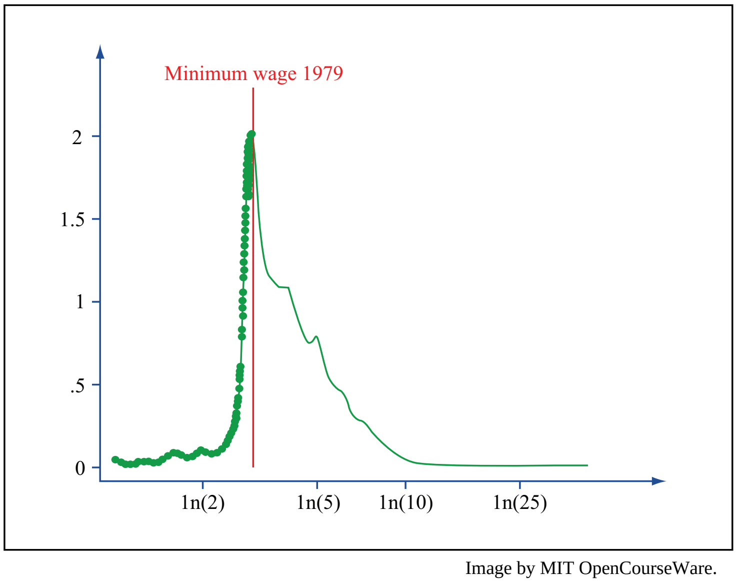

Figure 2: Log Wages for Female High-School Dropouts in 1979

2 The Cumulative Distribution Function (c.d.f.)

Definition 2 *The* cumulative distribution function (c.d.f.) F_X *of a random variable* X *is defined for each real number as*

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

Note that this definition is the same for discrete, continuous or mixed random variables. In particular, since we allow for X to be discrete, note that $P(X \leq x)$ may be different from $P(X < x)$, so it's important to distinguish the corresponding events. In the definition of the c.d.f., we'll always use X "less or equal to" x.
Since the c.d.f. is a probability, it inherits all the properties of probability functions, in particular

Property 1 *The c.d.f. only takes values between 0 and 1*

$$0 \leq F_X(x) \leq 1 \quad \text{for all } x \in \mathbb{R}$$

Also, since for $x_1 < x_2$, the event $X \leq x_1$ is *included* in $X \leq x_2$, we have

Property 2 F_X *is nondecreasing in* x*, i.e.*

$$F_X(x_1) \leq F_X(x_2) \quad \text{for } x_1 < x_2$$

If we let $x \to -\infty$, the event $(X \leq x)$ becomes "close" (here I'm very sloppy about what that means) to the impossible event in terms of its probability of occurring, whereas if $x \to \infty$, the event $(X \leq x)$ becomes almost certain, so that we have

Property 3

$$\lim_{x \to -\infty} F(x) = 0$$
$$\lim_{x \to \infty} F(x) = 1$$

Note that a c.d.f. doesn't have to be continuous: if we define the *left limit*

$$F(x^-) = \lim_{h>0, h\to 0} F(x-h)$$

and the *right limit*

$$F(x^+) = \lim_{h>0, h\to 0} F(x+h)$$

Recall that in order to be continuous at x, $F(x)$ must satisfy $F(x^-) = F(x^+)$. This need not be true in general, as the following example shows:

Example 3 *Consider again the experiment of rolling a die, where the random variable X corresponds to the number we rolled. Then the c.d.f. of X is given by*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 1 \\ \frac{1}{6} & \text{if } 1 \le x < 2 \\ \cdots & \cdots \\ \frac{5}{6} & \text{if } 5 \le x < 6 \\ 1 & \text{if } x \ge 6 \end{cases}$$

which has discontinuous jumps at the the values $1, 2, \ldots, 6$.

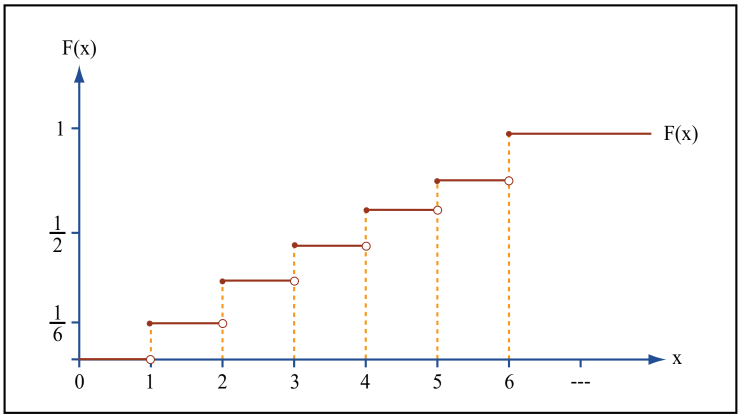

Figure 3: c.d.f. of a die roll

However, by a general result from real analysis, any monotone function (hence the c.d.f. F_X in particular) can only have *countably many points of discontinuity.*

