Economics 113

Agenda

1. Course information
2. Define Statistics/Econometrics.
3. Key things we'll use throughout the semester
4. Summary measures: Central Tendency
5. Summary measures: Variability
6. Covariance and Correlation
7. Data Scaling
8. Quote of the Day:

   I'm a great believer in luck, and I find the harder I work, the more I have of it.

   -Thomas Jefferson

Course information

- The syllabus is the main source of information for the course. Please check the syllabus before asking questions.
- Website: http://people.ucsc.edu/~aspearot
- Exams:
  - Exam 1: Wednesday, January 23rd In-class
  - Exam 2: Friday, February 8th In-class
  - Exam 3: Friday, March 1st In-class
  - Exam 4: Wednesday, March 20th 8-11AM
  - *You have a week from now to let me know of any issues with any exam.*
- Grades:
  - 20% Homework, 20% each exam.
- Book:
  - *Introductory Econometrics* by Jeffrey Wooldridge
- Office hours: 9:30-11:30AM Wednesdays, 459 Engineering 2
- Email: aspearot@ucsc.edu

Course information

Grades

- Yes, I do curve.
- No, it is not consistent from quarter to quarter.
- The curve will be worse if you disrespect your classmates, TAs, or me.
- Don't cheat. It will not be tolerated. If you're caught, you will receive a failing grade and be reported for academic misconduct.

Computer Program Stata

- We will eventually use the statistics program "Stata" to work on computer problems
- You are expected to either buy the program or use it at a lab.
  - It is available in limited UCSC computer labs (Class Folders/Economics).
  - There is a link on the course website to the Stata "Grad plan". The cheapest version is $30, which will be sufficient for the class. I HIGHLY RECOMMEND THIS OPTION.

Course wisdom

On Soapbox

- Emails
  - Do not send frivolous emails.
  - If you have something important to talk about, it's always best to do it in person.
  - If your email can be answered by looking at the website or syllabus, it likely won't be answered.
- Other
  - Exams won't (and shouldn't) be the homework with different numbers.
  - Now is the time to work hard.
  - You will solve problems in this course.
  - This course is useful. Believe it!!!

Basics

- What is statistics?
    1. Collecting raw data
    2. Manipulating raw data
    3. Summarizing data
- Types of statistics
    1. Descriptive statistics
    2. Inferential Statistics

Basics

Descriptive

- Descriptive statistics
  1. Numbers that summarize or describe
  2. Plots
- Descriptive Example:
  1. The Detroit foreclosure rate was 5% in 1997.
  2. Economics graduates 12 students per faculty member.

Basics

Inferential

- Inferential statistics
  1. Estimation
  2. Hypothesis testing
  3. Draw conclusions, given randomness
- Example
  1. Higher tariffs have a statistically significant effect on trade.
  2. At 99% confidence, living in an area with a significant cancer risk lowers housing prices between 11% and 20%.
- *Inferential statistics deal with hypotheses!*

Basics

- What is econometrics?

  $\Rightarrow$ *The application of statistics to economic issues*

- What are economic issues?

  $\Rightarrow$ *Anything dealing with the interactions of agents*

- What's the trick?

  1. Finding the right data.
  2. Settings are often *non-experimental*

Key Terms

- Population

  The "universe". All items of interest.

  Rarely view the entire population.

- Sample

  A random sample of the population

- Parameter

  Summary value of the population

- Statistic

  Summary measure of the sample

Summary measures

Preliminaries

- Σ is shorthand for addition
- Suppose x_i is the *ith* observation. $\Longrightarrow$

$$x_1 + x_2 + x_3 = \sum_{i=1}^{3} x_i$$

- If a is a constant $\Longrightarrow$

$$\sum_{i=1}^{3} a = 3a$$

- Mixed example: $\Longrightarrow$

$$\sum_{i=1}^{3} a x_i = a \sum_{i=1}^{3} x_i$$

Summary measures

Preliminaries

- b a constant
- n observations
- Simplify $\sum_{i=1}^{n}(a + bx_i)$
- Group terms! $\Longrightarrow$

$$\begin{aligned}
\sum_{i=1}^{n}(a + bx_i) &= a + bx_1 + a + bx_2 + \ldots + a + bx_n \\
&= a + a + a \ldots . bx_1 + bx_2 + \ldots bx_n \\
&= an + b\sum_{i=1}^{n} x_i
\end{aligned}$$

