EEP/IAS 118 - Introductory Applied Econometrics, Lecture 2

Gregory Lane

June 2017

Today

- Functional Form Reveiw
- Random Variable Review
  - Distribution of Random Variables (PDF, CDF)
  - Two Random Variables
- Assignments:
  - Problem Set 1 Posted

Math Review: Functional Forms

Yesterday, we discussed how econometrics is fundamentally about estimating relationships between variables

- This course will focus on the linear regression model
- However, linear regression model can handle exponentials, squares, logarithmics, etc.
    - Linear regression only ensures the model is linear in the parameters β_j
- We will go through some of the common relationships between x and y
    - What these relationships look like in data
    - How to interpret β

Proportion, Percentages, Elasticity

First, review of some concepts:

- *Proportional change*: $\frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$
- *Percentage change*: $\frac{x_1 - x_0}{x_0} \times 100 = \frac{\Delta x}{x_0} \times 100$
- *Elasticity*: $\frac{\Delta z / z}{\Delta x / x} = \frac{\partial z}{\partial x} \frac{x}{z}$

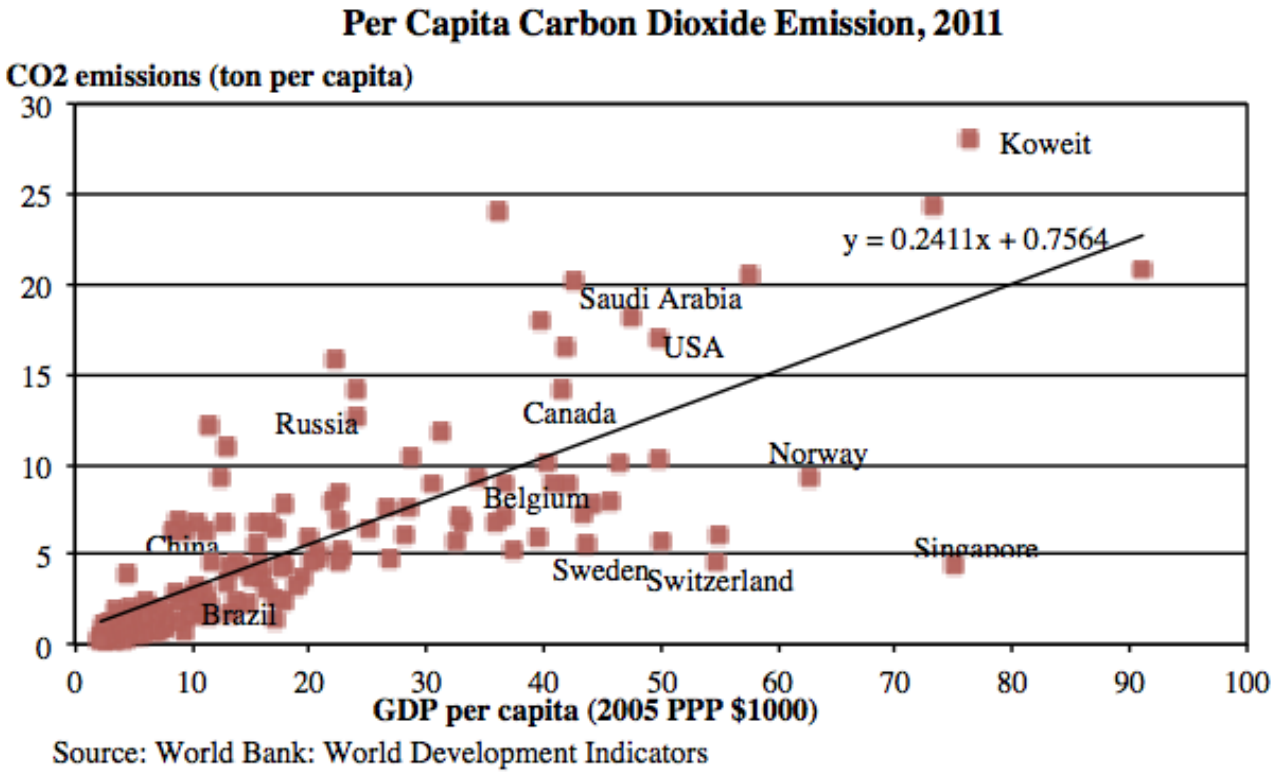
Note, percent change is just proportional change time 100.