Furthermore, we always have

Property 4 *Any c.d.f. is right-continuous, i.e.*

$$F(x) = F(x^+)$$

We can now use our knowledge about probabilities to show some more properties of c.d.f.s

Proposition 1 *For any given x,*

$$P(X > x) = 1 - F_X(x)$$

PROOF: From properties of probabilities,

$$P(X > x) = 1 - P(X \leq x) = 1 - F_X(x)$$

Similarly,

Proposition 2 *For two real numbers $x_1 < x_2$,*

$$P(x_1 < X \leq x_2) = F_X(x_2) - F_X(x_1)$$

Proposition 3 *For any x,*

$$P(X < x) = F(x^-)$$

Proposition 4 *For any x,*

$$P(X = x) = F(x^+) - F(x^-)$$

This last result means in particular that for continuous variables, $P(X = x) = 0$ for all values of x.

Example 4 *Let's check whether the function $G_X(x)$ in the following graph is a c.d.f. The function is*

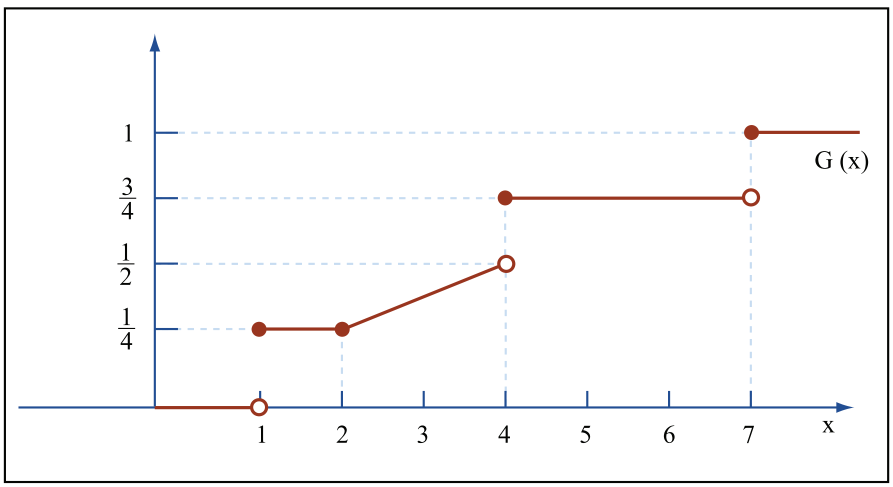

Image by MIT OpenCourseWare.

between 0 and 1, monotonically increasing, and right-continuous. Let's apply the last four propositions to this example (just reading numbers off the graph):

- $P(X > 4) = 1 - F(4) = 1 - \frac{3}{4} = \frac{1}{4}$
- $P(3 < X \leq 4) = \frac{3}{4} - \frac{1}{4} = \frac{1}{2}$
- $P(X < 4) = F(4^-) = \frac{1}{2}$
- $P(X = 4) = F(4^+) - F(4^-) = \frac{3}{4} - \frac{1}{2} = \frac{1}{4}$

Example 5 *Unlike for continuous random variables, where we have a one-line formula linking the p.d.f. and the c.d.f., in discrete case, have to use the results on deriving probabilities from c.d.f.s we just discussed. Let's look at the relationship in another graphical example*