Summary measures

Preliminaries

- Let $\{x_i, y_i\}$ be paired observations
- k another constant.
- Try this one.

$$\sum_{i=1}^{n} \left(ak + bx_i + c\left(x_i y_i\right)\right) = ???$$

- Treat $x_i y_i$ as any other random variable:

$$\sum_{i=1}^{n} \left(ak + bx_i + c\left(x_i y_i\right)\right) = nak + \sum_{i=1}^{n} bx_i + \sum_{i=1}^{n} \left(cx_i y_i\right)$$

- Pull out the constants:

$$nak + \sum_{i=1}^{n} bx_i + \sum_{i=1}^{n} \left(cx_i y_i\right) = nak + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} \left(x_i y_i\right)$$

Summary measures

Preliminaries

- Standard notation

	Estimated from the sample	**Parameter from the population**
Mean	$\widehat{\mu}_x$	μ_x
Variance	$\widehat{\sigma}_x^2$	σ_x^2
Standard Deviation	$\widehat{\sigma}_x$	σ_x
Size	n	N

- We estimate parameters to describe a population using the sample.
- The "hats" represent statistics estimated using the sample.
- What properties should these statistics have?

Summary measures

An example

- We want to estimate the height of all UCSC students.
- Is this expensive?
- Where should we go to sample?
  1. Bus stop?
  2. Bay tree?
  3. Bar?
  4. Office hours?
- Are there problems with these sampling points?

Summary measures

Statistics that describe distributions

- Measures of central tendency
  1. Mean
  2. Median
  3. Mode
- Measures of variation
  1. Range
  2. Variance
  3. Standard Deviation
- Measures of shape
  1. "Skew"

Summary measures

Measures of central tendency

- **Mean**
  - Most common
  - Central tendency

$$\widehat{\mu}_x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Median**
  - Mid-point of the data.
  - To calculate:
    1. Order the data.
    2. Calculate $(n+1)/2$
    3. *Take the value at $(n+1)/2$, or the average of the two closest (if $(n+1)/2$ is not whole)*

Summary measures

Measures of central tendency (cont.)

- **Mode**
  - The value that occurs most often
- Question: which measure(s) of central tendency are not affected by extreme values?

  $\Rightarrow$ *Median and Mode*

- US income distribution?

- **Range**
  - A measure of data dispersion, though not used for many applications
  - To calculate:
    1. Identify the largest observation
    2. Identify the smallest observation
    3. Take the difference

Summary measures

Measures of variation (cont.)

- **Sample Variance**
  - The most commonly used measure of dispersion.
  - Summarizes the how far a typical observation is from the mean

$$\widehat{\sigma}_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \widehat{\mu}_x)^2$$

  - Why do we divide by $n - 1$ instead of n?

$\Rightarrow$ *We used one "piece" of information to calculate* $\widehat{\mu}_x$

- **Sample Standard deviation**

$$\widehat{\sigma}_x = \sqrt{\widehat{\sigma}_x^2}$$

- This is more desirable than the sample variance. Why?

Covariance

Relationships

- Covariance describes the relationship between two random variables

$$\widehat{\sigma}_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \widehat{\mu}_x\right) \left(y_i - \widehat{\mu}_y\right)$$

- When will covariance be positive/negative?
  - $\widehat{\sigma}_{xy} > 0 \Rightarrow$ tends to have $x_i > \widehat{\mu}_x$ when $y_i > \widehat{\mu}_y$ (and vice versa)
  - $\widehat{\sigma}_{xy} < 0 \Rightarrow$ tends to have $x_i > \widehat{\mu}_x$ when $y_i < \widehat{\mu}_y$ (and vice versa)
- Covariance describes a "linear" relationship
- Any non-linear relationships with zero covariance?

