Statistics Review

October 14, 2019

Random Variables

- An **experiment** is any procedure that can (at least in theory) be infinitely repeated and has a well-defined set of outcomes

Random Variables

- An **experiment** is any procedure that can (at least in theory) be infinitely repeated and has a well-defined set of outcomes
- Example: Flipping a coin two times
  - Well defined set of outcomes: 2 heads, 1 heads, 0 heads
  - Can be done over and over again

Random Variables

- An **experiment** is any procedure that can (at least in theory) be infinitely repeated and has a well-defined set of outcomes
- Example: Flipping a coin two times
  - Well defined set of outcomes: 2 heads, 1 heads, 0 heads
  - Can be done over and over again
- A **random variable** is a variable that takes on numerical values and has an outcome that is determined by an experiment

Random Variables

- An **experiment** is any procedure that can (at least in theory) be infinitely repeated and has a well-defined set of outcomes
- Example: Flipping a coin two times
  - Well defined set of outcomes: 2 heads, 1 heads, 0 heads
  - Can be done over and over again
- A **random variable** is a variable that takes on numerical values and has an outcome that is determined by an experiment
- Example: The number of heads appearing in two coin flips
  - Before you flip the coins, you do not know how many heads will appear

Random Variables

- An **experiment** is any procedure that can (at least in theory) be infinitely repeated and has a well-defined set of outcomes
- Example: Flipping a coin two times
  - Well defined set of outcomes: 2 heads, 1 heads, 0 heads
  - Can be done over and over again
- A **random variable** is a variable that takes on numerical values and has an outcome that is determined by an experiment
- Example: The number of heads appearing in two coin flips
  - Before you flip the coins, you do not know how many heads will appear
  - Once we flip the coins, we obtain the number of heads as an outcome of the experiment

Examples

- **Example 1:** Airline decides how many reservations to accept for a flight with 100 seats
  - If more than 100 want reservations, could be safe and accept 100
  - But some people might now show up
  - X=Number of people showing up for the flight

Examples

- **Example 1:** Airline decides how many reservations to accept for a flight with 100 seats
  - If more than 100 want reservations, could be safe and accept 100
  - But some people might now show up
  - X=Number of people showing up for the flight
- **Example 2:** Experiment is tossing a coin once
  - $X = \begin{cases} 1 & \text{if outcome is heads} \\ 0 & \text{if outcome is tails} \end{cases}$

Discrete random variables

- A discrete random variable is a variable taking on a finite number of values
- So far, our examples of random variables were discrete

Discrete random variables

- A discrete random variable is a variable taking on a finite number of values
- So far, our examples of random variables were discrete
- A **Bernoulli random variable** takes on only two values (0 or 1)

Discrete random variables

- A discrete random variable is a variable taking on a finite number of values
- So far, our examples of random variables were discrete
- A **Bernoulli random variable** takes on only two values (0 or 1)
- 1=some event happens
- 0= some event does not happen

Discrete random variables

- A discrete random variable is a variable taking on a finite number of values
- So far, our examples of random variables were discrete
- A **Bernoulli random variable** takes on only two values (0 or 1)
- 1=some event happens
- 0= some event does not happen
- All we need to completely describe a Bernoulli random variable is the probability of the event happening

Example

- Consider again the problem of studying how many customers show up to a flight with 100 seats

Example

- Consider again the problem of studying how many customers show up to a flight with 100 seats
- We can analyze the problem be defining a Bernoulli random variable for each customer as follows:

$$X = \begin{cases} 1 & \text{if customer shows up} \\ 0 & \text{if customer does not show up} \end{cases}$$

Example

- Consider again the problem of studying how many customers show up to a flight with 100 seats
- We can analyze the problem be defining a Bernoulli random variable for each customer as follows:

$$X = \begin{cases} 1 & \text{if customer shows up} \\ 0 & \text{if customer does not show up} \end{cases}$$

- To solve the airline's problem, we need to assign probabilities to these events :

$$P(X = 1) = p, \qquad P(X = 0) = 1 - p, \qquad p \in (0, 1)$$

Probability density functions

- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J.

Probability density functions

- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J.
- For each possible value x_j, the probability of the value occurring is $p_j = P(X = x_j)$.
- Of course, $0 \leq p_j \leq 1$ for all j and $p_1 + p_2 + ... + p_J = 1$.

Probability density functions

- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J.
- For each possible value x_j, the probability of the value occurring is $p_j = P(X = x_j)$.
- Of course, $0 \leq p_j \leq 1$ for all j and $p_1 + p_2 + ... + p_J = 1$.
- The **probability density function (pdf)** assigns to each possible value of x_j its probability p_j of occurring:

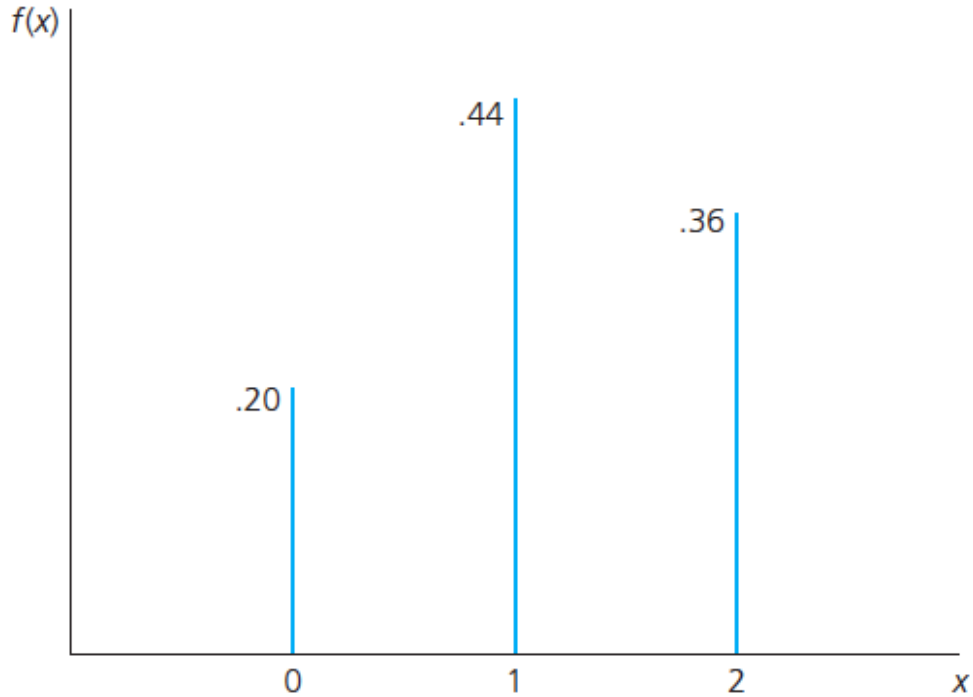
$$f(x_j) = p_j \quad j = 1, 2, 3, ..., J$$

Example

- X=number of free throws made by a basketball player out of two attempts

Example

- X=number of free throws made by a basketball player out of two attempts
- **pdf:** $f(0) = 0.20$, $f(1) = 0.44$, and $f(2) = 0.36$



Expected value

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring

Expected value

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring

- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J

Expected value

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring
- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J
- $f(x_j) = p_j \quad j = 1, 2, 3, ..., J$

Expected value

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring
- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J
- $f(x_j) = p_j \quad j = 1, 2, 3, ..., J$
- Then,

$$E(X) = \mu = x_1 f(x_1) + x_2 f(x_2) + x_3 f(x_3) + ... + x_J f(x_J)$$
$$= \sum_{i=1}^{J} x_j f(x_j)$$

Expected value

- **Example:** Suppose X takes on the values -1, 0, and 2 with probabilities 1/8, 1/2, and 3/8. What is the expected value of X?