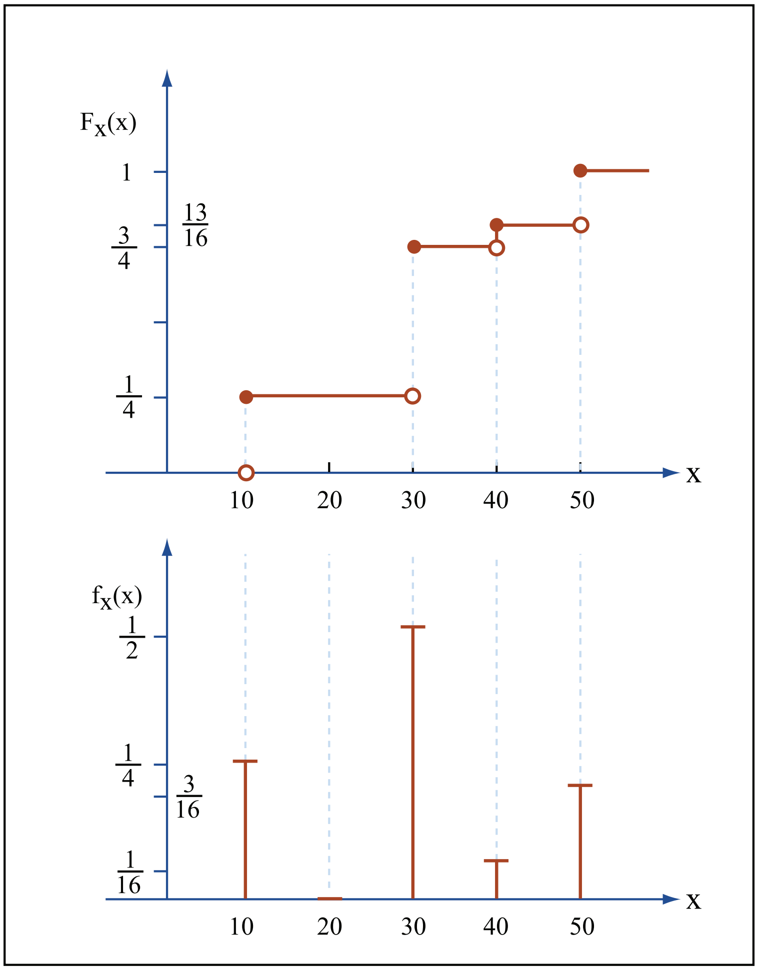

Figure 4: c.d.f. and p.d.f. for a discrete random variable

2.1 p.d.f. and c.d.f for Continuous Random Variables

If X has a continuous distribution with p.d.f. $f(x)$ and $F(x)$ (I'll drop the X subscript from now on wherever there are no ambiguities), then

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

From the fundamental theorem of calculus, we can in this case write the relationship between c.d.f. and p.d.f. as

$$F'(x) = \frac{d}{dx}F(x) = f(x)$$

Example 6 *Let*

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x}{1+x} & \text{if } x \geq 0 \end{cases}$$

Is $F(x)$ a c.d.f.? - let's check basic properties:

- $\lim_{x\to-\infty} F(x) = 0$
- $\lim_{x\to\infty} F(x) = 1$
- *$F(\cdot)$ is nondecreasing (can check derivative below)*

What is the p.d.f. $f(x)$?

$$f(x) = F'(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{(1+x)^2} & \text{otherwise} \end{cases}$$

Is $f(x)$ a p.d.f.? - well, we've essentially already shown that $F(x)$ is a c.d.f. We can see right away that

$$f(x) \geq 0 \quad \text{for all } x$$

and also,

$$\int_{-\infty}^{\infty} f(t)dt = \lim_{x\to\infty} F(x) - \lim_{x\to-\infty} F(x) = 1 - 0 = 0$$

Example 7 *If $X \sim U[0,1]$, then its c.d.f. is*

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

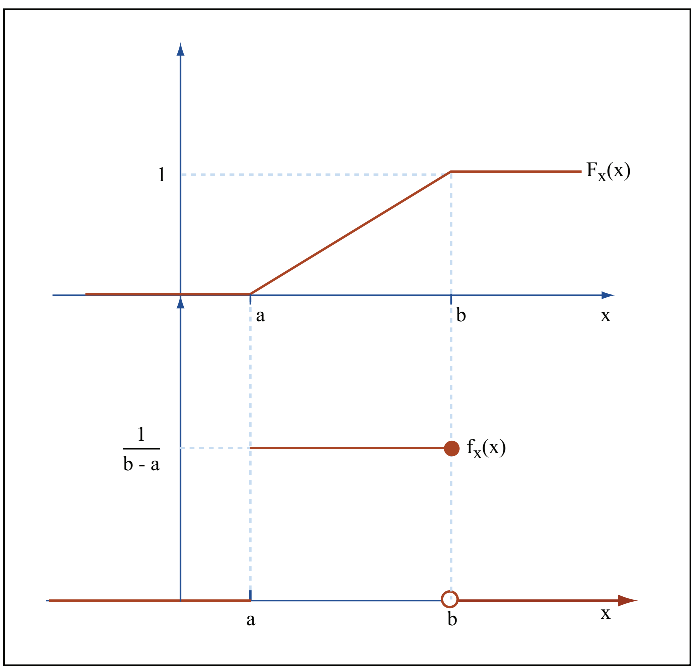

Image by MIT OpenCourseWare.

