Statistics Review

Christopher Taber

Wisconsin

Spring Semester, 2011

Outline

Outline

Random variables

Lets forget about the details that arise in dealing with data for a while. Most objects that economists think about are random variables.

Informally, a *random variable* is a numerical outcome or measurement with some element of chance about it. That is, it makes sense to think of it as having possibly had some value *other* than what is observed.

Econometrics is a tool that allows us to learn about these random variables from the data at our disposal.

Examples of random variables:

- Gross Domestic Product
- Stock Prices
- Wages of Workers
- Years of Schooling Attained by Students
- Numeric Grade in a Class
- Number of Job Offers Received
- Demand for a new product at a given price

From one perspective, there are two types of random variables, *discrete* and *continuous*.

A discrete random variable can only take on a finite number of values. Days worked last week is a nice example. It can only take on the 8 different values, 0 to 7.

By contrast, a continuous random variable takes on a continuum of values. Literally, this would mean there are an infinite number of values that the variable can take; but often a variable that can take on a very large number of values is treated as continuous because it is convenient.

Most random variables we will think about are approximately continuous, but we will start with a consideration of the characterization of discrete random variables because it is easier to follow.

Outline

Probability Density Functions

Suppose that X is a random variable that takes on J possible values $x_1, x_2, \ldots x_J$.

The probability density function (pdf), $f(\cdot)$ of X is defined as:

$$f(x_j) = \Pr(X = x_j)$$

Some conventions: capital letters are used to denote the variable, small letters realizations or possible values; a pdf is a lower-case letter (often f)

Now it follows that if X can only take on the values $x_1, x_2, \ldots x_J$, we have

$$\sum_{j=1}^{J} f(x_j) = 1$$

So, for example, if we regard grades as a random variable, and assign a 4 to A's, 3 to B's, etc., we might have

$$\begin{aligned} f(4) &= 0.30 \\ f(3) &= 0.40 \\ f(2) &= 0.25 \\ f(1) &= 0.04 \\ f(0) &= 0.01 \end{aligned}$$

Summarizing this full distribution is very complicated if it takes on many values.

We need some way of characterizing it. The first question we might ask is "Is this typically a big number or a small number?"

Outline

The expectation is also called the mean or the average.

$$E(X) = \sum_{j=1}^{N} x_j f(x_j)$$

(The expectation is thought of as a ‘typical value’ or measure of central tendency, though it has shortcomings for each of these purposes.)

Some properties of expected values:

1. If b is a nonstochastic (not random),
$$E(b) = b.$$
2. If X and Y are two random variables,
$$E(X + Y) = E(X) + E(Y)$$
3. If a is nonstochastic,
$$E(aX) = aE(X).$$

Also,
$$E(aX + b) = aE(x) + b.$$

However in general
$$E(g(X)) \neq g(E(X))$$

An example: Grades

Using the distribution from the example above, the expected grade is:

$$\begin{aligned} E(G) =& 0.3 \times 4 + 0.4 \times 3 + 0.25 \times 2 + 0.04 \times 1 + 0.01 \times 0 \\ =& 2.94 \end{aligned}$$

Interpretation of Expectation as Bet

One interpretation of an expectation is the value of a bet. I would break even if the expected value of the bet was zero.

Example 1: Coin Flip

I get

$$\begin{cases} \$1.00 & \text{if heads} \\ -\$1.00 & \text{if tails} \end{cases}$$

Expected payoff is

$$0.5 \times 1 + 0.5 \times -1 = 0.$$

Example 2: Die Roll

Suppose you meet a guy on the street who charges you $3.50 to play a game. He rolls a die and gives you that amount in dollars, i.e. if he rolls a 1 you get $1.00, etc. Is this a good bet?

The expected payoff from the bet is

$$\begin{aligned} E(Y - 3.5) &= \frac{1}{6}1 + \frac{1}{6}2 + \frac{1}{6}3 + \frac{1}{6}4 + \frac{1}{6}5 + \frac{1}{6}6 - 3.5 \\ &= 0 \end{aligned}$$

Example 3: Occupation

Suppose you are choosing between being a doctor or a lawyer. You may choose to go into the profession where the expected earnings are highest.