Expected value

- **Example:** Suppose X takes on the values -1, 0, and 2 with probabilities 1/8, 1/2, and 3/8. What is the expected value of X?

- **Example 2:** Suppose X takes on the values -1, 0, and 2 with probabilities 0, 1/2, and 1/2. What is the expected value of X?

Expected value

- **Example:** Suppose X takes on the values -1, 0, and 2 with probabilities 1/8, 1/2, and 3/8. What is the expected value of X?
- **Example 2:** Suppose X takes on the values -1, 0, and 2 with probabilities 0, 1/2, and 1/2. What is the expected value of X?
- Expected value tells us how large the random variable is "on average"
- If X takes on higher values with higher probabilities, it will have a larger expected value

Properties of expected values

- If c is a constant, $E(c) = c$
- For any constants a and b, $E(aX + b) = aE(x) + b$

Properties of expected values

- If c is a constant, $E(c) = c$
- For any constants a and b, $E(aX + b) = aE(x) + b$

Properties of expected values

- If c is a constant, $E(c) = c$
- For any constants a and b, $E(aX + b) = aE(x) + b$
- In econometrics, we will often consider several random variables at ones. If $X_1, X_2, ..., X_N$ are random variables, then

$$\begin{aligned} E\left(\sum_{i=1}^{L} X_n\right) &= E(X_1 + X_2 + ... + X_N) = \\ &= E(X_1) + E(X_2) + ... + E(X_N) \\ &= \sum_{i=1}^{N} E(X_i) \end{aligned}$$

Variance

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring

- The **variance** of a random variable X is the expected value of squared deviations from $E(X)$

Variance

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring
- The **variance** of a random variable X is the expected value of squared deviations from $E(X)$
- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J

Variance

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring
- The **variance** of a random variable X is the expected value of squared deviations from $E(X)$
- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J
- $f(x_j) = p_j \quad j = 1, 2, 3, ..., J$

Variance

- The **expected value** of a random variable X is the sum of all of its possible values weighted by the probability of each value occurring

- The **variance** of a random variable X is the expected value of squared deviations from $E(X)$

- Consider a discrete random variable X that can take on the values x_1, x_2, x_3, ..., x_J

- $f(x_j) = p_j \quad j = 1, 2, 3, ..., J$

- Then,

$$V(X) = \sigma^2 = (x_1 - E(X))^2 f(x_1) + (x_2 - E(X))^2 f(x_2) + ...$$
$$= \sum_{i=1}^{J} (x_j - E(X))^2$$

Variance

- **Example:** Suppose X takes on the values -1, 0, and 1 with equal probabilities. Compute $E(X)$ and $V(X)$

Variance

- **Example:** Suppose X takes on the values -1, 0, and 1 with equal probabilities. Compute $E(X)$ and $V(X)$

- **Example:** Suppose X takes on the values -2,-1, 0, 1, and 2 with equal probabilities. Compute $E(X)$ and $V(X)$

Variance

- **Example:** Suppose X takes on the values -1, 0, and 1 with equal probabilities. Compute $E(X)$ and $V(X)$
- **Example:** Suppose X takes on the values -2,-1, 0, 1, and 2 with equal probabilities. Compute $E(X)$ and $V(X)$
- Variance tells us how much the random variable is spread out
- If X is more spread out, it will have a larger variance

Variance

- **Example:** Suppose X takes on the values -1, 0, and 1 with equal probabilities. Compute $E(X)$ and $V(X)$
- **Example:** Suppose X takes on the values -2,-1, 0, 1, and 2 with equal probabilities. Compute $E(X)$ and $V(X)$
- Variance tells us how much the random variable is spread out
- If X is more spread out, it will have a larger variance
- The **standard deviation** of a random variable is the square root of the variance:

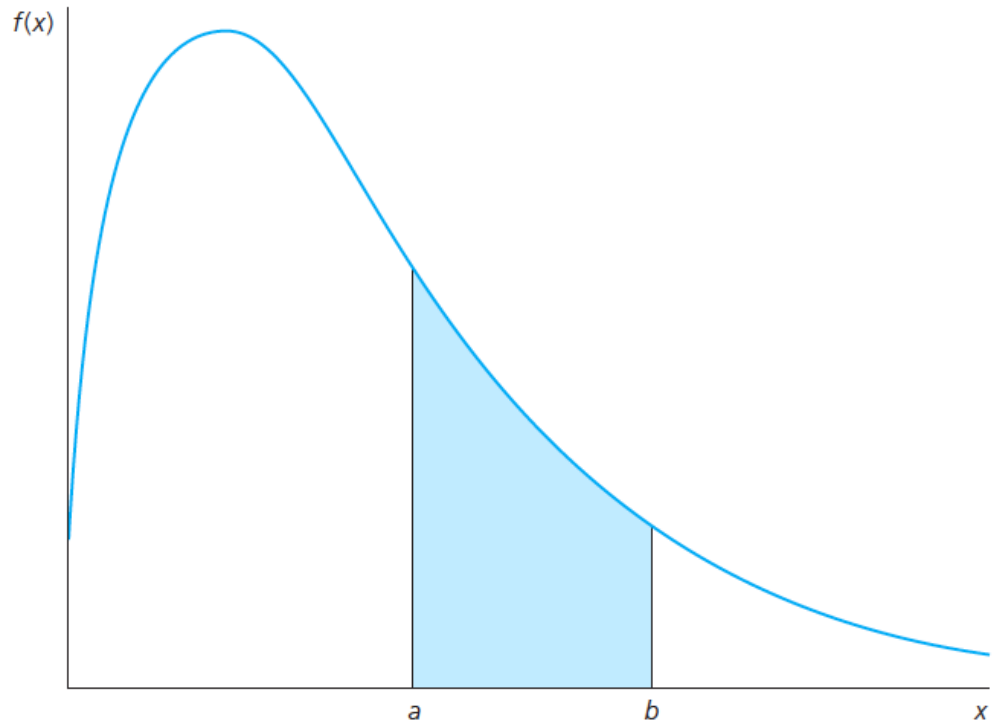
$$\sigma = \sqrt{\sum_{i=1}^{J}(x_j - E(X))^2}$$

Continuous random variables

- So far, we have been considering random variables that take on a finite number of values
- A random variable can also take on an infinite number of values
  - E.g., normal, uniform
- For continuous random variables, the probability of any particular value is **zero**
- But we can still talk about pdfs!