Figure 5: p.d.f. and c.d.f. for $X \sim U[a,b]$

3 Joint Distributions of 2 Random Variables X, Y

In many situations, we are interested not only in a single random variable, but may care about relationship between two or more variables, e.g. whether the outcome of some process affects the outcome of another. E.g. we could look at

- IQs of identical twins - i.e. X would be one kid's IQ, and Y that of her/his sibling
- educational attainment X and income Y: while we could look at the distributions of income or education separately, we can also plot both variables together for observations from a data set. And in the graph it looks like there is in fact a non-trivial relationship between the variables.

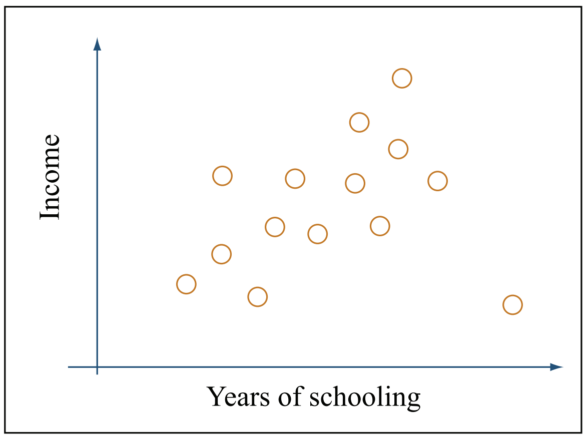

Image by MIT OpenCourseWare.

Figure 6: Schooling and Income

- relapse times: since it is often not possible to remove a cancer completely by surgery, we may want to evaluate the effectiveness of a medical procedure, by looking at how long it takes until either (a) a new operation becomes necessary (X), or (b) the patient dies (Y). While we are interested in either outcome, both outcomes are interdependent: if the patient dies before a new operation, we simply don't observe when he would have had to undergo surgery otherwise.

In this part of the class, we will consider the properties of two (or more) random variables simultaneously, including their relationship. We will also introduce concepts analogous to "independence" and "conditional probabilities" of events.

We let (X, Y) be a pair random variables that (jointly) takes values (x, y), and either variable can be continuous, discrete, or mixed.

3.1 Discrete Random Variables

In the discrete case, the joint p.d.f. is given by

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

for any $(x, y) \in \mathbb{R}^2$. If $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ contains all possible values of (X, Y), then

$$\sum_{i=1}^{n} f_{XY}(x_i, y_i) = 1$$

For any subset $A \subset \mathbb{R}^2$,

$$P((X,Y) \in A) = \sum_{(x,y)\in A} f_{XY}(x,y)$$

Example 8 *In a supermarket, let X be the number of people in the regular checkout line, and Y the number of people in the express line. Then the joint p.d.f. of X and Y could look like this: A table of this*

f_{XY}		Y					
		0	1	2	3	4	Total
	0	0.1	0.05	0.05	0	0	0.2
X	1	0.05	0.2	0.2	0.05	0	0.5
	2	0	0	0.1	0.1	0.05	0.25
	3	0	0	0	0	0.05	0.05
		0.15	0.25	0.35	0.15	0.1	1

form, summarizing the cell-probabilities from the joint p.d.f. of (X,Y), and the marginal probabilities on the sides, is also called a contingency table. *As argued before, the probabilities in the table should add up to one, and they do.*
We can see from the entries that there seems to be some relationship between the two variables: when the number of individuals at the regular checkout is high, then the number of persons in the express line also tends to be high.
We can also calculate probabilities for different events based on the p.d.f. as given in the table:

$$\begin{aligned}
P(X=2) &= 0+0+0.1+0.1+0.05 = 0.25 \\
P(X \geq 2, Y \geq 2) &= \sum_{x=2}^{3}\sum_{y=2}^{4} f_{XY}(x,y) = 0.1+0.1+0.05+0+0+0.05 = 0.3 \\
P(|X-Y| \leq 1) &= P(X=Y) + P(|X-Y|=1) \\
&= 0.1+0.2+0.1+0+0.05+0.05+0.2+0+0.1+0+0.05 = 0.85
\end{aligned}$$