Outliers and the Expectation

Outliers are very influential in expected values. Suppose you are a high school basketball player with a remote chance of becoming Lebron James. Your distribution of income might look like

$$Y = \begin{cases} \$20,000 & \frac{1}{3} \\ \$30,000 & \frac{199}{300} \\ \$50,000,000 & \frac{1}{300} \end{cases}$$

Then $E(Y) = \$166,748.$

One point made by this example is that the mean or expected value is not always a good representation of the typical value or central tendency. The median–the value at or above which half the realization fall–is \$30,000 in the example and is arguably more typical or central.

Estimation of Expected Value

You should have learned that the way we estimate the expected value is to use the sample mean

That is suppose we want to estimate $E(X)$ from a sample of data $X_1, X_2, ..., X_N$

We estimate using

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Example

Suppose data is (as in Wooldridge Example C.1)

City	Unemployment Rate
1	5.1
2	6.4
3	9.2
4	4.1
5	7.5
6	8.3
7	2.6
8	3.5
9	5.8
10	7.5

In this case

$$\bar{X} = \frac{5.1 + 6.4 + 9.2 + 4.1 + 7.5 + 8.3 + 2.6 + 3.5 + 5.8 + 7.5}{10}$$

$$= 6.0$$

A fact about sample means

There is a particular feature of sample means that we will use a lot in our course.

Forget about random variables for now and just think about the algebra.

Notice that for any variable a,

$$\begin{aligned}\sum_{i=1}^{N} a(X_i - \bar{X}) &= \sum_{i=1}^{N} aX_i - \sum_{i=1}^{N} a\bar{X} \\ &= aN\left[\frac{1}{N}\sum_{i=1}^{N} X_i\right] - Na\bar{X} \\ &= Na\bar{X} - Na\bar{X} \\ &= 0\end{aligned}$$

Here is one important example of this

$$
\begin{aligned}
s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})[X_i - \bar{X}] \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})X_i - \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})\bar{X} \\
&= \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})X_i
\end{aligned}
$$

Outline

Variance

As a first approximation, as the mean is a measure of central tendency, the variance is a measure of dispersion.

In deciding to make a bet, which occupation to pursue, or which stocks to buy, we might care not only about the expected payoff, but also its variability. 'Variability' captures some notion of *risk.*

The variance of a random variable is given by:

$$\begin{aligned} Var(X) =& E(X - \mu_X)^2 \\ =& \sum_{j=1}^{J} (x_j - \mu_X)^2 f(x_j) \end{aligned}$$

where $\mu_X = E(X)$.

If X is nonstochastic $x_i = \mu_X$ for all i, so $Var(X) = 0$.

$$\begin{aligned} Var(aX+b) &= E(aX+b-(a\mu_X+b))^2 \\ &= E(aX-a\mu_X)^2 \\ &= a^2E(X-\mu_X)^2 \\ &= a^2Var(X) \end{aligned}$$

The standard deviation is the square root of the variance.

Interpretation of Variance

The higher the variance, the less confident you are about whether the outcome will be near the mean (or expectation).

Suppose you bet on a coin toss. Case 1:

$$X = \begin{cases} \$1.00 & heads \\ -\$1.00 & tails \end{cases}$$

$$\begin{aligned} E(X) &= 1 \times \frac{1}{2} + (-1) \times \frac{1}{2} = 0 \\ V(X) &= (1-0)^2 \frac{1}{2} + (-1-0)^2 \frac{1}{2} = 1 \end{aligned}$$

Case 2, you bet \$10,000:

$$X = \begin{cases} \$10,000 & heads \\ -\$10,000 & tails \end{cases}$$

$$E(X) = 10,000 \times \frac{1}{2} + (-10,000) \times \frac{1}{2} = 0$$

$$V(X) = (10,000 - 0)^2 \frac{1}{2} + (-10,000 - 0)^2 \frac{1}{2} = 100,000,000$$

In considering a bet, an investment in a risky project, a 'life choice' (education, occupation, marriage, etc.), the variance of the 'payoff' is likely to be relevant.