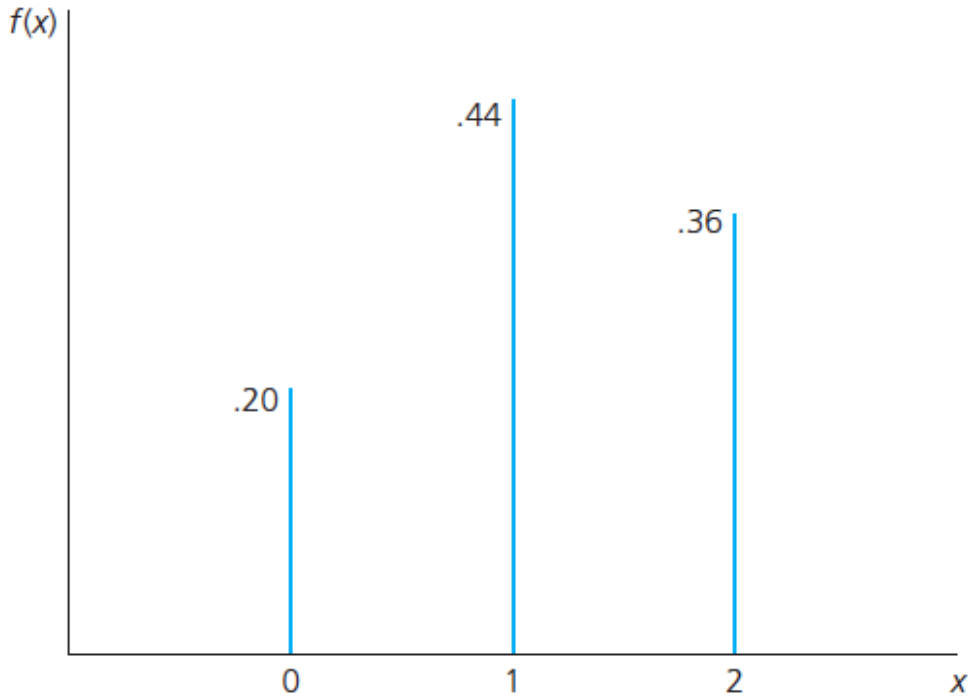
Continuous random variables

- A **pdf** of a continuous random variable is a function $f(x)$ such that the probability that x **falls in some range** is the area under the curve:



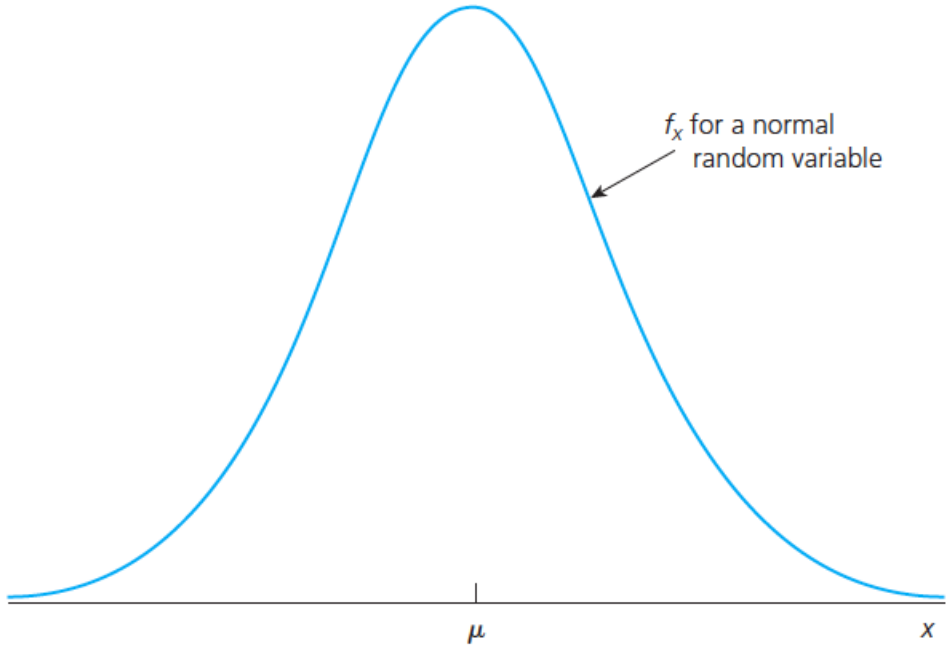
Comparison to discrete pdfs

- Notice that for a discrete random variable we can think think about ranges too:



- High values of the pdf mean that the associated values are more likely
- Same for continuous random variables

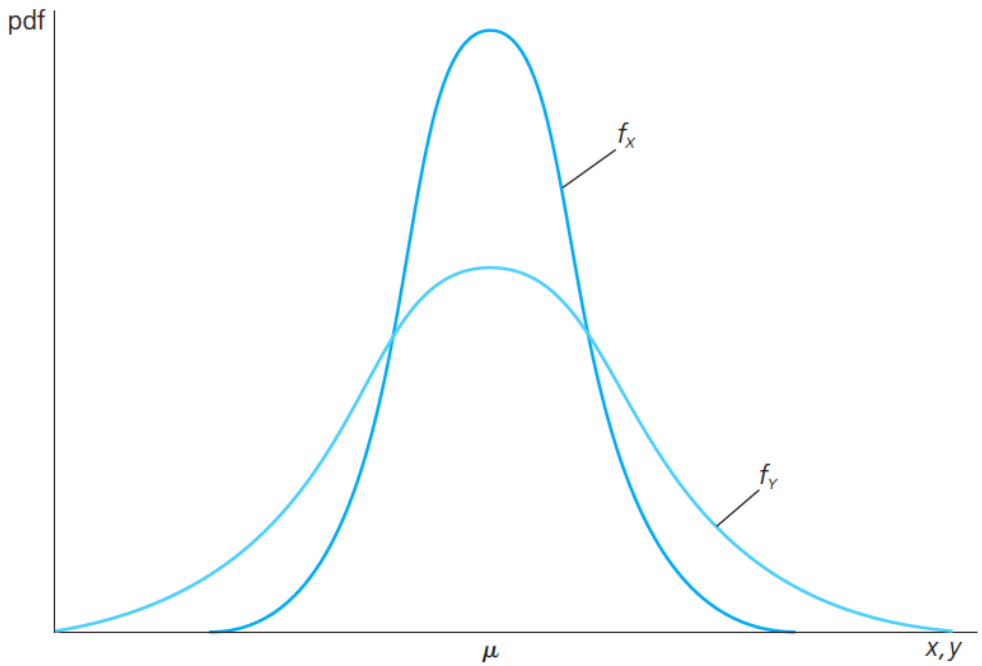
Pdf of a normal distribution



- One of the most commonly used distributions in econometrics
- We will rely on it heavily later in the course

Mean and variance of continuous random variables

- The definition of the expected value of a continuous random variable requires integrals, so we will skip it here
- Intuitively, it represents the same thing: how large the random variable is "on average"
- Variance of a continuous random variable has the same definition: $E((X - E(X))^2)$
- ...and same interpretation: how much the variable is spread out:



Conditional distributions

- Easy to think about for discrete random variables:

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

Conditional distributions

- Easy to think about for discrete random variables:

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

- Basketball example: Y=second free throw, X=first free throw
- 1=success, 0=failure

Conditional distributions

- Easy to think about for discrete random variables:

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

- Basketball example: Y=second free throw, X=first free throw
- 1=success, 0=failure

$$f_{Y|X}(1|1) = .85, \quad f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(1|0) = .70, \quad f_{Y|X}(0|0) = 0.30$$

Conditional distributions

- Easy to think about for discrete random variables:

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

- Basketball example: Y=second free throw, X=first free throw
- 1=success, 0=failure

$$f_{Y|X}(1|1) = .85, \quad f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(1|0) = .70, \quad f_{Y|X}(0|0) = 0.30$$

$\Rightarrow$ Distribution of Y depends on distribution of X

Conditional distributions

- Easy to think about for discrete random variables:

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

- Basketball example: Y=second free throw, X=first free throw
- 1=success, 0=failure

$$f_{Y|X}(1|1) = .85, \quad f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(1|0) = .70, \quad f_{Y|X}(0|0) = 0.30$$

$\Rightarrow$ Distribution of Y depends on distribution of X

$\Rightarrow$ Y and X are not independent

Conditional distributions

- Easy to think about for discrete random variables:

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

- Basketball example: Y=second free throw, X=first free throw
- 1=success, 0=failure

$$f_{Y|X}(1|1) = .85, \quad f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(1|0) = .70, \quad f_{Y|X}(0|0) = 0.30$$

$\Rightarrow$ Distribution of Y depends on distribution of X

$\Rightarrow$ Y and X are not independent

- If X and Y are independent, then $f_{Y|X}(y|x) = f(Y)$ (the distribution of Y does not depend on X)

Conditional expectations

- We can use conditional distributions to form conditional expectations:

Conditional expectations

- We can use conditional distributions to form conditional expectations:
- $E(Y) = \sum_{i=1}^{J} y_j f(y_j)$

Conditional expectations

- We can use conditional distributions to form conditional expectations:
- $E(Y) = \sum_{i=1}^{J} y_j f(y_j)$
- $E(Y|X) = \sum_{i=1}^{J} y_j f_{Y|X}(y_j|X)$

Conditional expectations

- We can use conditional distributions to form conditional expectations:
- $E(Y) = \sum_{i=1}^{J} y_j f(y_j)$
- $E(Y|X) = \sum_{i=1}^{J} y_j f_{Y|X}(y_j|X)$
- Go back to our example:

$$f_{Y|X}(1|1) = .85, \quad f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(1|0) = .70, \quad f_{Y|X}(0|0) = 0.30$$

Conditional expectations

- We can use conditional distributions to form conditional expectations:
- $E(Y) = \sum_{i=1}^{J} y_j f(y_j)$
- $E(Y|X) = \sum_{i=1}^{J} y_j f_{Y|X}(y_j|X)$
- Go back to our example:

$$f_{Y|X}(1|1) = .85, \quad f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(1|0) = .70, \quad f_{Y|X}(0|0) = 0.30$$

- $E(Y|X = 0) = ???$, $E(Y|X = 1) = ???$

Conditional expectations

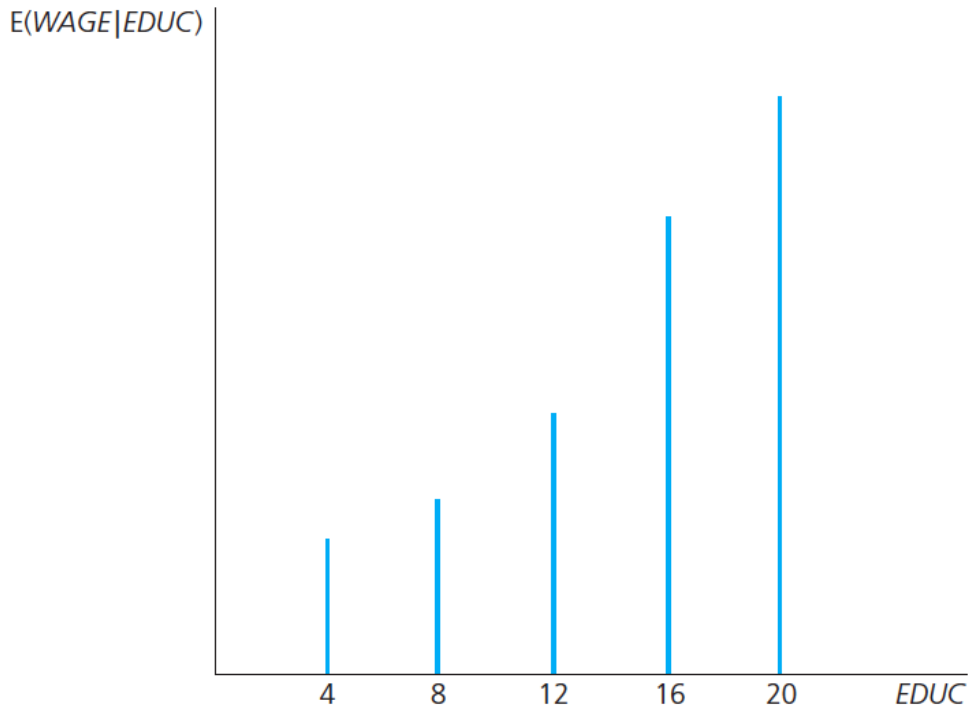
- We can use conditional distributions to form conditional expectations:
- $E(Y) = \sum_{i=1}^{J} y_j f(y_j)$
- $E(Y|X) = \sum_{i=1}^{J} y_j f_{Y|X}(y_j|X)$
- Go back to our example:

$$f_{Y|X}(1|1) = .85, \quad f_{Y|X}(0|1) = 0.15$$
$$f_{Y|X}(1|0) = .70, \quad f_{Y|X}(0|0) = 0.30$$

- $E(Y|X = 0) = ???$, $E(Y|X = 1) = ???$
- If X and Y are independent, then $E(Y|X) = E(Y)$ (the expectation of Y does not depend on X)

Conditional expectations

- Example: WAGE=wage (in dollars), EDUC=education in years
- $E(WAGE|EDUC)$:



Conditional expectations

- When working with data, we will often assume that conditional expectations are linear:

$$E(Y|X) = \beta_0 + \beta_1 X$$

- For example, WAGE=wage (in dollars), EDUC=education in years,

$$E(WAGE|EDUC) = 1.05 + .45 EDUC$$

Population

- A **population** is a random variable represented by some density $f(X;\theta)$
- θ is a vector of parameters (e.g., mean and variance)

Population

- A **population** is a random variable represented by some density $f(X;\theta)$
- θ is a vector of parameters (e.g., mean and variance)
- In our airline example, the population is the number of people showing up for the flight

Population

- A **population** is a random variable represented by some density $f(X;\theta)$
- θ is a vector of parameters (e.g., mean and variance)
- In our airline example, the population is the number of people showing up for the flight
- If we are interested in modeling wages, our population is all possible wages

Population

- A **population** is a random variable represented by some density $f(X;\theta)$
- θ is a vector of parameters (e.g., mean and variance)
- In our airline example, the population is the number of people showing up for the flight
- If we are interested in modeling wages, our population is all possible wages
- If we are interested in modeling wages of the residents of Moscow, our population is all possible wages in Moscow, etc

The point of statistics

- The point of statistics (econometrics) is to learn about the population $f(X;\theta)$
- Usually, we make some assumption about f but we never know θ

The point of statistics

- The point of statistics (econometrics) is to learn about the population $f(X;\theta)$
- Usually, we make some assumption about f but we never know θ
- Learning about the population means
  - Estimating θ (**Estimation**)
  - Testing hypotheses about it (**Hypothesis testing**)

Sample

- A **sample** consists of realizations of a random variable coming from the population