Covariance

Example

- US Imports and Exports, 2007, percentage change
- Do you think they positively or negatively covary?
- $IM = \{4, -3, 4 - 1\}$, $EX = \{1, 8, 19, 6\}$
  (rounded percentage changes)
- $\widehat{\sigma}_{EX,IM} = \frac{1}{n-1} \sum_{i=1}^{n} \left(EX_i - \widehat{\mu}_{EX}\right) \left(IM_i - \widehat{\mu}_{IM}\right)$
- Compute mean of $\widehat{\mu}_{IM}$:

$$\widehat{\mu}_{IM} = \frac{1}{4}(4 - 3 + 4 - 1) = \frac{1}{4}(4) = 1$$

- Compute mean of $\widehat{\mu}_{EX}$:

$$\widehat{\mu}_{EX} = \frac{1}{4}(1 + 8 + 19 + 6) = \frac{1}{4}(34) = 8.5$$

Covariance

Example

$$
\begin{aligned}
\widehat{\sigma}_{EX,IM} &= \frac{1}{4-1}\sum_{i=1}^{4}\left(EX_i - \widehat{\mu}_{EX}\right)\left(IM_i - \widehat{\mu}_{IM}\right) \\
&= \frac{1}{4-1}\left(\begin{array}{c} (4-1)(1-8.5)+(-3-1)(8-8.5)+ \\ (4-1)(19-8.5)+(-1-1)(6-8.5) \end{array}\right) \\
&= \frac{1}{4-1}(-3\cdot 7.5+4\cdot 0.5+3\cdot 10.5+2\cdot 2.5) \\
&= \frac{1}{4-1}(-22.5+2+31.5+5)=5.33
\end{aligned}
$$

Correlation

Basic

- Correlation describes a linear relationship
- $\widehat{\rho}_{xy} = \frac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x \widehat{\sigma}_y}$
- $\widehat{\rho}_{xy} \in [-1, 1]$

Correlation

- $IM = \{4, -3, 4, -1\}$, $EX = \{1, 8, 19, 6\}$
- $\widehat{\mu}_{IM} = 1$, $\widehat{\mu}_{EX} = 8.5$, $\widehat{\sigma}_{EX,IM} = 5.33$
- Compute $\widehat{\sigma}_{IM}$:

$$\begin{aligned}
\widehat{\sigma}_{IM} &= \sqrt{\frac{1}{3}((4-1)^2 + (-3-1)^2 + (4-1)^2 + (-1-1)^2)} \\
&= \sqrt{\frac{1}{3}(9+16+9+4)} = \sqrt{\frac{1}{3}(38)} = 3.559026
\end{aligned}$$

- Compute $\widehat{\sigma}_{EX}$:

$$\begin{aligned}
\widehat{\sigma}_{EX} &= \sqrt{\frac{1}{3}((1-8.5)^2 + (8-8.5)^2 + (19-8.5)^2 + (6-8.5)^2)} \\
&= \sqrt{\frac{1}{3}((-7.5)^2 + (-0.5)^2 + (10.5)^2 + (-2.5)^2)} \\
&= 7.593857
\end{aligned}$$

- $\widehat{\rho}_{zy} = \frac{\widehat{\sigma}_{zy}}{\widehat{\sigma}_z \widehat{\sigma}_y} = \frac{5.33}{3.559 * 7.59} = 0.1973124$

Data Scaling

Mean

- What happens when we scale variables by either adding a constant or multiplying by a constant?
- Sample: $X = \{2, 3, 4\}$
- Calculate the mean of X:

$$
\begin{aligned}
\widehat{\mu}_x &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{3}(2+3+4) \\
&= 3
\end{aligned}
$$