Elasticity (η) is the "percent change in one variable in response to a given (one) percent change in another variable"

- If $0 < \eta < 1$ then the elasticity is inelastic. If $\eta > 1$ then it is elastic.

Linear Relationships

From last class, we go back to relate CO2 and GDP:



We've drawn the line:

$$y = \beta_0 + \beta_1 x \quad \Rightarrow CO2/cap = 0.75 + 0.24\ GDP/cap$$

Linear Relationships

$$CO2/cap = 0.75 + 0.24\ GDP/cap$$

How do we interpret this:

- β_1 is the slope parameter and reflects the *marginal effect* of GDP/cap on CO2/cap
  - Increase x (GDP/cap) by one unit, then the y (CO2/cap) will increase by 0.24 units
  - This is the main thing we care about: *if x changes, how does y change*. Useful to think about this as the partial derivitive
- **Units:** In order to interpret β_1 correctly, we need to pay attention to units!
  - x is measured in \$1000, while y is measured in tons. Therefore $\triangle x = 1(\$1000) \rightarrow \triangle y = 0.24$ tons
  - I.e. $\beta_1 = \frac{\triangle y}{\triangle x}$ measured in tons/\$1000

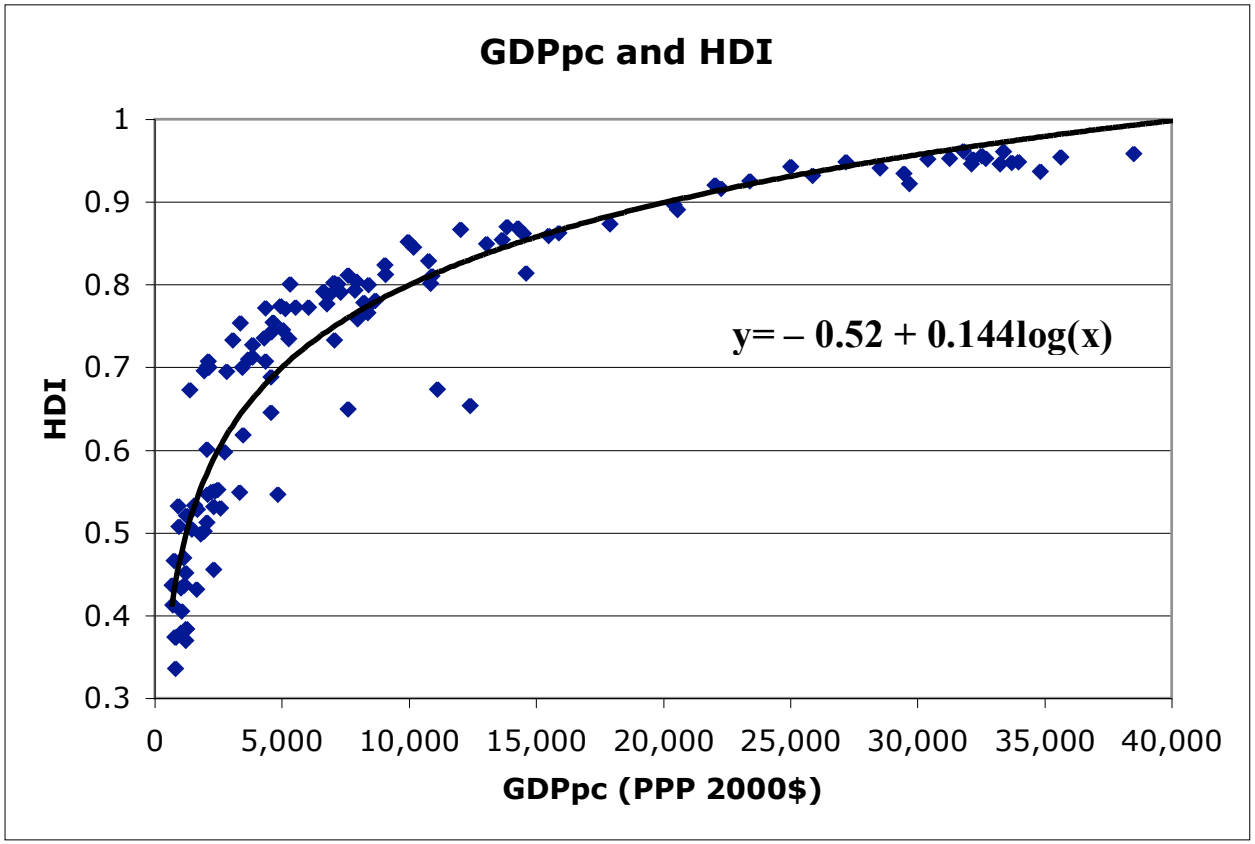
Linear Relationships

$$CO2/cap = 0.75 + 0.24\ GDP/cap$$

β_0 is the intercept parameter

- Reflects level of emissions for a country with GDP of zero
- This doesn't mean much, no country has an actual GDP of zero
- In many (most) cases β_0 does not have a meaningful interpretation