Sample

- A **sample** consists of realizations of a random variable coming from the population
- **Example:** $f(X;\theta)$ is the distribution of airline customers

Sample

- A **sample** consists of realizations of a random variable coming from the population
- **Example:** $f(X;\theta)$ is the distribution of airline customers
- You observe $X_1 = 58$ customers show up on day one and $X_2 = 110$ customers show up on day to

Sample

- A **sample** consists of realizations of a random variable coming from the population
- **Example:** $f(X;\theta)$ is the distribution of airline customers
- You observe $X_1 = 58$ customers show up on day one and $X_2 = 110$ customers show up on day to
- Because these are realizations of a random variable from $f(X;\theta)$, we say that 58 and 110 is the sample

Sample

- A **sample** consists of realizations of a random variable coming from the population
- **Example:** $f(X;\theta)$ is the distribution of airline customers
- You observe $X_1 = 58$ customers show up on day one and $X_2 = 110$ customers show up on day to
- Because these are realizations of a random variable from $f(X;\theta)$, we say that 58 and 110 is the sample
- A different sample would result in different values of X_1 and X_2

Example

- Assume we have a sample of unemployment rates:

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

- These come from some population with mean μ
- How do we estimate μ?

Sample mean

- Consider a population with mean μ
- Sample $X_1, X_2, ..., X_N$ from the population
- We estimate the population mean using the sample mean:

$$\bar{X} = \frac{1}{N} \sum_n X_n$$

Sample mean

- Consider a population with mean μ
- Sample $X_1, X_2, ..., X_N$ from the population
- We estimate the population mean using the sample mean:

$$\bar{X} = \frac{1}{N} \sum_n X_n$$

- Unbiased estimator: $E(\bar{X}) = \mu$

Sample mean

- Consider a population with mean μ
- Sample $X_1, X_2, ..., X_N$ from the population
- We estimate the population mean using the sample mean:

$$\bar{X} = \frac{1}{N} \sum_n X_n$$

- Unbiased estimator: $E(\bar{X}) = \mu$
- We usually use μ to denote the population mean and $\bar{X}$ to denote the sample mean

The general problem

- In general, there is some parameter θ to estimate and some sample of $X_1, X_2, ..., X_N$

The general problem

- In general, there is some parameter θ to estimate and some sample of $X_1, X_2, ..., X_N$
- Our estimator of θ will be a function of $X_1, X_2, ..., X_N$:

$$W = h(X_1, X_2, ..., X_N)$$

The general problem

- In general, there is some parameter θ to estimate and some sample of $X_1, X_2, ..., X_N$
- Our estimator of θ will be a function of $X_1, X_2, ..., X_N$:

$$W = h(X_1, X_2, ..., X_N)$$

- Different samples of $X_1, X_2, ..., X_N$ will lead to different values of w

The general problem

- In general, there is some parameter θ to estimate and some sample of $X_1, X_2, ..., X_N$
- Our estimator of θ will be a function of $X_1, X_2, ..., X_N$:

$$W = h(X_1, X_2, ..., X_N)$$

- Different samples of $X_1, X_2, ..., X_N$ will lead to different values of w
- W is itself a random variable

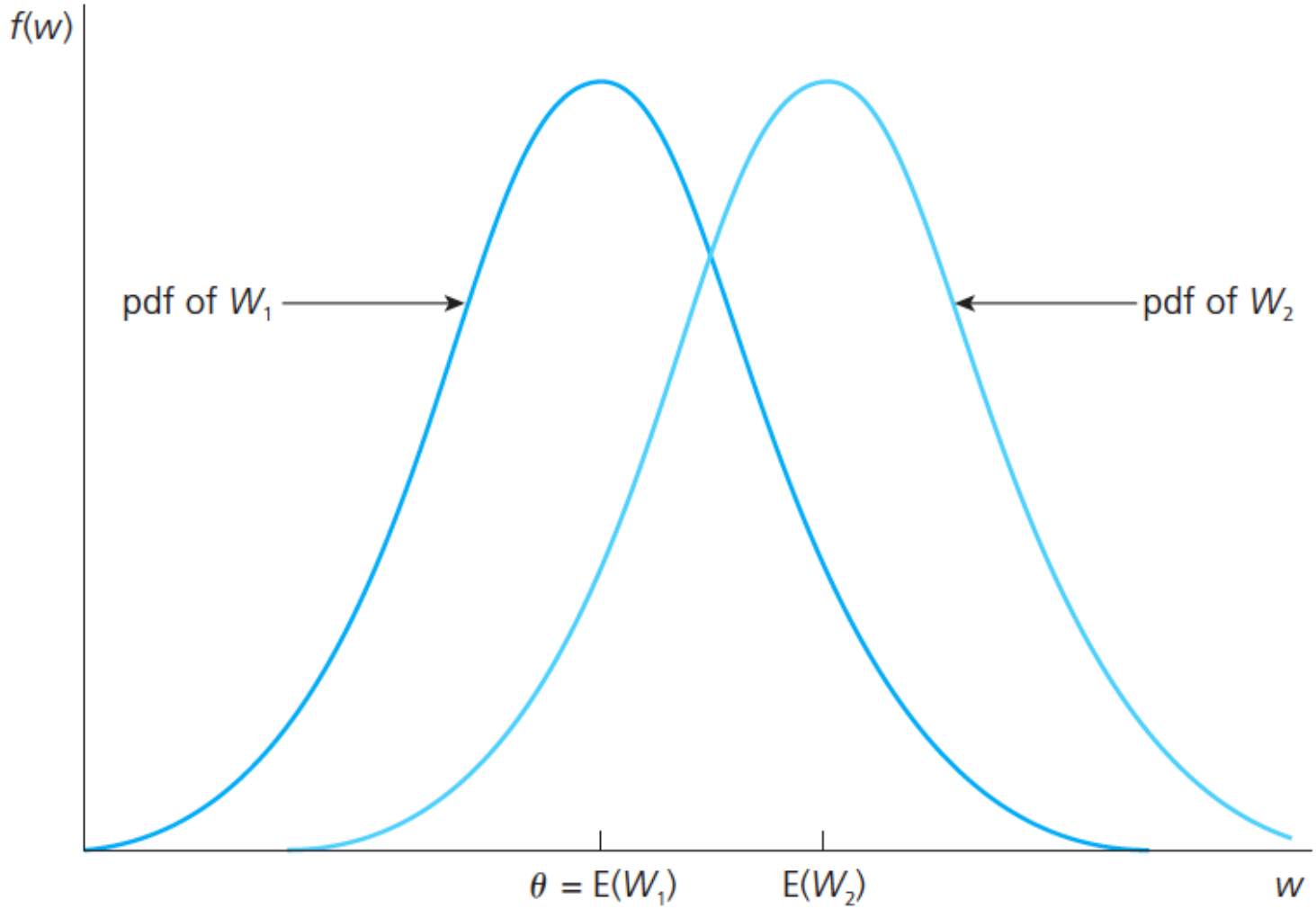
The general problem

- In general, there is some parameter θ to estimate and some sample of $X_1, X_2, ..., X_N$
- Our estimator of θ will be a function of $X_1, X_2, ..., X_N$:

$$W = h(X_1, X_2, ..., X_N)$$

- Different samples of $X_1, X_2, ..., X_N$ will lead to different values of w
- W is itself a random variable
- We say that W is an **unbiased** estimator of θ if $E(W) = \theta$

Biased estimation



Unbiased doesn't mean our estimate is always good! But better than biased...