- Calculate the variance of X:

$$
\begin{aligned}
\widehat{\sigma}_x^2 &= \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \widehat{\mu}_x)^2 \\
\widehat{\sigma}_x^2 &= \frac{1}{2}\left((2-3)^2 + (3-3)^2 + (4-3)^2\right) = \frac{1}{2}(1+0+1) \\
&= 1
\end{aligned}
$$

Scaling

Adding a constant

- What if we define a new variable, $Z = X + 3$
- Calculate the mean of $Z = \{5, 6, 7\}$
- What will happen to the mean, variance?
- $\widehat{\mu}_z = \frac{1}{3}(5 + 6 + 7) = 6$
- $\widehat{\sigma}_z^2 = \frac{1}{2}\left((5-6)^2 + (6-6)^2 + (7-6)^2\right) = \frac{1}{2}(1 + 0 + 1) = 1$
- Adding a constant **does** affect central tendency.
- Adding a constant **does not** affect dispersion.
- $1 \quad \underline{\mathbf{2} \quad \mathbf{3} \quad \mathbf{4}} \quad 5 \quad 6 \quad 7 \quad 8 \quad 9$
- $1 \quad 2 \quad 3 \quad 4 \quad \underline{\mathbf{5} \quad \mathbf{6} \quad \mathbf{7}} \quad 8 \quad 9$

Scaling

Multiplying by a constant

- What if we define a new variable, $J = 3X$
- Calculate the mean/variance of $J = \{6, 9, 12\}$
- What will happen to the mean, variance?
- $\widehat{\mu}_j = \frac{1}{3}(6+9+12) = 9$
- $\widehat{\sigma}_j^2 = \frac{1}{2}\left((6-9)^2 + (9-9)^2 + (12-9)^2\right) = \frac{1}{2}(9+0+9) = 9$
- Multiplying by a constant affects both central tendency and dispersion.

$$1 \quad \underline{\mathbf{2} \quad \mathbf{3} \quad \mathbf{4}} \quad 5 \quad 6 \quad 7 \quad 8 \quad 9 \quad 10 \quad 11 \quad 12$$

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad \underline{\mathbf{6} \quad 7 \quad 8 \quad \mathbf{9} \quad 10 \quad 11 \quad \mathbf{12}}$$

- Generally, if $J = aX$, then $\widehat{\sigma}_j^2 = a^2\widehat{\sigma}_x^2$, $\widehat{\mu}_j = a\widehat{\mu}_x$

- Suppose $Z = aX$. What is $\widehat{\sigma}_{zy}$?
- Write $\widehat{\sigma}_{zy}$

$$\widehat{\sigma}_{zy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(z_i - \widehat{\mu}_z\right)\left(y_i - \widehat{\mu}_y\right)$$

- Substitute for $\widehat{\mu}_z$ and z_i
- This gives:

$$\widehat{\sigma}_{zy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(ax_i - a\widehat{\mu}_x\right)\left(y_i - \widehat{\mu}_y\right)$$

- Factor out a and simplify:

$$\widehat{\sigma}_{zy} = a\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \widehat{\mu}_x\right)\left(y_i - \widehat{\mu}_y\right) = a\widehat{\sigma}_{xy}$$

- Covariance is sensitive to scale!! Is this a problem?

Correlation

Scaling

- Is correlation sensitive to scale?
- Suppose $Z = aX$, $a > 0$.
- $\widehat{\rho}_{zy} = \dfrac{\widehat{\sigma}_{zy}}{\widehat{\sigma}_z\widehat{\sigma}_y} = \dfrac{a\widehat{\sigma}_{xy}}{a\widehat{\sigma}_x\widehat{\sigma}_y} = \dfrac{\widehat{\sigma}_{xy}}{\widehat{\sigma}_x\widehat{\sigma}_y} = \widehat{\rho}_{xy}$