Outline

Continuous Random Variables

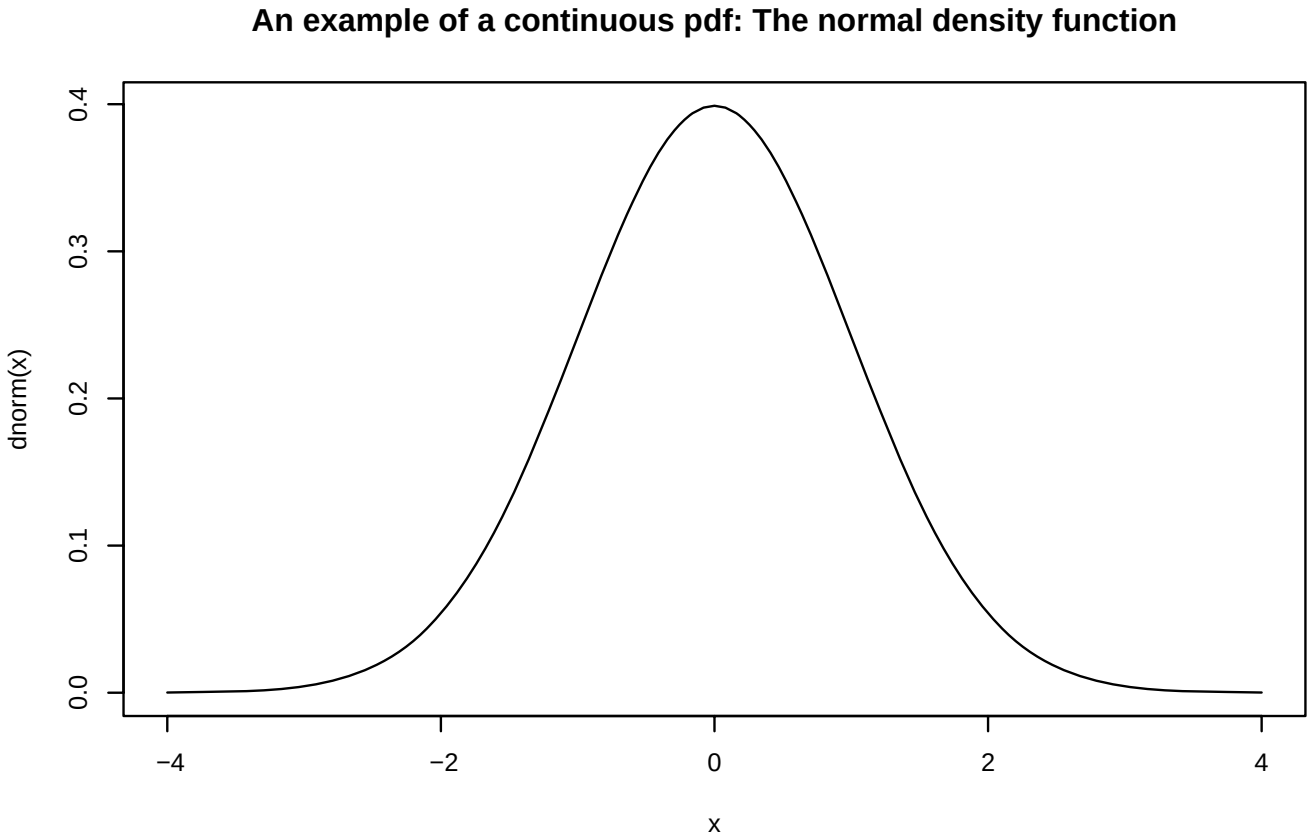
The probability density function shows, heuristically, the 'relative probability' of a value. Since the random variable is continuous it can take on an infinite number of values and no exact value has non–negligible probability. Instead, there is a probability of falling between two points a and b, which is given by

$$\begin{aligned} \Pr(a \leq X \leq b) &= \text{ area under the curve between } a \text{ and b} \\ &= \int_a^b f(x)dx \end{aligned}$$

Thus we also have:

$$f(x) = \lim_{\delta \to 0} \frac{\Pr(x \leq X \leq x + \delta)}{\delta}$$

The particular value of the density is not interesting in its own right. It is only interesting if you integrate over it.


An example of a continuous pdf: The normal density function
dnorm(x)
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
x

Analogous to the properties of the pdf in the discrete case, we have:

$$\int f(x)dx = 1$$

$$E(X) = \int xf(x)dx$$

$$Var(X) = \int (x - E(x))^2 f(x)dx,$$

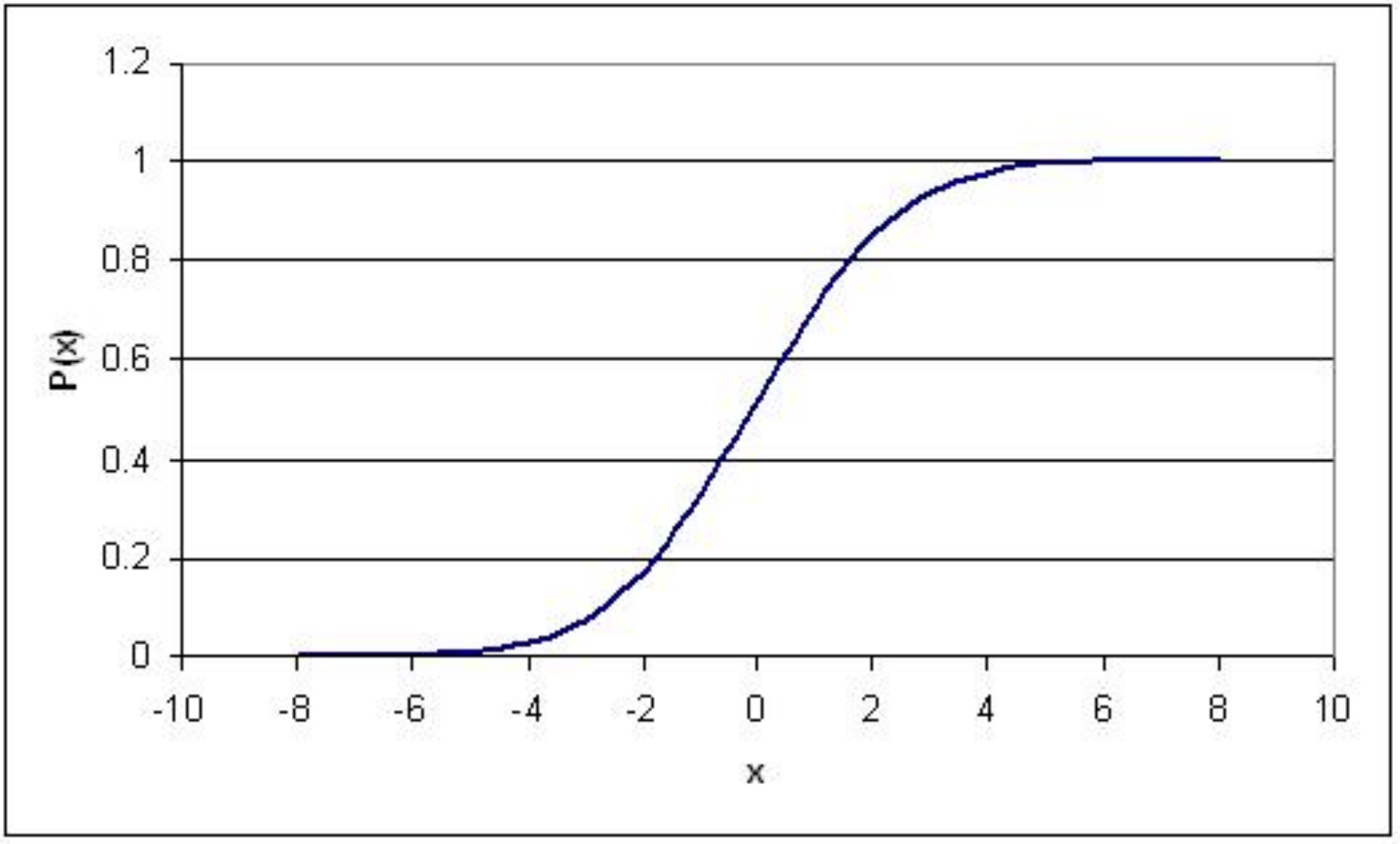
and the properties of expectations that we discussed for the discrete case carry over.

Another representation of the distribution of a random variable is the cumulative distribution function, usually denoted $F(x)$ and defined to be the probability that X falls at or below the value x:

$$F(x) = \Pr(X \leq x)$$

For continuous random variables that can take on any value (i.e. positive $f(x)$ for any x between $-\infty$ and $+\infty$) this can be written as:

$$F(a) = \int_{-\infty}^{a} f(x)dx$$


1.2
1
0.8
0.6
0.4
0.2
0
P(x)
-10
-8
-6
-4
-2
0
2
4
6
8
10
x

Joint Distributions of Random Variables

For two random variables X and Y we can define their *joint* distribution. For discrete random variables

$$f(x,y) = \Pr(X = x, Y = y)$$

For continuous random variables:

$$\Pr(a_1 \leq X \leq a_2, b_1 \leq Y \leq b_2) = \int_{a_1}^{a_2} \int_{b_1}^{b_2} f(x,y) dy dx$$

Outline

Covariance and Correlation

When we have two random variables, the first question one may ask is whether they move together. That is when X is high, is Y high?