- β_0 has the units of the y variable (tons)

Logarithmic Relationships



$$y = \beta_0 + \beta_1 log(x) \rightarrow HDI = -0.52 + 0.14 log(GDPpc)$$

Good model for relationships that experience rapid growth early, but there is saturation at a certain level

Logarithmic Relationships

How do we think about the marginal effect (remember $\frac{\partial ln(x)}{\partial x} = \frac{1}{x}$):

$$\Delta HDI = 0.14 \Delta log(GDP/c) \approx 0.14 \cdot \frac{\Delta GDP/cap}{GDP/cap}$$

Now suppose the following

$$\begin{aligned} & \Delta logGDP/cap = 0.10 \\ \Rightarrow & \frac{\Delta GDP/cap}{GDP/cap} = 0.10 \\ \Rightarrow & \Delta HDI = 0.14 * 0.10 = 0.014 \end{aligned}$$

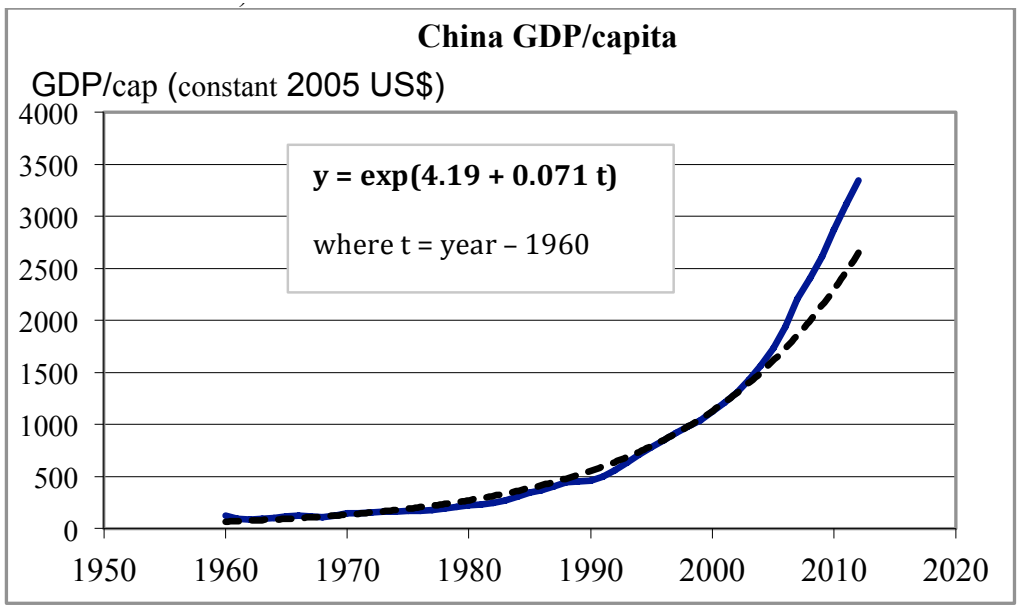
So, here we would say that a 10% increase in GDPpc leads to an increase of 0.014 HDI points

Logarithmic Relationships

General rule for Lin-log models:

- Interpret β as a 1% increase in x will lead to a $\beta/100$ unit increase in y
- A useful way to remember this is that in general when you see "log" you should think *percent*

Exponential Relationships



$$y = e^{\beta_0 + \beta_1 x}$$

$$\log y = \beta_0 + \beta_1 x$$

$$\log(GDPpc) = 4.19 + 0.07t$$

Exponential Relationships

How do we think about marginal effects:

$$\frac{\Delta y}{y} = \beta_1 \Delta x$$

In our example:

$$log(GDP/cap) = 4.19 + 0.071t$$
$$\Delta t = 1 \rightarrow \Delta \log GDP/c = 0.07$$

If t changes by 1 units, then $\log GDP/cap$ changes by 7%

Exponential Relationships

General rule for log-lin models:

- Interpret β as a 1 unit increase in x will lead to a $\beta * 100$ *percent* increase in y
- As before, for the variable that is in "log" you should think *percent*

Log-log Relationships

What if we have *both* variables in log form? For example:

$$log(y) = \beta_0 + \beta_1 log(x)$$
$$log(food) = \beta_0 + \beta_1 log(income)$$

Let's look at the marginal effect:

$$\frac{\Delta Food}{Food} = \beta_1 \frac{\Delta income}{income}$$

Question: How do we interpret this?

Log-log Relationships

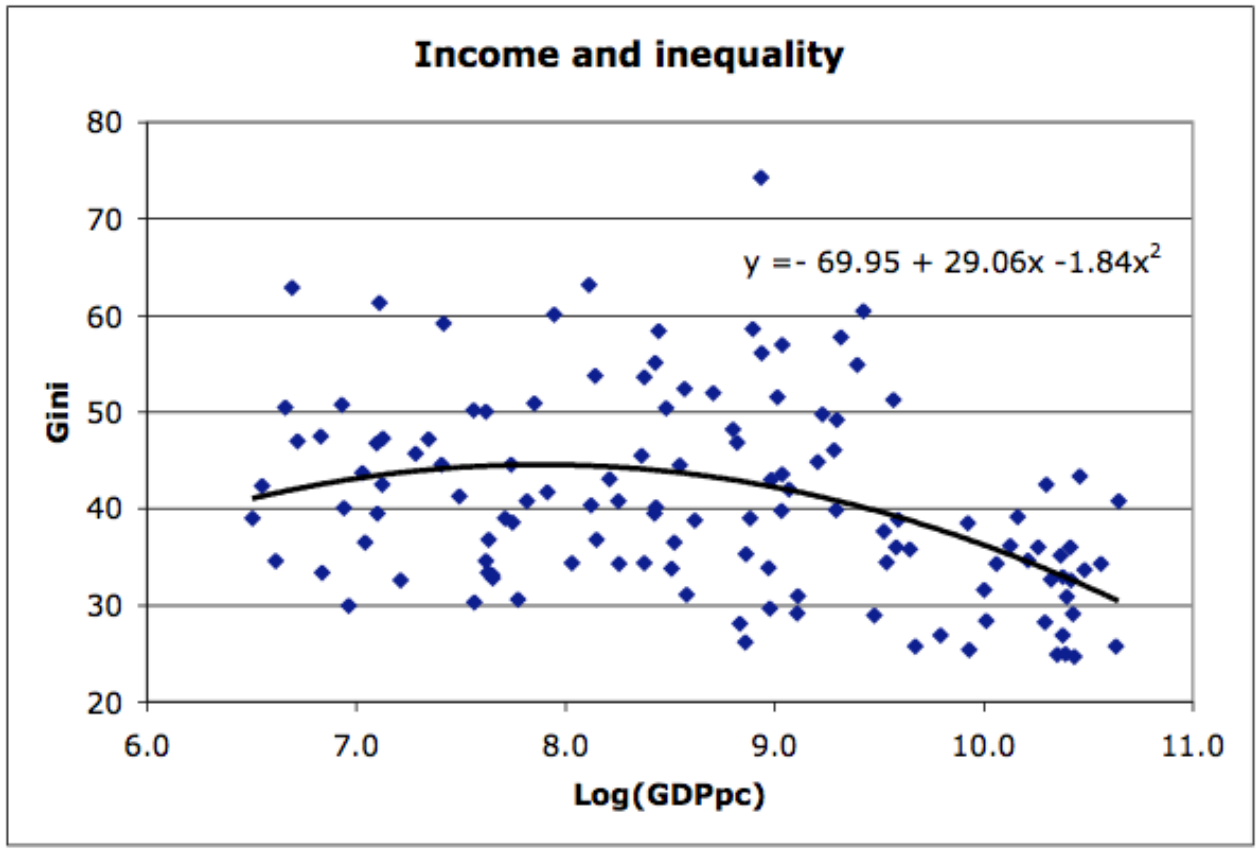
$$\frac{\Delta Food}{Food} = \beta_1 \frac{\Delta Income}{Income}$$

Question: How do we interpret this?

$$\frac{\frac{\Delta Food}{Food}}{\frac{\Delta Income}{Income}} = \beta_1$$

- So, β_1 is the income elasticity of food consumption
- That is, a 1% increase in income will lead to a β_1 percent increase in food conusmption
- *Both* variables are in logs, so we think about *both* variables in terms of percent change

Quadratic Relationships



We've drawn the line:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \quad \Rightarrow Gini = -70 + 29x - 1.84x^2$$

Quadratic Relationships

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$GINI = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$GINI = -70 + 29x - 1.84x^2$$

- Want to know the shape (concave up or concave down) : β_2 negative concave down, and β_2 positive concave up
- Need to retrieve the marginal effect

$$\begin{aligned} \Delta y &= \beta_1 \Delta x + 2\beta_2 x \Delta x \\ &= (\beta_1 + 2\beta_2 x)(\Delta x) \end{aligned}$$

Note that the marginal effect *changes* depending on the starting value of x

Quadratic Relationships

Next we may also want to find the turning point:

$$\beta_1 + 2\beta_2 x = 0$$

$$x = -\frac{\beta_1}{2\beta_2} = \frac{29.06}{2(-1.84)} \approx 8$$

Interpreting a quadratic is a little tricky - in order to state the marginal effect of increasing x on y we need to choose a starting value of x. Typically we choose the mean value of x

Functional forms and Marginal Effects Overview

This Table (Table 2.3 in Wooldridge) is meant to provide a summary of the various functional forms and the associated β interpretations (found on page 5 of notes).

Model	DepVar	Ind. Var	Δy relates to Δx?	Interpretation
Linear	y	x	$\Delta y = \beta_1 \Delta x$	$\Delta y = \beta_1 \Delta x$
Logarithmic	y	$\log(x)$	$\Delta y = \beta_1 \frac{\Delta x}{x}$	$\Delta y = (\beta_1/100)\%\Delta x$
Exponential	$\log(y)$	x	$\frac{\Delta y}{y} = \beta_1 \Delta x$	$\%\Delta y = (100\beta_1)\Delta x$
Log-Log	$\log(y)$	$\log(x)$	$\frac{\Delta y}{y} = \beta_1 \frac{\Delta x}{x}$	$\%\Delta y = \beta_1 \%\Delta x$

Ex: $\%\Delta y = (100\beta_1)\Delta x$ - Read this as "$\hat{y}$ increases by $100 * \beta_1\%$ for a one unit increase in x."

Example Interpreting Marginal Effects

Suppose you've collected data on household gasoline consumption (gallons) in the Bay Area and gas prices ($ per gallon), and you estimate the following model:

$$\log(gasoline) = 12 - 0.21 price$$

According to the model, how does gas consumption change when $price$ increases by $1?

Example Interpreting Marginal Effects

Suppose you've collected data on household gasoline consumption (gallons) in the Bay Area and gas prices ($ per gallon), and you estimate the following model:

$$\log(gasoline) = 12 - 0.21 price$$

According to the model, how does gas consumption change when $price$ increases by $1?

If $price$ increases by $1, then predicted $gasoline$ consumption will decrease by 21%

Example Interpreting Marginal Effects

In a strange (but real) example, a researcher used scanner data from a national grocery store to investigate how chicken consumption was affected by gas prices. Specifically, she looked at the share of chicken purchases that were made while the chicken was on sale. The following model was estimated:

$$\log(chickenshare) = 0.83 + 0.491 \log(gasprice)$$

How does *chickenshare* change if gas prices rise by 2%?

Example Interpreting Marginal Effects

In a strange (but real) example, a researcher used scanner data from a national grocery store to investigate how chicken consumption was affected by gas prices. Specifically, she looked at the share of chicken purchases that were made while the chicken was on sale. The following model was estimated:

$$\log(chickenshare) = 0.83 + 0.491 \log(gasprice)$$

How does *chickenshare* change if gas prices rise by 2%?

This is a log-log model, so if the price of gas increases by 2%, then the predicted share of chicken sold on sale increases by 0.98%.

$$\%\Delta y = 0.491 * 2\% = 0.98\%$$

Example Interpreting Marginal Effects

Suppose you've collected data on CEO salaries (hundred thousand \$) and annual firm sales (million \$), and you estimate the following model:

$$salary = 2.23 + 1.1 \log(sales)$$

According to the model, how does $salary$ change if annual firm sales increase by 10%?

Example Interpreting Marginal Effects

$$salary = 2.23 + 1.1 \log(sales)$$

According to the model, how does $salary$ change if annual firm sales increase by 10%?

Sol. If annual firm sales increase by 10%, the model predicts that CEO salary increases by \$11,000.

If annual firm sales increase by 10%, then we know $\%\Delta x = 10$.

$$\Delta y = (\beta_1/100)\%\Delta x$$

We can plug this and our estimate of β_1 into the formula from the table to see that $\Delta y = \frac{1.1}{100} * 10 = 0.11$. Since the units of CEO salaries is \$100,000, an increase of 0.11 units is an increase of \$11,000.

Populations and Samples

An important distinction in this class will be between the *population* and a *sample*

- We distinguish between the population (the universe of adults) and a sample (the subset of this population that you observe in your data)
- The population at large has a distribution for each variable of interest: educ, income, number of cars,...
- A random draws of an single observation gives us a random variable from this true distribution found in the population. In other words a random variable is a number that is taken from some set of possible outcomes. It is the value of a characteristic (age, education) of an observation (firm, household, city) drawn randomly

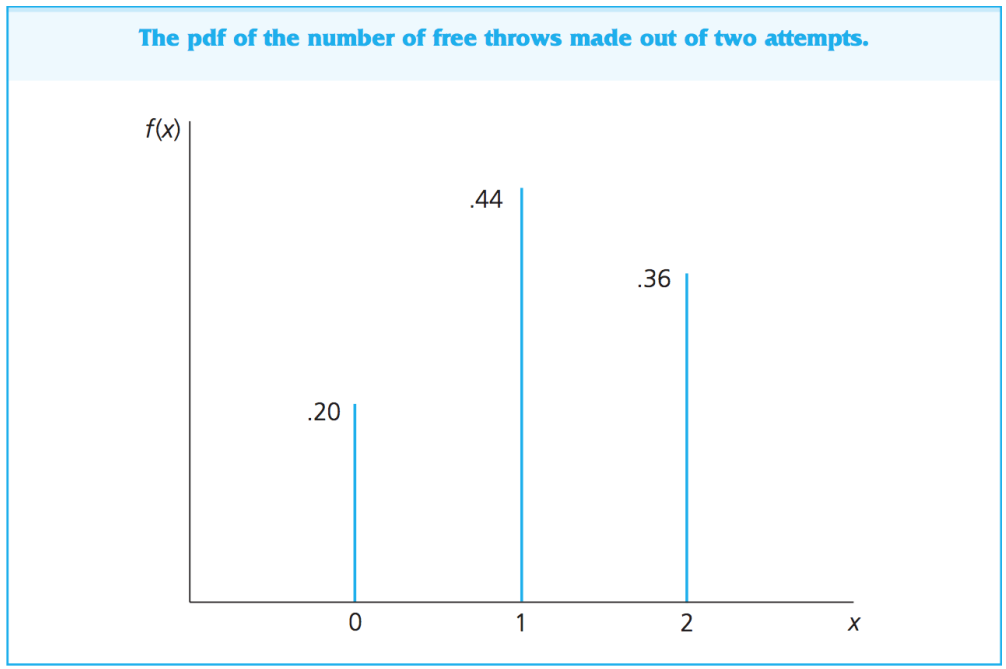
Stat Review: Random Variables

Random variables are numbers that are taken from a distribution of possible outcomes. A fundamental way to describe a random variable is through its probability distribution function.

Discrete random variable pdf:

$$f(x_j) = P(X = x_j), \;\; j = \{1, 2, 3, 4, 5, ...k\}$$

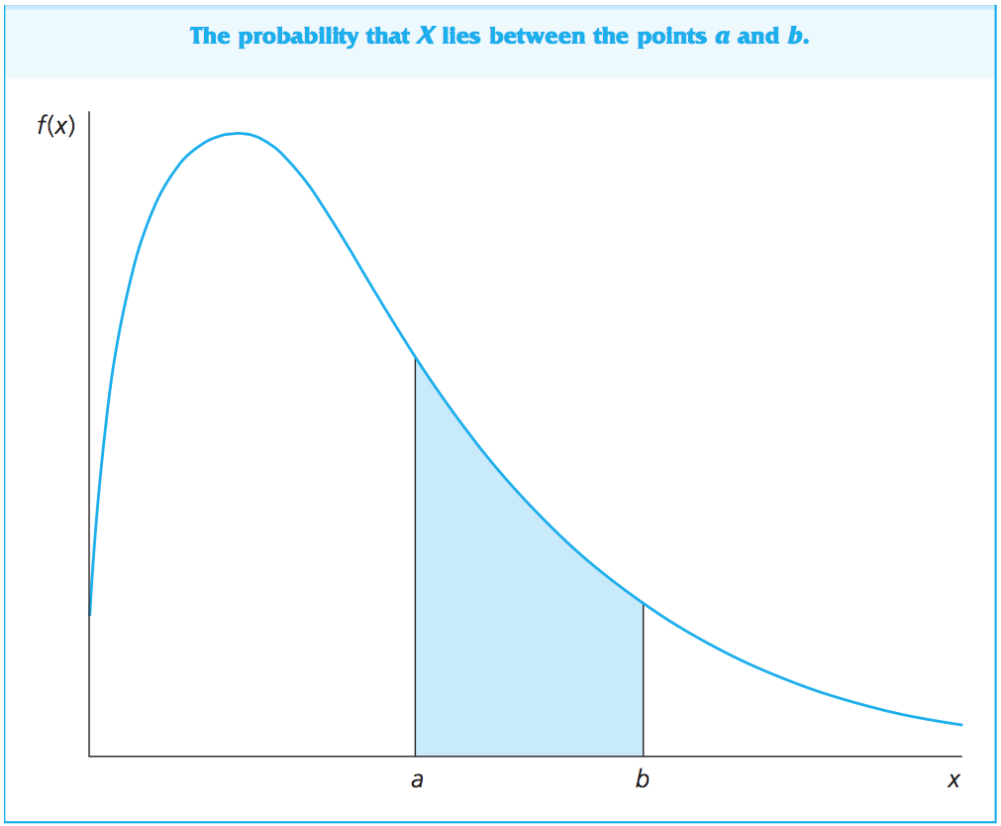
$$f(0) = 0.20 \; ; f(1) = 0.44 \; ; f(2) = 0.36$$



Stat Review: Random Variables

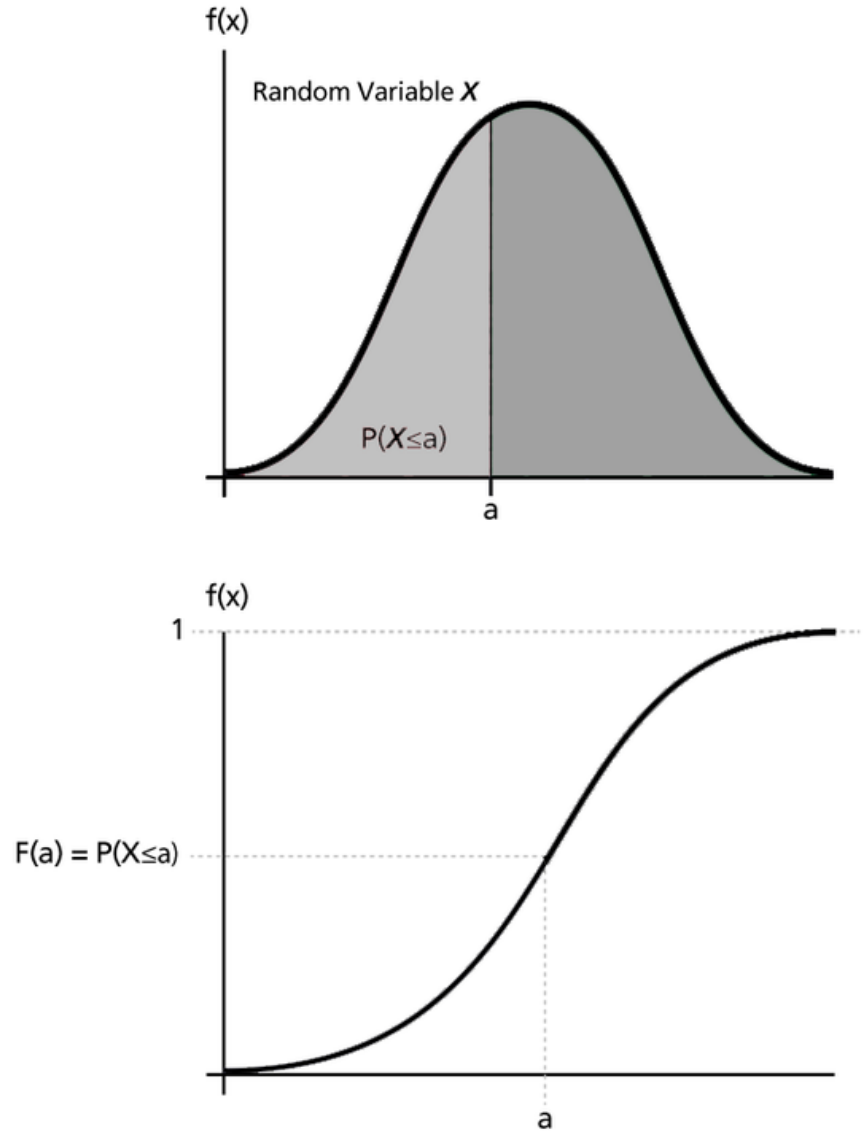
Continuous variable (pdf):

$$Pr(a < X < b) = \int_a^b xf(x)dx$$



Stat Review: Random Variables

The cumulative distribution function is another useful way to visualize a random variable:



Stat Review: Two Random Variables

- If we have two discrete random variables X and Y, we can define the **joint probability density function** of (X,Y):

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

- Two variables are **independent** if the joint PDF is equal to the product of the individual variables' pdf.

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

- The conditional distribution of Y given X, which is described by the **conditional probability density function** :

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

Stat Review: Two Random Variables

Let's do an example using survey data:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is joint probability that a random person is a head of household and did NOT complete primary school?

Stat Review: Two Random Variables

Let's do an example using survey data:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is joint probability that a random person is a head of household and did NOT complete primary school?

$$f(Incomplete, yes) = 0.05$$

Stat Review: Two Random Variables

Let's do an example using survey data:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is the conditional probability that a randomly drawn head of household did NOT complete primary school?

Stat Review: Two Random Variables

Let's do an example using survey data:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is the conditional probability that a randomly drawn head of household did NOT complete primary school?

$$f(Incomplete|yes) = 30/197 = 0.15$$

Features of Probability Distributions

- **The expected value of X:**

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k) = \sum_{j=1}^{k} x_j f(x_j)$$

If X is continuous

$$E(X) = \int_{-\infty}^{+\infty} x f(x) d(x)$$

- **The variance of X:**

$$Var(X) = E[(X - E(X))^2]$$

- **The standard deviation of X**

$$sd(X) = \sqrt{Var(X)}$$

Sample Properties

We can never know the real pdf or cdf of the population at large, instead we can only infer things about the population based on the samples we do observe

We can calculate the statistical properties of these samples:

- **Sample Mean:**

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

- **Sample Variance:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

These two terms are sample **estimators** for the true population sample. What an estimator is and its properties will be a key concept in this class!

Sample Properties

The goal in calculating these sample properties is that they can inform us about the analogous properties found in the population

- The sample mean is informative about the population mean
- The sample variance is information about the population variance
- The sample correlation between two variables is informative about the population correlation

The problem is that the value of the sample statistic *will not* be equal to it's analogue in the population.

- How do we deal with this problem?

Sample Properties: Law of Large Numbers

- In small samples, the sample mean can be quite different from the true population mean
    - For example, if I roll a five and a six on a die the sample mean will be $\frac{1}{2}(6+5) = 5.5$, even when we know the true *population* expected value of a die roll is 3.5:

$$E(X) = 1\left(\tfrac{1}{6}\right) + 2\left(\tfrac{1}{6}\right) + 3\left(\tfrac{1}{6}\right) + 4\left(\tfrac{1}{6}\right) + 5\left(\tfrac{1}{6}\right) + 6\left(\tfrac{1}{6}\right) = 3.5$$

- Usefully, the **law of large numbers** says that if we draw a sample consisting of *n* realizations of our random variable, and take the average, this sample mean will approach the population mean as *n* approaches infinity.
    - This means that if I roll a die more and more, my sample mean will approach the true population mean of 3.5