Sample variance

- Recall that the variance of a random variable is given by $E((X - E(X))^2)$
- You observe a sample $X_1, X_2, X_3, ..., X_N$.
- How do you estimate the variance from the sample?

Sample variance

- Recall that the variance of a random variable is given by $E((X - E(X))^2)$
- You observe a sample $X_1, X_2, X_3, ..., X_N$.
- How do you estimate the variance from the sample?
- Variance is a kind of population mean

Sample variance

- Recall that the variance of a random variable is given by $E((X - E(X))^2)$
- You observe a sample $X_1, X_2, X_3, ..., X_N$.
- How do you estimate the variance from the sample?
- Variance is a kind of population mean
- Sample mean is an unbiased estimator of the population mean, so

$$\frac{1}{N}\sum_n (X_n - E(X))^2 \qquad ???$$

Sample variance

- The problem with using $\frac{1}{N}\sum_n (X_n - E(X))^2$ as an estimate of V(X) is that you don't know E(X)

Sample variance

- The problem with using $\frac{1}{N}\sum_n (X_n - E(X))^2$ as an estimate of V(X) is that you don't know E(X)
- But we can estimate it using $\bar{X}$ (the sample mean), so

$$\frac{1}{N}\sum_n (X_n - \bar{X})^2 \qquad ???$$

Sample variance

- The problem with using $\frac{1}{N}\sum_n (X_n - E(X))^2$ as an estimate of V(X) is that you don't know E(X)
- But we can estimate it using $\bar{X}$ (the sample mean), so

$$\frac{1}{N}\sum_n (X_n - \bar{X})^2 \qquad ???$$

- It turns out that we need to correct for the fact that we are using an estimate to define our estimate:

$$\text{Sample Variance} = S^2 = \frac{1}{N-1}\sum_n (X_n - \bar{X})^2$$

Sample variance

- The problem with using $\frac{1}{N}\sum_n (X_n - E(X))^2$ as an estimate of V(X) is that you don't know E(X)
- But we can estimate it using $\bar{X}$ (the sample mean), so

$$\frac{1}{N}\sum_n (X_n - \bar{X})^2 \qquad ???$$

- It turns out that we need to correct for the fact that we are using an estimate to define our estimate:

$$\text{Sample Variance} = S^2 = \frac{1}{N-1}\sum_n (X_n - \bar{X})^2$$

- Notice we usually use σ^2 to denote the population variance and S^2 to denote the sample variance

Covariance and correlation

- We will often be interested in studying how two variables move together
- Assume that our sample consists of *pairs* of variables (X_1, Y_1), (X_2, Y_2), etc
- Sample covariance of X and Y:

$$S_{xy} = \frac{1}{N-1} \sum_n (X_n - \bar{X})(Y_n - \bar{Y})$$

Covariance and correlation

- We will often be interested in studying how two variables move together
- Assume that our sample consists of *pairs* of variables (X_1, Y_1), (X_2, Y_2), etc
- Sample covariance of X and Y:

$$S_{xy} = \frac{1}{N-1} \sum_n (X_n - \bar{X})(Y_n - \bar{Y})$$

Sample correlation:

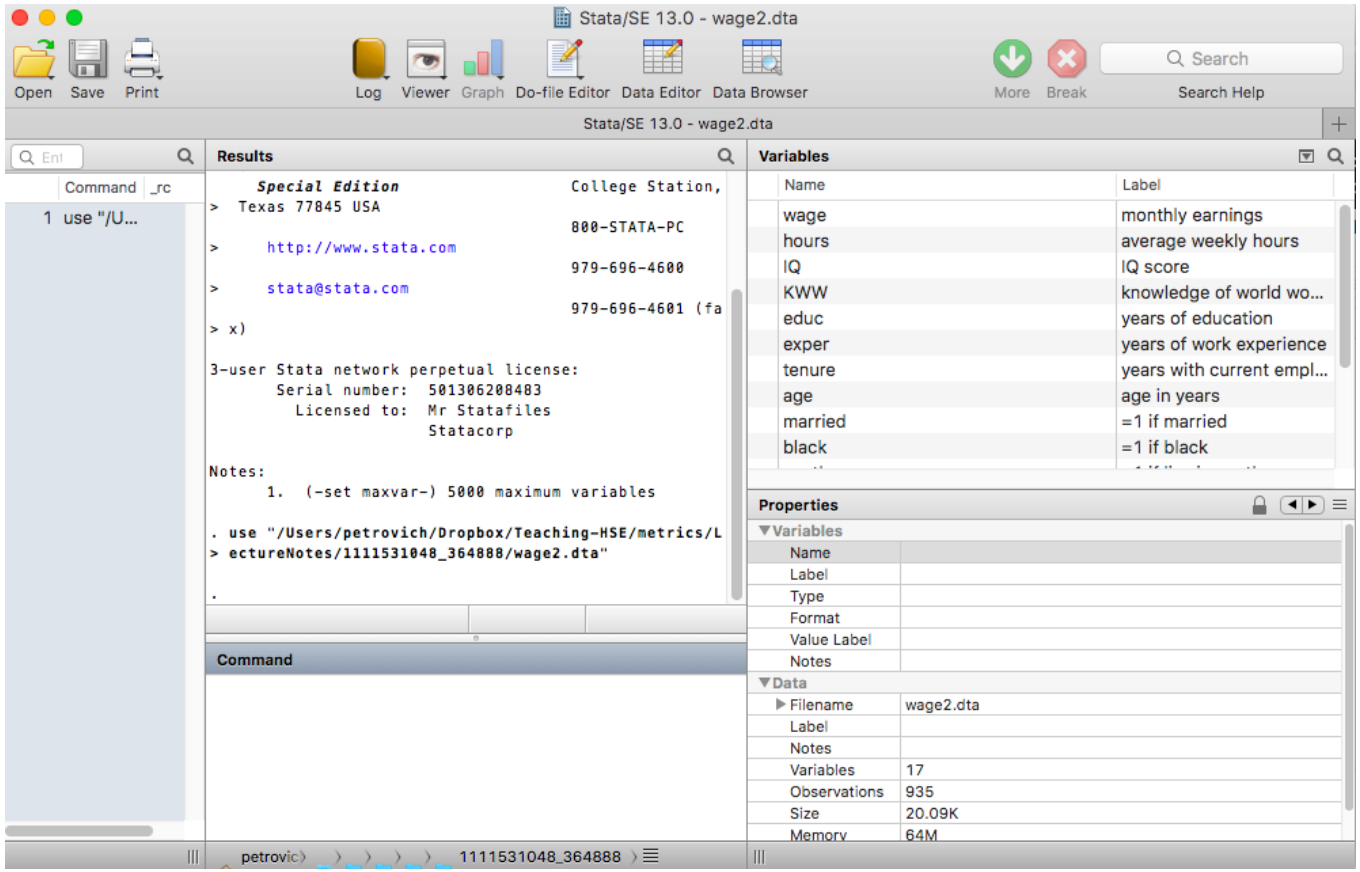
$$R_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{\sum_n (X_n - \bar{X})(Y_n - \bar{Y})}{\sum_n (X_n - \bar{X})^2 \sum_n (Y_n - \bar{Y})^2}$$