One measure of whether two variables move together is the *covariance.*

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

If, when X is large relative to μ_X Y tends to be large relative to μ_Y, then the covariance is positive.

Properties of covariance:

- $Var(X+Y) = Var(X) + Var(Y) + 2\ Cov(X,Y)$
- $Cov(X,X) = Var(X)$
- $Cov(a_1X + b_1, a_2Y + b_2) = a_1a_2Cov(X,Y)$

This last property means that if we change the units of measurement of X and/or Y, the covariance changes.

A measure of association that is 'unit–free' is the correlation coefficient:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}$$

It turns out that ρ is between -1 and 1. When X and Y are independent, so there is 'no relation' between X and Y, ρ is zero; if $\rho > 0$ then X and Y go up and down together, whereas if $\rho < 0$ then when X goes up, Y tends to go down and vice versa.

Examples of Correlation

Positively correlated random variables:

- Years of school, earnings
- Husband's wage, wife's wage
- Stock price, profit of firm
- GDP, country size

Negatively correlated random variables:

- GDP growth, unemployment
- Husband's income, wife's hours worked

Random variables with zero or near–zero correlation

- Number on first die, number on second die
- Michelle Obama's temperature, Number of questions asked in this class today
- Stock gain today, stock gain tomorrow

Correlation versus Causality

The difference between causation and correlation is a key concept in econometrics. We would like to identify causal effects and estimate their magnitude. It is generally agreed that this is very difficult to do; often a causal interpretation can be given that is consistent with results derived from an appropriate statistical procedure; having an economic model is often essential in establishing the causal interpretation.

These issues are confused all the time by politicians and the popular press

For some first thoughts, suppose X and Y are positively correlated.

Case 1: $X \rightarrow Y$ (or $Y \rightarrow X$)

- Money supply $\rightarrow$ Inflation
- Increase in minimum wage $\rightarrow$ Increase in wages
- Retirement $\rightarrow$ Decline in income

Case 2: $X \rightarrow Y$ *and* $Y \rightarrow X$

- Prices, quantities
- Earnings, hours worked

Case 3: $Z \to X$ and $Z \to Y$

- Earnings of wife, earnings of husband
- Education, earnings
- GDP growth, unemployment

Outline

Normal Random Variables

Normal random variables play a big role in econometrics–partly because they are tractable, partly because normality is a recurring phenomenon in estimation theory.

The distribution of a normal random variable depends only on its mean and variance, i.e. its first two moments. To say it again, if it's normal, and you know its mean, and you know its variance, you know everything about it. You know its *exact* distribution.

- The sum of normal random variables is normal.
- Many distributions (we encounter) are *approximately* normal.

The foregoing facts and what we already know about mean and variances are important for all that we do.

An Example of Calculations with Normal Random Variables

Suppose

$$Y = a_1 X_1 + a_2 X_2 + c,$$

where $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, and the covariance between X_1 and X_2 is σ_{12}.

What is the distribution of Y?

1. Y is the sum of normal random variables so if we know its mean and variance, then we know its distribution.
2. Using the rules for expectations to calculate the mean:

$$\begin{aligned} E(Y) &= E(a_1X_1 + a_2X_2 + c) \\ &= E(a_1X_1) + E(a_2X_2 + c) \\ &= a_1E(X_1) + a_2E(X_2) + c \\ &= a_1\mu_1 + a_2\mu_2 + c \end{aligned}$$

3. Using the rules for variances:

$$\begin{aligned} Var(Y) &= Var(a_1X_1 + a_2X_2 + c) \\ &= Var(a_1X_1 + a_2X_2) \\ &= a_1^2Var(X_1) + a_2^2Var(X_2) + a_1a_2 2Cov(X_1, X_2) \end{aligned}$$

Therefore we know that:

$$Y \sim N(\mu_Y, \sigma_Y^2)$$

where

$$\begin{aligned}
\mu_Y &= a_1\mu_1 + a_2\mu_2 + c \\
\sigma_Y^2 &= a_1^2 Var(X_1) + a_2^2 Var(X_2) + a_1 a_2 2Cov(X_1, X_2) \\
&= a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + 2a_1 a_2\sigma_{12}
\end{aligned}$$

Outline

Conditional Expectations

Often the goal of empirical research in economics is to uncover conditional expectations.

Formally, I could derive the conditional probability density function and derive conditional expectation from that. If you are interested you can find this in Appendix B of Wooldridge.

Instead I want to think of a conditional expectation in a looser and informal way

The question we often care about is if I could gather everyone in the world for whom X is some particular value, what would be the expected value of Y

We define conditional expectation

$$E(Y \mid X)$$

to mean: if I "condition" X to be some value, what is the expected value of Y?"

In almost all interesting cases

- Y is a random variable so after choosing X we don't know exactly what Y will be
- $E(Y \mid X)$ depends on X, so changing X will change the expected value of Y

Very often in Economics we care about conditional expectations.

Examples

Lets consider some examples.

Note that all of this fits in the “descriptive” type of analysis we consider

We are not saying anything about causation

Wages and Gender

Let X be the gender of an individual, we may be very interested in how wages vary with gender

That is we are interested in

$$E(wage \mid male)$$
$$E(wage \mid female)$$

How do we estimate this?

Here it is pretty clear. To estimate an expectation we use the mean, to estimate the conditional expectation we use the conditional mean

That is, just take the mean value of wages for men and the mean value of wages for women

I do this using the file CPS78_85 from the textbook website

I only look at the year 1985

In this year the values were:

Average Wages 1985	
Men	Women
$9.99	$7.88

This is the data

Smoking and Birthweight

How does the birthweight of a newborn depend on smoking of the mother?

Here I use the data set bwght to look at the relationship

I just look at whether you smoked at all

Birth weight is measured in ounces

Birthweights	
Smoked	Didn't Smoke
111.15	120.06

Baseball Position and Salaries

How do the salaries of baseball players vary with the position that they play?

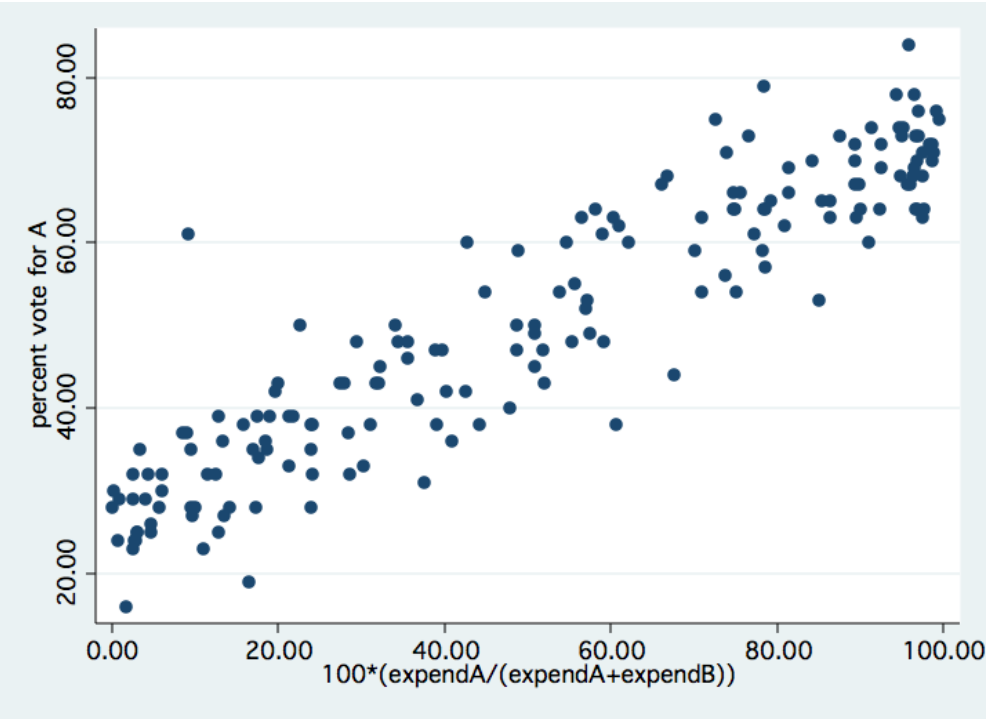
Baseball Salaries

Position	Salary
First Base	$1,586,781
Second Base	$1,309,641
Third Base	$1,382,647
Shortstop	$1,069,21
Catcher	$892,519
Outfielder	$1,539,324

Voting Outcomes and Campaign Expenditures

How does the fraction of votes you get depend on campaign expenditure?

The raw data looks like this



How do I estimate this?

Clearly I can't just condition on all levels of expenditure and take the mean

We need a model to help us think about this