Working in Stata

- We can use Stata to apply some of the concepts we learned today to real data
- First, we load a dataset into Stata. Type:

```
use http://qcpages.qc.cuny.edu/~rvesselinov/statadata/WAGE2.DTA
```

- You will see a screen that looks like this:



Working in Stata

- You can use the tabulate command to obtain information about the distribution of a variable Type:

```
tab educ
```

- This will produce output that looks like this:

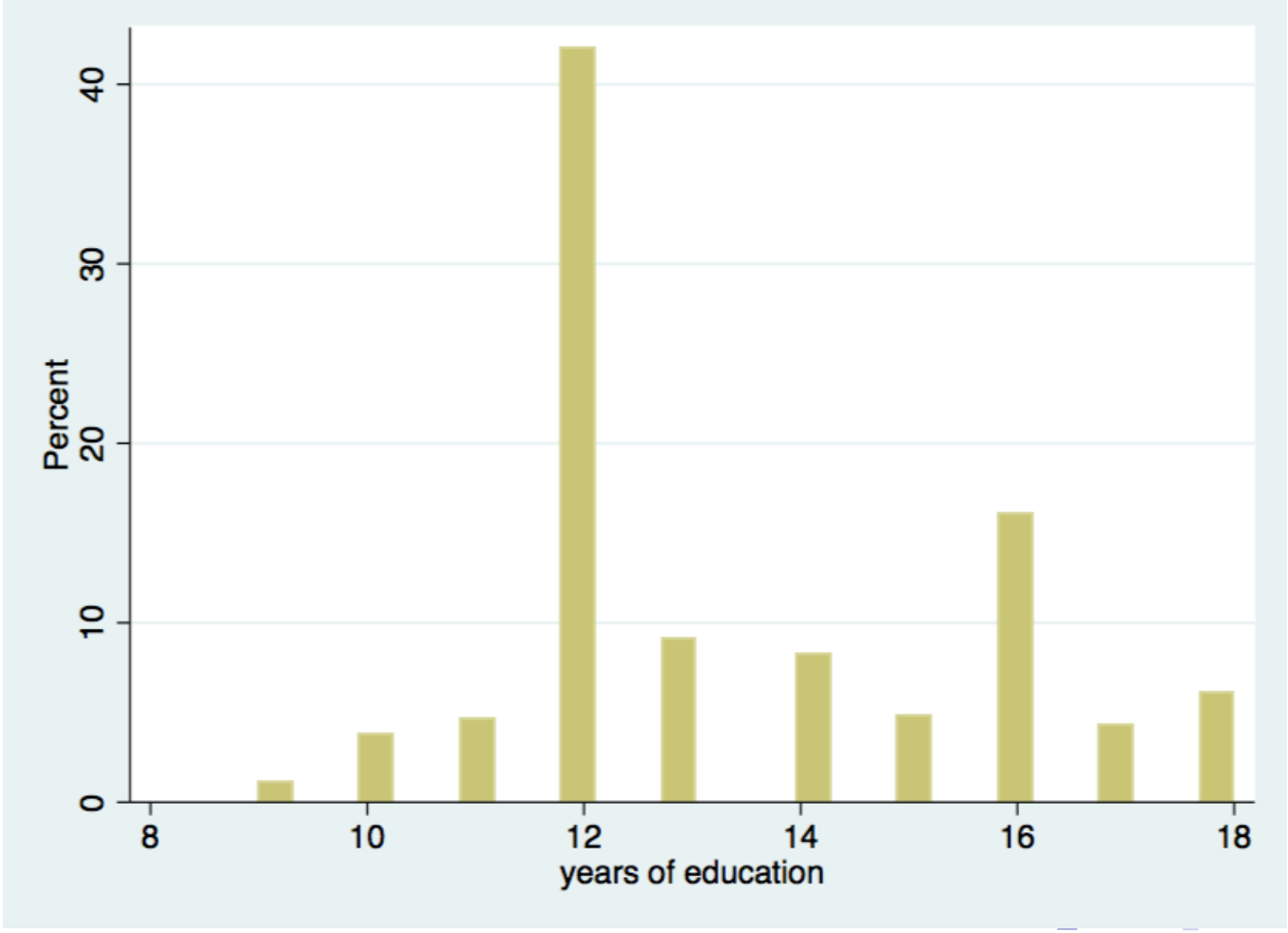
```
 years of |
education |      Freq.     Percent        Cum.
------------+-----------------------------------
          9 |         10        1.07        1.07
         10 |         35        3.74        4.81
         11 |         43        4.60        9.41
         12 |        393       42.03       51.44
         13 |         85        9.09       60.53
         14 |         77        8.24       68.77
         15 |         45        4.81       73.58
         16 |        150       16.04       89.63
         17 |         40        4.28       93.90
         18 |         57        6.10      100.00
------------+-----------------------------------
      Total |        935      100.00
```

Working in Stata

- You can use the histogram command to represent the distribution visually: Type:

```
hist educ, percent
```

- This will produce output that looks like this:



Working in Stata

- You can use the summarize command to obtain information about the mean and standard deviation Type:

```
sum educ
```

- This will produce output that looks like this:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        educ |       935    13.46845    2.196654          9         18
```

Working in Stata

- You can use the detail option to obtain more detailed information: Type:

```
sum educ, detail
```

- This will produce output that looks like this:

```
                    years of education
-------------------------------------------------------------
      Percentiles      Smallest
 1%           9              9
 5%          11              9
10%          12              9       Obs                 935
25%          12              9       Sum of Wgt.         935

50%          12                      Mean           13.46845
                     Largest         Std. Dev.      2.196654
75%          16             18
90%          17             18       Variance       4.825288
95%          18             18       Skewness       .5477959
99%          18             18       Kurtosis       2.262651
```

Y is the X-th percentile $\Rightarrow$ X% of the data takes on values smaller than Y

50-th percentile has a special name: **the median**

Working in Stata

- You can use the pwcorr command to obtain correlation between two variables: Type:

```
pwcorr wage educ
```

- This will produce output that looks like this:

```
             |     wage     educ
-------------+------------------
        wage |   1.0000
        educ |   0.3271   1.0000
```

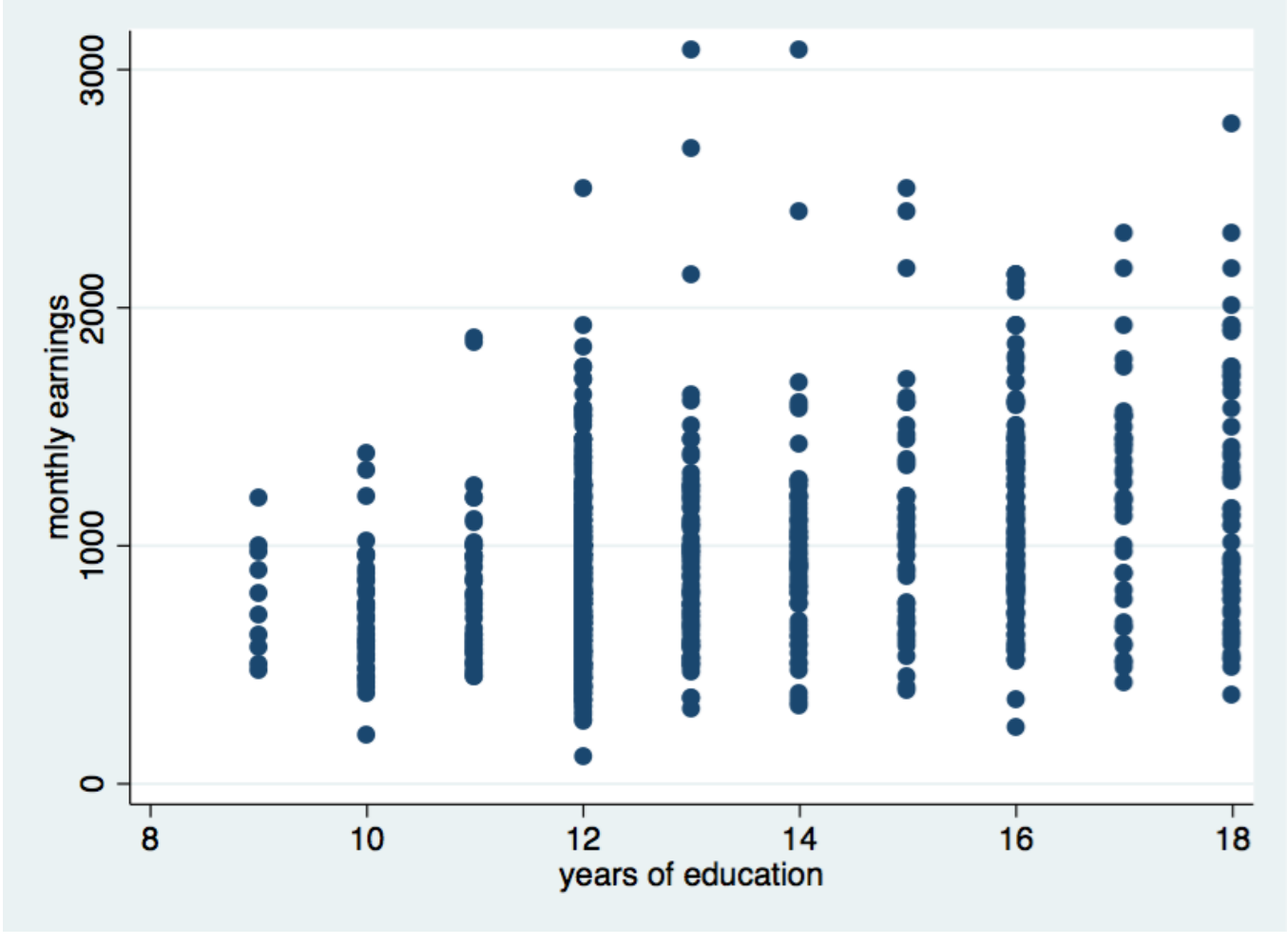
- The results tell us that wage and education are positive correlated

Working in Stata

- We can represent the correlation graphically using the scatter command: Type:

```
scatter wage educ
```

- This will produce output that looks like this:

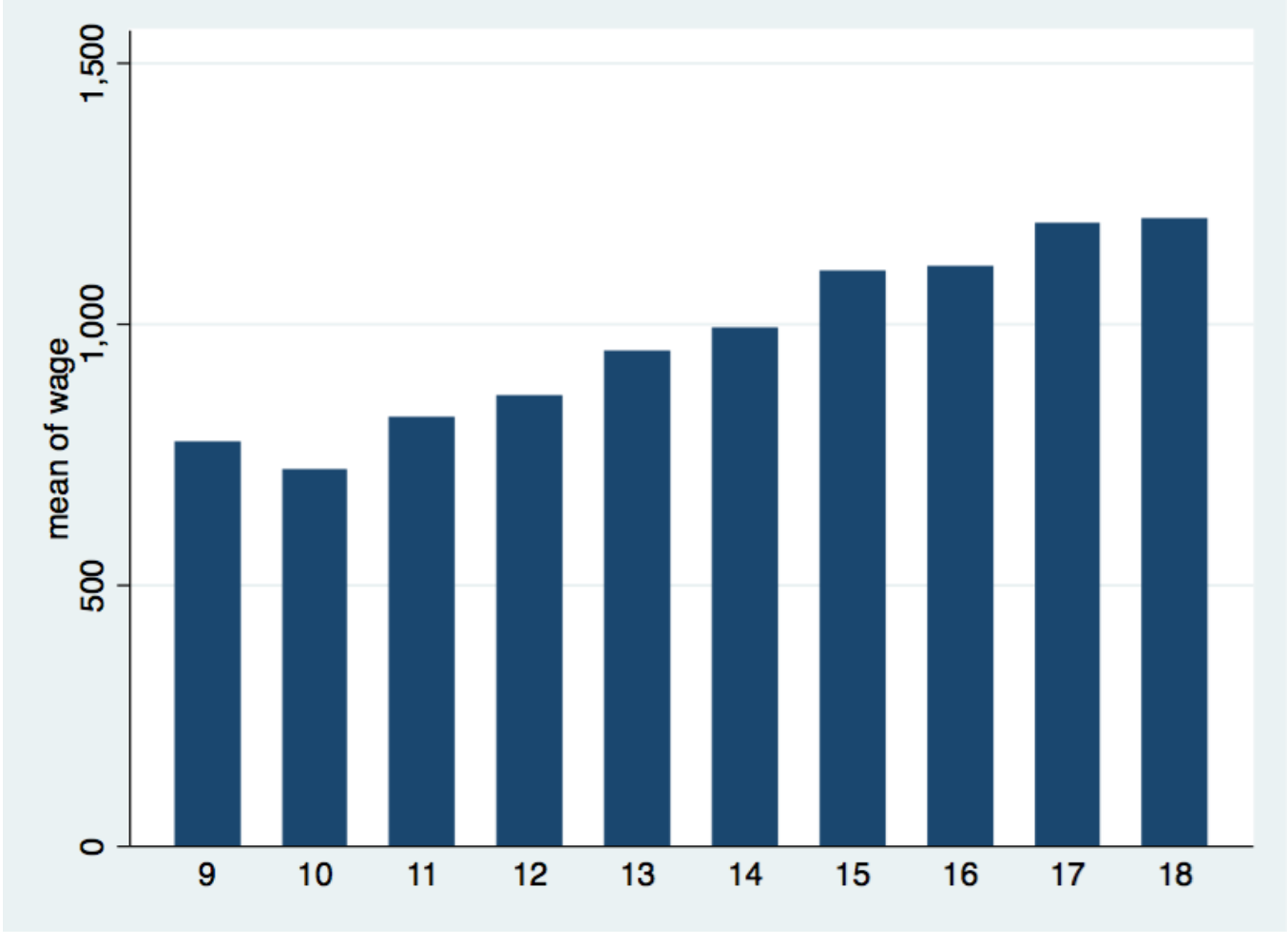


Working in Stata

- We can also represent the mean wage for each education level by using the bar plot: Type:

```
graph bar wage, over(educ)
```

- This will produce output that looks like this:



Working in Stata

- We can also make separate bars for people in and out of the south: Type:

```
gen wage1=wage if south==0
gen wage2=wage if south==1
graph bar wage1 wage2, over(educ)
```

- This will produce output that looks like this: